Survival Modelling of Goal Arrival Times in Champions League

Mathsport 2019

Ilias Leriou, Ioannis Ntzoufras, Dimitris Karlis

Athens University of Economics and Business

Outline

1. Motivation

Outline

1. Motivation
2. Champions League 2017-2018 data layout

Outline

1. Motivation
2. Champions League 2017-2018 data layout
3. Initial Model

Outline

1. Motivation
2. Champions League 2017-2018 data layout
3. Initial Model
4. Final Model

Outline

1. Motivation
2. Champions League 2017-2018 data layout
3. Initial Model
4. Final Model
5. Issues and Further work

Acknowledgments

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under the HFRI PhD Fellowship grant (1809).

Motivation

- Modelling the number of events (goals) has been thoroughly addressed.

Motivation

- Modelling the number of events (goals) has been thoroughly addressed.
- Little research on modelling the goal arrival times

Motivation

- Modelling the number of events (goals) has been thoroughly addressed.
- Little research on modelling the goal arrival times
  - Thomas (2007)

    Analysis of inter-arrival times of goals in ice hockey using Weibull and Plateau-Hazard distributions.

Motivation

- Modelling the number of events (goals) has been thoroughly addressed.
- Little research on modelling the goal arrival times
    - Thomas (2007)

      Analysis of inter-arrival times of goals in ice hockey using Weibull and Plateau-Hazard distributions.
    - Nevo and Ritov (2013)

      Cox model for 1st & 2nd goal (760 Premier League games).

Motivation

- Modelling the number of events (goals) has been thoroughly addressed.
- Little research on modelling the goal arrival times
  - Thomas (2007)

    Analysis of inter-arrival times of goals in ice hockey using Weibull and Plateau-Hazard distributions.
  - Nevo and Ritov (2013)

    Cox model for 1st & 2nd goal (760 Premier League games).

AIM

Motivation

- Modelling the number of events (goals) has been thoroughly addressed.
- Little research on modelling the goal arrival times
  - Thomas (2007)

    Analysis of inter-arrival times of goals in ice hockey using Weibull and Plateau-Hazard distributions.
  - Nevo and Ritov (2013)

    Cox model for 1st & 2nd goal (760 Premier League games).

AIM

Since we are considering two arrival times, investigate the possible modeling of those goal arrival times using Bivariate distributions under a survival analysis framework.

Champions League 2017-2018 data layout

Let t_{1im} and t_{2im} be the event times for team 1 and team 2 respectively with $1 = 1, 2, ...n$ and $m = 1, 2, ..., M$ the game indicator. To be more precise, part of the data layout in our case is presented as follows:

Game	t_1	t_2	Home Team	Away Team
1	50	NA	Benfica	PFC CSKA Moskva
1	NA	13	Benfica	PFC CSKA Moskva
1	NA	8	Benfica	PFC CSKA Moskva
1	NA	NA	Benfica	PFC CSKA Moskva
...				
...				

Table 1: Data layout for survival modelling of Champions' League Data which consisted of 528 times (events and censored), 32 teams and 125 games

Champions League 2017-2018 data layout

Let t_{1im} and t_{2im} be the event times for team 1 and team 2 respectively with $1 = 1, 2, ...n$ and $m = 1, 2, ..., M$ the game indicator. To be more precise, part of the data layout in our case is presented as follows:

Game	t_1	t_2	Home Team	Away Team
1	50	NA	Benfica	PFC CSKA Moskva
1	NA	13	Benfica	PFC CSKA Moskva
1	NA	8	Benfica	PFC CSKA Moskva
1	NA	NA	Benfica	PFC CSKA Moskva
...				
...				

Table 1: Data layout for survival modelling of Champions' League Data which consisted of 528 times (events and censored), 32 teams and 125 games

Properties of the data

– Teams are competing with one another.

Champions League 2017-2018 data layout

Let t_{1im} and t_{2im} be the event times for team 1 and team 2 respectively with $1 = 1, 2, ...n$ and $m = 1, 2, ..., M$ the game indicator. To be more precise, part of the data layout in our case is presented as follows:

Game	t_1	t_2	Home Team	Away Team
1	50	NA	Benfica	PFC CSKA Moskva
1	NA	13	Benfica	PFC CSKA Moskva
1	NA	8	Benfica	PFC CSKA Moskva
1	NA	NA	Benfica	PFC CSKA Moskva
...				
...				

Table 1: Data layout for survival modelling of Champions' League Data which consisted of 528 times (events and censored), 32 teams and 125 games

Properties of the data

- Teams are competing with one another.
- After a team scores, **time resets to zero**.

Champions League 2017-2018 data layout

Let t_{1im} and t_{2im} be the event times for team 1 and team 2 respectively with $1 = 1, 2, ...n$ and $m = 1, 2, ..., M$ the game indicator. To be more precise, part of the data layout in our case is presented as follows:

Game	t_1	t_2	Home Team	Away Team
1	50	NA	Benfica	PFC CSKA Moskva
1	NA	13	Benfica	PFC CSKA Moskva
1	NA	8	Benfica	PFC CSKA Moskva
1	NA	NA	Benfica	PFC CSKA Moskva
...				
...				

Table 1: Data layout for survival modelling of Champions' League Data which consisted of 528 times (events and censored), 32 teams and 125 games

Properties of the data

- Teams are competing with one another.
- After a team scores, **time resets to zero**.
- NA-NA means that we are unable to observe at what time would a team have scored from the time that the last team scored until the end of the game

Champions League 2017-2018 data layout

Let t_{1im} and t_{2im} be the event times for team 1 and team 2 respectively with $1 = 1, 2, ...n$ and $m = 1, 2, ..., M$ the game indicator. To be more precise, part of the data layout in our case is presented as follows:

Game	t_1	t_2	Home Team	Away Team
1	50	NA	Benfica	PFC CSKA Moskva
1	NA	13	Benfica	PFC CSKA Moskva
1	NA	8	Benfica	PFC CSKA Moskva
1	NA	NA	Benfica	PFC CSKA Moskva
...				
...				

Table 1: Data layout for survival modelling of Champions' League Data which consisted of 528 times (events and censored), 32 teams and 125 games

Properties of the data

- Teams are competing with one another.
- After a team scores, **time resets to zero**.
- NA-NA means that we are unable to observe at what time would a team have scored from the time that the last team scored until the end of the game

Initial Model

Marshall Olkin Bivariate Weibull Distribution

Let U_0, U_1 and U_2 be independent Weibull random variables with the same shape parameter γ and scale parameters λ_0, λ_1 and λ_2 respectively.
Define

$$T_1 = U_0 \wedge U_1 \quad T_2 = U_0 \wedge U_2.$$

Then

$$(T_1, T_2) \sim MOBW(\gamma, \lambda_0, \lambda_1, \lambda_2)$$

Initial Model

Marshall Olkin Bivariate Weibull Distribution

Let U_0, U_1 and U_2 be independent Weibull random variables with the same shape parameter γ and scale parameters λ_0, λ_1 and λ_2 respectively.
Define

$$T_1 = U_0 \wedge U_1 \quad T_2 = U_0 \wedge U_2.$$

Then

$$(T_1, T_2) \sim MOBW(\gamma, \lambda_0, \lambda_1, \lambda_2)$$

The *Joint Probability Density Function* of the Marshall Olkin Bivariate Weibull distribution is given by

$$f_{T_1,T_2}(t_1, t_2) = \begin{cases} f_W(t_1; \gamma, \lambda_1) f_W(t_2; \gamma, \lambda_0 + \lambda_2) & \text{if } 0 < t_1 < t_2 \\ f_W(t_1, \gamma, \lambda_0 + \lambda_1) f_W(t_2, \gamma, \lambda_2) & \text{if } 0 < t_2 < t_1 \\ \frac{\lambda_0}{\lambda_0+\lambda_1+\lambda_2} f_W(t; \gamma, \lambda_0 + \lambda_1 + \lambda_2) & \text{if } 0 < t_1 = t_2 = t \end{cases}$$

where

$$f_W(x; \gamma, \lambda) = \gamma \lambda x^{\gamma-1} e^{-\lambda x^{\gamma}}$$

Initial Model

Marshall Olkin Bivariate Weibull Distribution

Let U_0, U_1 and U_2 be independent Weibull random variables with the same shape parameter γ and scale parameters λ_0, λ_1 and λ_2 respectively.
Define

$$T_1 = U_0 \wedge U_1 \quad T_2 = U_0 \wedge U_2.$$

Then

$$(T_1, T_2) \sim MOBW(\gamma, \lambda_0, \lambda_1, \lambda_2)$$

The *Joint Probability Density Function* of the Marshall Olkin Bivariate Weibull distribution is given by

$$f_{T_1,T_2}(t_1, t_2) = \begin{cases} f_W(t_1; \gamma, \lambda_1) f_W(t_2; \gamma, \lambda_0 + \lambda_2) & \text{if } 0 < t_1 < t_2 \\ f_W(t_1, \gamma, \lambda_0 + \lambda_1) f_W(t_2, \gamma, \lambda_2) & \text{if } 0 < t_2 < t_1 \\ \frac{\lambda_0}{\lambda_0+\lambda_1+\lambda_2} f_W(t; \gamma, \lambda_0 + \lambda_1 + \lambda_2) & \text{if } 0 < t_1 = t_2 = t \end{cases}$$

where

$$f_W(x; \gamma, \lambda) = \gamma \lambda x^{\gamma-1} e^{-\lambda x^{\gamma}}$$

The *Joint Survivor Function* is given by

$$S_{T_1,T_2}(t_1, t_2) = S_W(t_1; \gamma, \lambda_1) S_W(t_2; \gamma, \lambda_2) S_W(t_1 \vee t_2\ \gamma, \lambda_0), \quad \forall \lambda_0, \lambda_1, \lambda_2, \gamma, t_1, t_2 > 0$$

Initial Model

Marshall Olkin Bivariate Weibull Distribution

Theoretical problems:

- Not straightforward representation of the BWMO distribution using latent variables unlike the bivariate Poisson when modelling the number of goals.

Initial Model

Marshall Olkin Bivariate Weibull Distribution

Theoretical problems:

- Not straightforward representation of the BWMO distribution using latent variables unlike the bivariate Poisson when modelling the number of goals.
- Unclear how λ_0 affects the correlation between T_1 and T_2.

Initial Model

Marshall Olkin Bivariate Weibull Distribution

Theoretical problems:

- Not straightforward representation of the BWMO distribution using latent variables unlike the bivariate Poisson when modelling the number of goals.
- Unclear how λ_0 affects the correlation between T_1 and T_2.
- λ_0 is always positive and therefore the dependence between the goal arrival times is always positive (not realistic).

Initial Model

Marshall Olkin Bivariate Weibull Distribution

Theoretical problems:

- Not straightforward representation of the BWMO distribution using latent variables unlike the bivariate Poisson when modelling the number of goals.
- Unclear how λ_0 affects the correlation between T_1 and T_2.
- λ_0 is always positive and therefore the dependence between the goal arrival times is always positive (not realistic).

Computational Problem:

- The BWMO distribution is not available in the well establised Bayesian platforms (like BUGS or STAN).

Initial Model

Marshall Olkin Bivariate Weibull Distribution

Theoretical problems:

- Not straightforward representation of the BWMO distribution using latent variables unlike the bivariate Poisson when modelling the number of goals.
- Unclear how λ_0 affects the correlation between T_1 and T_2.
- λ_0 is always positive and therefore the dependence between the goal arrival times is always positive (not realistic).

Computational Problem:

- The BWMO distribution is not available in the well establised Bayesian platforms (like BUGS or STAN).

To avoid these problems we assumed that the two arrival scoring times are coming from independent Weibull Distributions truncated at the censoring times of each team.

Final Model

Independent Weibull Model: Formulation

Let t_{i1} and t_{i2} be the goal arrival times (in the sense that was presented above) by home (HT) and away teams (AT) $i = 1, 2, ..., n$. Then the "independent Weibull" model can be expressed by

$$T_{ij} \sim Weibull(\gamma, \lambda_{ij}), \quad j = 1, 2, i = 1, 2, ..., n$$

Final Model

Independent Weibull Model: Formulation

Let t_{i1} and t_{i2} be the goal arrival times (in the sense that was presented above) by home (HT) and away teams (AT) $i = 1, 2, ..., n$. Then the "independent Weibull" model can be expressed by

$$T_{ij} \sim Weibull(\gamma, \lambda_{ij}), \quad j = 1, 2, i = 1, 2, ..., n$$

with

$$\begin{aligned} log\Big(E(T_{i1})\Big) &= \mu + home + a_{HT_i} + d_{AT_i} + ge_{GDescr_i} + re_{GDescr_i} \\ &+ \beta_1 gd1_i + \beta_2(hf_i - 1) + \beta_3 gd2_i + \beta_4 rt_i + \beta_5 gs_i \end{aligned}$$

$$\begin{aligned} log\Big(E(T_{i2})\Big) &= \mu + a_{AT_i} + d_{HT_i} + ge_{GDescr_i} + re_{GDescr_i} \\ &- \beta_1 gd1_i + \beta_2(hf_i - 1) - \beta_3 gd2_i + \beta_4 rt_i + \beta_5 gs_i \end{aligned}$$

with

$$E(T_{ij}) = \lambda_{ij}^{\frac{1}{\gamma}} \Gamma(1 + \frac{1}{\gamma})$$

Offline Covariates

- Team Effects.
- Game Effect.
- Round Effect.

Final Model

Independent Weibull Model: Covariates

Offline Covariates

- Team Effects.
- Game Effect.
- Round Effect.

Online Covariates

- Indicator for one goal difference.
- Different effect for goal difference that is higher than 2.
- Half Time indicator.
- Remaining Time.
- Goal Scored by each even time.

Final Model

Independent Weibull Model: Covariates

Offline Covariates

- Team Effects.
- Game Effect.
- Round Effect.

Online Covariates

- Indicator for one goal difference.
- Different effect for goal difference that is higher than 2.
- Half Time indicator.
- Remaining Time.
- Goal Scored by each even time.

Other Parameters

- Home Effect.
- Intercept.

Final Model

Independent Weibull Model: Prior Distributions

The prior distributions that were assigned to the parameters of our model, are weakly informative and are presented as follows:

$$a_k, d_k \sim Normal(0, 10^{-3})$$

$$\mu, home, ge_G descr_i, gs_G descr_i \sim Normal(0, 10^{-3})$$

The coefficients in our model are also assumed to have a weakly informative prior namely:

$$\beta_j \sim Normal(0, 10^{-3})$$

Finally, since the shape parameter γ is a positive parameter, a Gamma distribution as follows

$$\gamma \sim Gamma(10^{-3}, 10^{-3})$$

In order to make the model identifiable and make comparisons of the ability of each team with an overall level of attacking and defensive abilities we imposed Sum-To-Zero constrains on those parameters. In particular we assumed the following

$$\sum_{k=1}^{K} a_k = 0, \quad \sum_{k=1}^{K} d_k = 0$$

Final Model

Independent Weibull Model: Bayesian Estimation and Model Fitting.

Final Model

Independent Weibull Model: Bayesian Estimation and Model Fitting.

- Use MultiBUGS to fit out model and sample from the required posterior distributions using MCMC.

Final Model

Independent Weibull Model: Bayesian Estimation and Model Fitting.

- Use MultiBUGS to fit out model and sample from the required posterior distributions using MCMC.
- Conduct Gibbs Variable Selection (Dellaportas et al., 2002) to select a final model.

Final Model

Independent Weibull Model: Bayesian Estimation and Model Fitting.

- Use MultiBUGS to fit out model and sample from the required posterior distributions using MCMC.
- Conduct Gibbs Variable Selection (Dellaportas et al., 2002) to select a final model.
- For illustration, make comparisons with the null model.

Final Model

Independent Weibull Model: Bayesian Estimation and Model Fitting.

- Use MultiBUGS to fit out model and sample from the required posterior distributions using MCMC.
- Conduct Gibbs Variable Selection (Dellaportas et al., 2002) to select a final model.
- For illustration, make comparisons with the null model.
- Make the required interpretations.

Final Model

Independent Weibull Model: Results

Parameter	*Mean*	*SD*	*2.5%*	*97.5%*
home	-0.197	0.071	-0.345	-0.050
μ	1.045	0.451	0.370	2.301
γ	1.357	0.057	1.247	1.468
hf	-0.533	0.112	-0.762	-0.314
gd2	-0.110	0.038	-0.179	-0.041
rt	-0.031	0.002	-0.035	-0.026

Model	DIC	Covariates
Null Model	4069.7	None
GVS Model	3826.7	hf + gd2 + rt

Final Model

Independent Weibull Model: Attacking and Defensive abilities' estimates

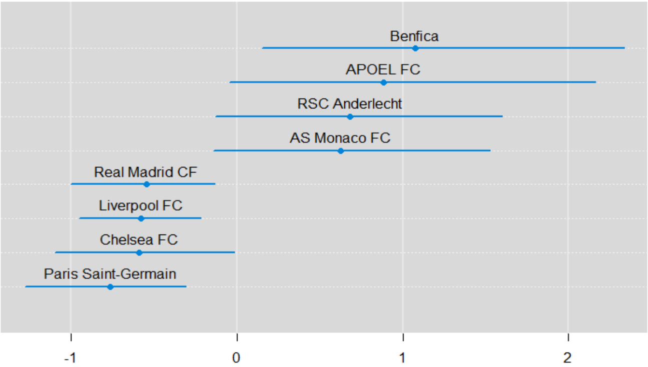

Figure 1: Credible intervals for attacking ability for best and worst teams.

Final Model

Independent Weibull Model: Attacking and Defensive abilities' estimates

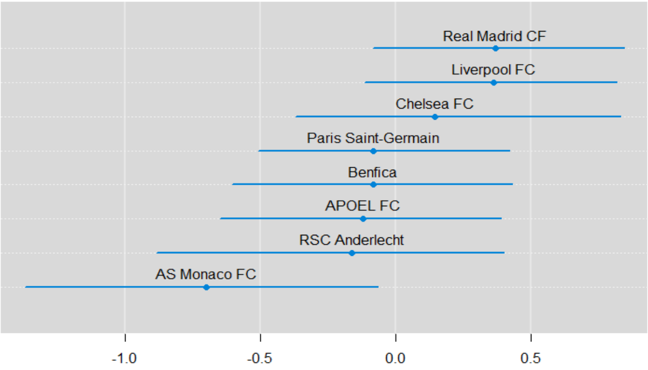

Figure 2: Credible intervals for defensive ability for best and worst teams.

Final Model

Survival Curve for the final game.

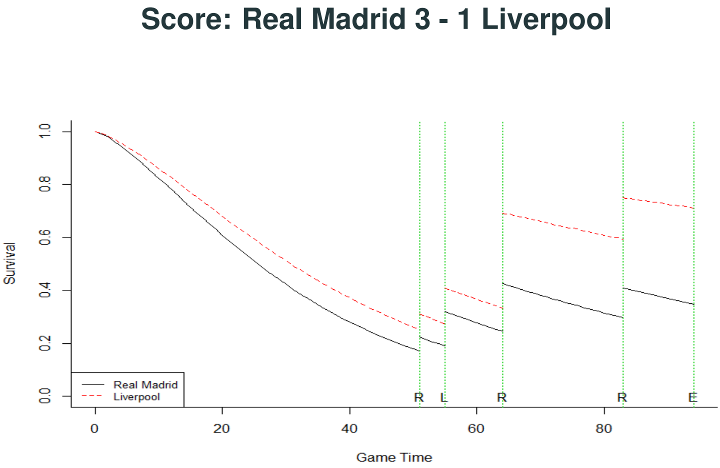

Figure 3: Survival Curves for the Champions League's final game. The vertical green lines represent the goal times.

Further work

- To reset or not to reset?

Further work

- To reset or not to reset?
- Use only the gap times and an indicator for H/A team (logistic approach).

Further work

- To reset or not to reset?
- Use only the gap times and an indicator for H/A team (logistic approach).
- Model the goal arrival times as Recurrent Events for the same subject (game).

Further work

- To reset or not to reset?
- Use only the gap times and an indicator for H/A team (logistic approach).
- Model the goal arrival times as Recurrent Events for the same subject (game).
- Investigate the use of alternative bivariate distributions for modelling allowing for positive and negative dependence between the goal arrival times.

Further work

- To reset or not to reset?
- Use only the gap times and an indicator for H/A team (logistic approach).
- Model the goal arrival times as Recurrent Events for the same subject (game).
- Investigate the use of alternative bivariate distributions for modelling allowing for positive and negative dependence between the goal arrival times.
- Use Copulas to model the dependence.

Further work

- To reset or not to reset?
- Use only the gap times and an indicator for H/A team (logistic approach).
- Model the goal arrival times as Recurrent Events for the same subject (game).
- Investigate the use of alternative bivariate distributions for modelling allowing for positive and negative dependence between the goal arrival times.
- Use Copulas to model the dependence.
- Goodness of Fit.

Further work

- To reset or not to reset?
- Use only the gap times and an indicator for H/A team (logistic approach).
- Model the goal arrival times as Recurrent Events for the same subject (game).
- Investigate the use of alternative bivariate distributions for modelling allowing for positive and negative dependence between the goal arrival times.
- Use Copulas to model the dependence.
- Goodness of Fit.
- Predictions.

Further work

- To reset or not to reset?
- Use only the gap times and an indicator for H/A team (logistic approach).
- Model the goal arrival times as Recurrent Events for the same subject (game).
- Investigate the use of alternative bivariate distributions for modelling allowing for positive and negative dependence between the goal arrival times.
- Use Copulas to model the dependence.
- Goodness of Fit.
- Predictions.


thank you
adidas

References

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12(1):27–36.

Nevo, D. and 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(2):165–177.

Thomas, A. C. (2007). Inter-arrival times of goals in ice hockey. *Journal of Quantitative Analysis in Sports*, 3(3).