A Competing Risks Model for the Time to First Goal

Petr Volf

Institute of Information Theory and Automation, Czech Ac. Sci., Prague
volf@utia.cas.cz

Abstract

In the contribution the time to the first goal in a football (soccer) match is analyzed, in the framework of competing risks scheme. Potential random times to the first goals scored by both teams are modelled by exponential distributions with parameters depending on attack and defence strengths of teams. Mutual dependence of these two times is described with the aid of a conveniently chosen copula ensuring the model identifiability. As a real example the data from the 2014 World Championship are analyzed. It is shown that the correlation is, as a rule, negative, and is absolutely larger in more competitive matches. Possible extensions of the approach are discussed, too.

1 Introduction

A basic probability model for final score of a football (soccer) match, presented for instance already in [3], consist 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 certain more frequent results. Another generalization can consist in considering a time development of model parameters as well as covariates during the match (see for instance [6]), in such a way a model based on counting process scheme is obtained. The present contribution concerns yet another direction of basic model improvement, namely to models considering an explicit form of dependence of both teams scoring distributions. Thus, in [2] a special case of bivariate Poisson distribution was employed. In the same context, McHale and Scarf [4] 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 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 the 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 the present contribution we analyze continuous distribution of time to the first scored goal in a match. Consequently, we deal with the scheme of competing risks. On the one hand the use of copula for two-dimensional continuous distribution can lead to a 'nice' closed form of the model, on the other hand it is well known that in the competing risks setting the model may be non-identifiable. A proof and an example of this phenomenon is given in [5], some instances of identifiable (or not) models are treated in [1] – in these classical studies the notion of copula has not been used yet. Therefore we are facing the problem of reasonable copula selection. Fortunately, it is known (cf. [7]), that the selection of copula type is not so crucial, that the finding proper value of its parameter (connected with correlation) is much more important.

Potential times to the first goal scored by each team are modelled by exponential distribution following from the basic Poisson model proposed in [3]. The parameters again consist of attack

and defence parameters of both teams. Further, for joint survival function we use a copula derived from Tsiatis' [5] example, its form is convenient for work with exponential distribution. Such a combination of marginal and simultaneous distributions has already been analyzed in [1] and proved to be identifiable. From this fact the consistency of estimates follows.

The outline of the paper is the following: The next section recalls the scheme of competing risks and the problem of possible non-identifiability. Then the copula model and corresponding likelihood is formulated. Section 4 then contains a real example, namely the analysis of data from the 2014 Football World Championship (in Brazil). The results are quite comparable with conclusions in [4], namely that estimated correlation is, as a rule, negative, and is absolutely large in more competitive matches, i.e. the matches of teams with good defence and comparable attack abilities.

2 Competing Risks Scheme

Let us assume that certain event (e. g. a failure of a device) can be caused by K reasons. Therefore we consider K (possibly dependent) random variables - survival times $T_j, j = 1, ..., K$, sometimes plus 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' survivals T_j we 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 $Z = C$.

Such data lead to 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 **cause–specific hazard functions** for events $j = 1, 2, \ldots, K$:

$$h_j^*(t) = \lim_{d \to 0} \frac{P(t \le Z < t + d,\, \delta = j \mid Z \ge t)}{d},$$

and the **cumulative incidence functions**

$$F_j^*(t) = P(Z \le t, \delta = j) = \int_0^t S(s) \cdot h_j^*(s)\, \mathrm{d}s.$$

As both components, i.e. S and h_j^*, are estimable consistently by standard survival analysis methods, there also exist consistent estimates of F_j^*.

2.1 Problem of Non-Identifiability

However, in general, from data (Z_i, δ_i), $i = 1, \ldots, N$ it is not possible to identify neither marginal nor joint distribution of $\{T_j\}$. A. Tsiatis [5] 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 $h_j^*(t)$. In other words, even if the model is parametrized and the MLE yields consistent estimates, in general we do not know parameters of which model are estimated. The situation can be better in the case of a regression model, because the covariates provide an additional information, especially when their structure is rich enough. On the other hand, as a consequence of the Tsiatis [5] result, in competing risks models without regressors it is necessary to make certain functional form assumptions about both marginal and joint distribution in order to identify them. Several such cases are studied in [1] and in some other papers.

3 Competing Risks and Copula

In the sequel we shall consider just 2 random variables S, T and data $Z_i = min(S_i, T_i, C_i), \delta_i = 1, 2, 0$. The notion of copula offers a way how to model multivariate distributions, we prefer here to use it for modelling the joint survival function $\overline{F_2}(s,t)$ of S, T:

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

$\overline{F_S}$, $\overline{F_T}$ are marginal survival 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, which is, as a rule, uniquely connected with correlation of U, V, hence also with correlation 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 another point of view, the identifiability of the copula (and its parameter) and marginals can be considered as two separate steps.

Zheng and Klein [7] proved that when the copula is known, the marginal distributions are estimable consistently (and then the joint distribution, too, from (1)), even in non-parametric (so that quite general) setting. However, in general, also the knowledge of θ is needed. They also discussed importance of proper selection of copula form. As it has already been said, 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 its parameter.

3.1 Copula Based on Tsiatis' Example

Let us return to the example of Tsiatis [5], considering just $K = 2$ random variables S, T with exponential distribution and the following marginal and joint survival functions,

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

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:

$$C(u,v) = u \cdot v \cdot \exp(-\theta \cdot \ln u \cdot \ln v) \tag{2}$$

with $\theta \geq 0$, corresponding correlation $\rho(U,V) \leq 0$, $\theta = 0$ means independence of U, V. The parameters are connected in the following way: $\gamma = \theta \cdot \lambda \cdot \mu$. Figure 1 shows the dependence of correlation on parameters.

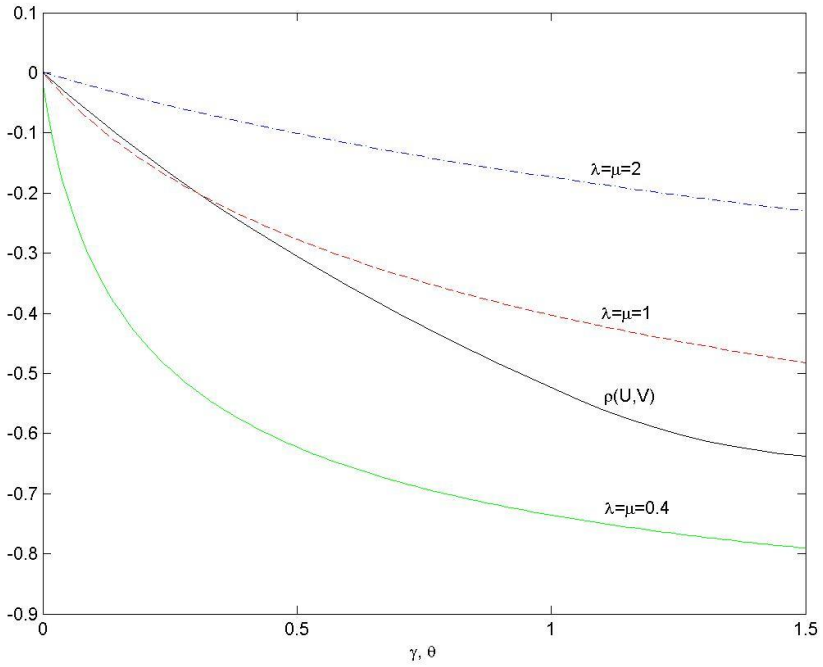

Figure 1: Dependence of $\rho(U,V)$ on parameter θ and $\rho(S,T)$ on γ, when $S \sim Exp(\lambda)$, $T \sim Exp(\mu)$.

It is easy to show that the case of competing risks with two exponential marginal distributions tied together by copula (2) is identifiable. It actually has been proven already by Basu and Ghosh [1] – though authors did not use a notion of copula yet. It is also easy to verify that the case fulfils the regularity conditions and therefore yields unique ML estimates of parameters. The likelihood function has the following form:

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

where again $Z_i = min(S_i, T_i, C_i), \delta_i = 1, 2, 0$.

3.2 Other Two-Dimensional Exponential Distributions

The identifiability results obtained above 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

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

It is easy to check that the corresponding competing risks model is 'over-parametrized', determined by any couple of parameters only, i.e. we cannot estimate λ, μ, θ uniquely.

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 [2].

4 Application to Time of the First Goal

We shall now use the competing risk model derived in Part 3.1 to modelling the time to first scored goal during a football (soccer) match. Marginal variables are the 'latent' times of 1-st goal of each time, however only the incidence of one of them is observed. Or, in the case of draw 0:0, we have censoring by a fixed value 90 minutes (or 120 minutes in the case of prolonged match). Except statistical estimation of model parameters, we are interested in the main question: How dependent are these 'latent' times to 1-st goal of both teams?

In our rather small study we shall use the data from the Football World Championship 2014 in Brazil. 32 participating teams played together 64 matches, some of them just 3 matches in a group. In order to improve this proportion (matches to team), we considered just 11 'teams': 8 teams passing to quarterfinal, then 'team' No 9 - aggregated data of 8 teams losing in eight-final matches, No 10 - teams taking 3-rd places in groups, No 11 - teams ending 4-th in groups. Let us recall also the final order of the championship: 1. Germany, 2. Argentina, 3. The Netherlands, 4. Brazil.

As regards marginal models, the source was the standard model of Maher [3]. More specifically, each team (i) was characterized by its attack parameter a_i and defense parameter b_i. The sequence of scoring in a match between teams i and j is then described by two Poisson processes with intensities $a_i \cdot b_j$, $a_j \cdot b_i$, respectively. Consequently, the time to the 1-st goal followed from two competing exponential random variables

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

Further, it was assumed that their mutual dependence can be expressed via 'Tsiatis' model described in Part 3.1.

Thus, we were facing the problem of the maximum likelihood estimation (MLE) of 23 parameters, a_i, b_i of 11 teams and γ characterizing the dependence. It was assumed that γ was the same for all couples of teams, i.e. in all matches. The results of the MLE of teams parameters are displayed in Table 1. For computational convenience, we estimated $\alpha_i = \ln a_i$, $\beta_i = \ln b_i$. Finally, the MLE of parameter γ was 0.605, with half-width of approximate 95% confidence interval 0.143.

It is possible to say that our result is comparable with the findings of McHale and Scarf [4]. Inspection of the graph on Figure 1 indicates that in the match of teams with very good defense (as for instance Germany and Argentina) the first goal really matters, correlation is large (absolutely), while in a opposite case of weaker defense and sufficiently good attack ability (here for instance Columbia and - rather surprisingly - Brazil) the correlation is smaller (but still negative).

Team	alpha		beta		a	b
Brazil	0.8408	(1.1756)	0.6683	(1.1519)	2.3181	1.9509
Netherlands	0.3580	(1.2784)	-0.9680	(2.1790)	1.4305	0.3799
Columbia	0.8542	(1.0408)	-0.2337	(2.0061)	2.3496	0.7916
Costa Rica	-0.9342	(2.3540)	-1.6044	(3.6409)	0.3929	0.2010
France	-0.2146	(1.5123)	-0.8994	(2.1512)	0.8068	0.4068
Argentina	0.4830	(0.9885)	-4.6885	(13.4755)	1.6209	0.0092
Germany	0.6888	(0.9692)	-5.0360	(17.2193)	1.9913	0.0065*
Belgium	-0.7982	(2.1896)	-0.2884	(1.5620)	0.4501	0.7495
No 9	-0.1707	(0.6335)	-0.3715	(0.6572)	0.8431	0.6897
No 10	0.1572	(0.6805)	0.4176	(0.6815)	1.1702	1.5183
No 11	-1.3428	(1.6872)	0.4444	(0.5148)	0.2611	1.5596

Table 1: **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

5 Conclusion

We have studied the dependence of random variables – latent times of scoring the first goal in a 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. 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 [6] for an overview of models).

Acknowledgement. The research has been supported by the project No 13-14445S of the Czech Scientific Foundation (GA ČR).

References

[1] A.P. Basu and J.K. Ghosh. Identifiability of the multinormal and other distributions under competing risks model. *Journal of Multivariate Analysis*, 8:413–429, 1978.

[2] D. Karlis and I. Ntzoufras. Analysis of sports data by using bivariate poisson models. *J. R. Stat. Soc. Ser. D*, 52:381–394, 2003.

[3] M.J. Maher. Modelling association football scores. *Stat. Neerl.*, 36:109–118, 1982.

[4] I. McHale and P.A. Scarf. Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling*, 11:219–236, 2011.

[5] A. Tsiatis. A nonidentifiability aspects of the problem of competing risks. *Proc. Nat. Acad. Sci. USA*, 72:20–22, 1975.

[6] P. Volf. A random point process model for the score in sport matches. *IMA Journal of Management Mathematics*, 20:121–131, 2009.

[7] M. Zheng and J.P. Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82:127–138, 1995.

Proceedings of the 5th International Conference on Mathematics in Sport

Loughborough University, U.K.
29 June – 1 July 2015

Editors: Anthony Kay, Alun Owen, Ben Halkon, Mark King