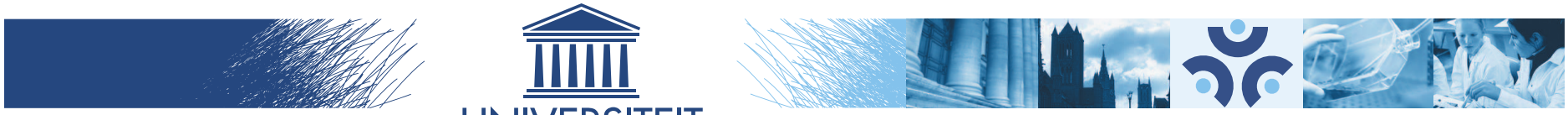

Statistical modelling in soccer

Hans Van Eetvelde

Napels, April 8, 2018

Overview

- Modelling soccer games
  - Directly Modelling soccer outcomes
  - Modelling scores
- Estimating parameters
  - Maximum Likelihood
  - Weighted Maximum Likelihood
- Applications
  - Prediction
  - Ranking

Bradley Terry

First assume that there are only two outcomes: a home win (H) or an away win (A)
We model the chances on these outcomes as follows:

$$P_H = \frac{H}{H + A};$$

$$P_A = \frac{A}{H + A};$$

where H is the strength of the home team and A is the strength of the away team.

Possibility of a draw

Of course, in soccer we have also the possibility of a draw, which we will model as follows:

$$P_H = \frac{p\frac{H}{H\ A} + A}{H + d\,p\frac{H}{H\ A} + A};$$

$$P_D = \frac{d\,p\frac{H\ A}{H\ A}}{H + d\,p\frac{H\ A}{H\ A} + A};$$

$$P_A = \frac{p\frac{A}{H\ A} + A}{H + d\,p\frac{H\ A}{H\ A} + A};$$

Exercise: Assume $\ _H$ is constant. Prove that $P_D(\ _A)$ is maximal for $\ _A = \ _H$.

Adding the Home advantage

A first simple addition to this model is adding a home effect. The strength of the home team is multiplied with h:

$$P_H = \frac{h\,\lambda_H}{h\,\lambda_H + d\sqrt{h\,\lambda_H\,\lambda_A} + \lambda_A};$$

$$P_D = \frac{d\sqrt{h\,\lambda_H\,\lambda_A}}{h\,\lambda_H + d\sqrt{h\,\lambda_H\,\lambda_A} + \lambda_{iA}};$$

$$P_A = \frac{\lambda_A}{h\,\lambda_H + d\sqrt{h\,\lambda_H\,\lambda_A} + \lambda_A};$$

Disadvantage of the Bradley-Terry Model

It does only consider the outcomes of the matches (Home win, Draw, away win), but not the scores.
Winning a game by 5-0 or by 1-0 is not the same!

) We will try to model the scores

Modelling scores

We can indirectly model the outcome of the game by modelling the home and away score. If G_H is the number of goals made by the home team and G_A is the number of goals made by the away team, then we have

$$P_H = P(G_H > G_A)$$
$$P_D = P(G_H = G_A)$$
$$P_A = P(G_H < G_A)$$

Poisson distribution

- How to model a number of goals?
- Most used 'count distribution': Poisson distribution

$$P(G = g) = \frac{^{g}}{g!}e$$

where the parameter is the mean.

- In this case: is the average number of goals scored

Soccer scores

Let G_H be the number of goals made by the home team and G_A the number of goals scored by the away team, then we have for the Poisson model:

$$P(G_H = x) = \frac{{}_H^x}{x!} e^{-{}_H}$$

$$P(G_A = x) = \frac{{}_A^x}{x!} e^{-{}_A}$$

where ${}_H$ is the expected number of goals scored by the home team and ${}_A$ is the expected number of goals scored by the away team.

From scores to outcome

If the variables G_H and G_A are Poisson distributed with parameters $_H$ and $_A$, then the distribution of the di erence $G_H \quad G_A$ is a known distribution, named the 'Skellam' distribution. Using this distribution we can easily compute the chances on every outcome:

$$P_H = P(G_H \quad G_A > 0)$$
$$P_D = P(G_H \quad G_A = 0)$$
$$P_A = P(G_H \quad G_A < 0)$$

Poisson model

Idea from Maher (1982)

$$\lambda_H = c \frac{H}{A}$$

$$\lambda_A = c \frac{A}{H}$$

where λ_H is again the strength of the home team, λ_A is the strength of the away team and c is the expected number of goals for teams with equal strengths.

Adding the home e ect (bis)

Idea from Maher (1982)

$$_H = c \quad h \quad \frac{H}{A}$$

$$_A = c \quad \frac{A}{H}$$

where $_H$ is again the strength of the home team, $_A$ is the strength of the away team and c is the expected number of goals for teams with equal strengths.

Attack and defence strengths

Instead of using one strength parameter per team, we could use separate parameters for a team's attacking strength and a teams defensive strength:

$$_H = c \quad h \quad \frac{H;att}{A;def}$$

$$_A = c \quad \frac{A;att}{H;def}$$

Correlation between home and away sco

Consider again the model with one strength parameter per team

$$_H = c \quad h \quad \frac{H}{A}$$

$$_A = c \quad \frac{A}{H}$$

There is clearly a (negative) correlation between the home and the away score.

Should we allow for other types of correlation? Bivariate Poisson

Bivariate Poisson Model

In the simple Poisson model we have

$$P(G_H = x; G_A = y) = P(G_H = x) \quad P(G_A = y)$$

$$= \frac{{}^{x}_{H}}{x!} e^{\ {}_{H}} \frac{{}^{y}_{A}}{y!} e^{\ {}_{A}}$$

$$= \frac{{}^{x}_{H} \ {}^{y}_{A}}{x!y!} e^{\ ({}_{H} + \ {}_{A})}$$

In the bivariate model, we allow for a positive covariance between G_H and G_A

$$P(G_H = x; G_A = y) = \frac{{}^{x}_{H} \ {}^{y}_{A}}{x!y!} e^{\ ({}_{H} + \ {}_{A} + \ {}_{C})} \sum_{k=0}^{\min(x;y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{{}_{C}}{{}_{H} \ {}_{A}} \right)^{k}$$

Bivariate Poisson Model

- The bivariate Poisson distribution is derived from the simple Poisson distribution as follows: If X_H, X_A en X_C are Poisson distributed with parameters $_H$, $_A$ en $_C$, then $G_H = X_H + X_C$ en $G_A = X_A + X_C$ are bivariate poisson distributed.
- So the number of goals for both teams now consist out of a team-speci c component and a common component.

How to determine $_C$?

We will determine $_H$ and $_A$ as before, for $_C$ we have several options:

1. $_C = b_C =$ constant
2. $_C = b_C \quad b_H$ depending on the home team
3. $_C = b_C \quad b_A$ depending on the away team
4. $_C = b_C \quad b_H \quad b_A$ depending on both teams

Other ways to model soccer goals

- One can use other count distributions instead of the Poisson distribution. (For example the Negative binomial or the Weibull Count Distribution are used)
- Other ways to model the dependency between home and away scores?
- We look for a functionc for which:

$$P(G_H = x; G_H = y) = c(P(G_H = x); P(G_A = y))$$

This kind of functions are called copula (density) functions.

Overview

- Modelling soccer games
  - Directly Modelling soccer outcomes
  - Modelling scores
- Estimating parameters
  - Maximum Likelihood
  - Weighted Maximum Likelihood
- Applications
  - Prediction
  - Ranking

Maximum Likelihood

- When we know the values of the parameters ($_H$, $_A$, h, c,...) we can calculate the chances for the outcome of a game
- but how do we get these parameters?
- We can estimate them by using maximum likelihood!
- We look for the optimal parameters using numerical optimization

Maximum likelihood Bradley Terry

Imagine we have collected the outcome $o_i \in \{H, D, A\}$ of N past games, including T teams

$$L(\theta_1, \ldots, \theta_T, d, h) = \prod_{i=1}^{N} P(O_i = o_i).$$

$$P(O_i = H) = \frac{h\theta_{iH}}{h\theta_{iH} + d\sqrt{h\theta_{iH}\theta_{iA}} + \theta_{iA}},$$

$$P(O_i = D) = \frac{d\sqrt{h\theta_{iH}\theta_{iA}}}{h\theta_{H} + d\sqrt{h\theta_{iH}\theta_{A}} + \theta_{iA}},$$

$$P(O_i = A) = \frac{\theta_{iA}}{h\theta_{iH} + d\sqrt{h\theta_{iH}\theta_{iA}} + \theta_{iA}},$$

Maximum likelihood Poisson

Let now x_{iH} en x_{iA} be the number of goals scored by the home and away team.

For example, if the result of the game is 1-2, $x_{iH} = 1$ and $x_{iA} = 2$.

$$L(\ _1; :::;\ _T; c; h) = \prod_{i=1}^{N} P(G_{iH} = x_{iH})\ \ P(G_{iA} = x_{iA}):$$

$$P(G_{iH} = g) = \frac{{}_{iH}^{g}}{g!} e^{\ {}_{iH}} \qquad P(G_{iA} = g) = \frac{{}_{iA}^{g}}{g!} e^{\ {}_{iA}}$$

$${}_{iH} = c\ \ h\ \ \frac{{}_{iH}}{{}_{iA}} \qquad {}_{iA} = c\ \ \frac{{}_{iA}}{{}_{iH}}$$

Time decay parameter

- How many matches do we consider for our MLE-estimates?
- Is every game equally important?

We introduce a weight such that more recent games have a higher weight in the maximum likelihood

$$w_{time}(t_i) = \left(\frac{1}{2}\right)^{\frac{t_i}{\text{Half Period}}};$$

where t_i is the time (in days) that has passed since the game is played and `Half Period' is a parameter whose value is to be chosen. (see later)

When we use the time weights, our likelihood formulas for the Bradley-Terry and the simple Poisson model change to

$$L = \prod_{i=1}^{N} P(O_i = o_i)^{w_{time}(t_i)};$$

$$L = \prod_{i=1}^{N} (P(G_{iH} = x_{iH})P(G_{iA} = x_{iA}))^{w_{time}(t_i)}$$

Importance weights

We could also use weights to give important games more weights than less important games. For example, if we would consider the national teams, we could use the weights used by the FIFA:

- World Cup: $w_{importance} = 4$
- Confederations Championships (e.g. EC): $w_{importance} = 3$
- Quali cation matches for tournaments above: $w_{importance} = 2.5$
- Friendlies and small tournaments: $w_{importance} = 1$

E.g. for Bradley-Terry the likelihood function changes to:

$$L = \prod_{i=1}^{N} P(O_i = o_i)^{w_{time}(t_i) w_{importance}}.$$

Overview

- Modelling soccer games
  - Directly Modelling soccer outcomes
  - Modelling scores
- Estimating parameters
  - Maximum Likelihood
  - Weighted Maximum Likelihood
- Applications
  - Prediction
  - Ranking

Prediction

- Since our models can produce probabilities for (future) soccer games, it is straightforward that they can be used for prediction.
- We repeatedly estimate our parameters on a number of games in a certain time period and we then predict the outcome of the next game.
- We measure if a model is good in predicting by using a loss function

loss functions

- Rank Probability Score:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{(P(O_i = H) \quad \mathbb{I}(o_i = H))^2 + (P(O_i = A) \quad \mathbb{I}(o_i = A))^2}{2}$$

The best model is the model with the lowest RPS.

Results Premier League

Table: Comparison table for the best performing models of each of the considered classes with respect to the RPS criterion. All of the second season half English Premier League matches in the period between the seasons 2000-2001 and 2016-2017 are considered.

Model Class	**Parameters per team**	**Half Period best model**	**Lowest RPS**
Independent Poisson,	1	HP = 200	0.1979
Bivariate Poisson, c=const.	1	HP = 180	0.1981
Bivariate Poisson, c=const.	2	HP = 240	0.1986
Independent Poisson	2	HP = 240	0.1988
Bradley-Terry	1	HP = 540	0.2020

Ranking

Another application of this modelling, is the making of rankings. Since we have estimated the strengths $_{i}$ for all teams, we can order them by (estimated) strength.

Example ranking national teams

	teams	points		team	points
1	Brazil	4.32	11	England	3.15
2	Spain	4.18	12	Uruguay	2.98
3	Germany	4.13	13	Denmark	2.97
4	Argentina	4.13	14	Croatia	2.95
5	France	3.53	15	Sweden	2.93
6	Belgium	3.50	16	Poland	2.92
7	Colombia	3.42	17	Italy	2.91
8	Chile	3.37	18	Peru	2.86
9	Netherlands	3.37	19	Ecuador	2.61
10	Portugal	3.36	20	Switzerland	2.60
131	Luxembourg	1.12	212	Vatican	0.29