Bivariate copula-based regression for modeling results of football matches

Rasmus Ekman

June 14, 2020

Abstract

This paper considers bivariate models of football match results, where the expected number of goals scored by each team depends on the estimated strength of the individual players. An important consideration is whether there is some kind of dependence between the scores for each team. We use copulas to try to answer that question. Models with Poisson marginals and different parametric copula families are applied to a large data set with tens of thousands of matches and players. It is found that the fitted models are able to forecast out-of-sample match outcomes with an accuracy comparable to bookmakers. In comparison with a model assuming independence, there are indications of a slight advantage in using copulas to model the dependence.

Contents

1 Introduction

Quantitative analysis of sports is an interesting and growing field. Historically, association football, the most popular sport in the world, has arguably lagged behind other sports such as basketball and American football in terms of applications of statistical models. However, in recent years it has started to attract more interest from people approaching it from an analytic point of view. Advanced statistical metrics such as Expected Goals have gained mainstream acceptance, while teams such as Liverpool have successfully applied quantitative modeling as support when making decisions on player transfers involving tens of millions of euro.

When modeling the score of a football match, one has to assume that some amount of randomness is involved. In fact, it has been claimed that the high degree of uncertainty of how a given match will end is a major reason for the popularity of the sport.. However, the skill levels of the teams also have a crucial influence on the outcome. Most successful mathematical models of football are based on some sort of estimation of team strength, offensive and defensive.

One approach towards estimating team strength is to consider it to depend on the quality of its individual players. While tactics, team organization and many other factors are also important, it is probably not a stretch to assume that having the better players is generally the most important advantage one can have in football. Estimating the strength of individual players in team sports can however be difficult. For sports like basketball and ice hockey, an established method to get such estimates is to use so-called plus/minus models. In recent years, such models have started to become popular also for football. However, one can argue that the sport is in many ways poorly suited for them. Large data sets are required to get a good result, and usually the number of parameters to estimate gets very large. Hence, a careful choice of estimation technique is needed to ensure that the parameter estimates do not get drowned by noise resulting from large standard errors in the estimates.

The score of a football match can be considered either as the univariate difference between the number of goals scored by each team, or as the bivariate variable consisting of both these numbers. In the latter case, one has to consider a possible dependence between the marginal variables. It turns out that the dependence has quite subtle properties. Generally, one finds a small negative correlation which is related to teams being unequally good. However, after taking the varying team strengths into account, the conditional dependence is instead slightly positive. Various methods for dealing with the dependence have been suggested. Perhaps the simplest is to use ad-hoc constructions to adjust the probability of match results to be more realistic than those given by a model assuming independence. Other more elaborate methods are to employ bivariate models that allow for dependence, such as so-called bivariate Poisson models, or to use a copula in conjunction with some model for the marginals.

In this master's thesis copulas are used together with an adjusted regularized plus/minus model. The model is applied to forecast football match results and try to find opportunities with positive expected value on the betting market.

2 Modeling football matches

In this section we give some general background on quantitative models of football matches, with examples from the scientific literature. While results of football matches, due to the very low-scoring nature of the sport, tend to be more decided by random chance than most other sports, the relative skill levels of the teams involved also have a crucial impact on the final result. Our approach in this paper is to consider the level of a team to be a function of the skill of its individual players. This requires us to somehow define and estimate an individual player's skill level as a single number rating. The basic idea of the approach we use is simple: a player is good if his team achieves good results when he is playing. This idea is formalized in a so-called adjusted plus/minus model.

2.1 Individual player ratings for football

Creating an accurate one-number measure of the ability of an individual football player is a nontrivial problem with diverse applications ranging from purely recreational to transfer decisions involving hundreds of millions of euros. Ratings can be based on subjective expert opinion, like the ratings used in computer game such as Football Manager and Fifa, or the market value estimates provided by sites as Transfermarkt. They can also be based on more objective quantitative measures. It can be useful to divide performance-based objective rating methods into bottom-up and top-down methods, respectively.

Bottom-up methods are based on registering individual actions performed by a player during a match. The player is then evaluated based on whether these actions are considered to have a beneficial or harmful effect on the player's team's chance of winning the game. This gives a possibility of getting a detailed picture of what exactly a particular player does well, and can be useful for finding players capable of fulfilling a particular tactical role in a team. A drawback with the approach is that it requires detailed tracking data that is usually not freely available to the public. It can also be somewhat hard to know exactly which actions performed by players are important and why. In particular, there is probably a risk of overemphasizing spectacular offensive actions that are easily noticeable, whereas good defensive positioning and marking can be harder too measure.

Top-down methods in contrast are based upon the performance of the team as a whole, which is distributed in some way to the individual players. This allows one to take an agnostic stance towards exactly which actions are beneficial, which gives a less detailed but potentially fairer picture of a player's contribution. However, since football is a sport with few substitutions per match, the same player's tend to play together in many matches. For that reason it can be diffcult to distribute the team performances in an accurate way to the right players. For example, there is a risk of overrating players playing for good teams and underrating good players in less successful teams. A clear advantage compared to bottom-up methods is that much simpler data is required, generally only match results and team lineups.

2.2 Plus/minus models for football and other sports

A particular class of top-down models used for rating individual players in team sports consists of the so-called plus/minus models, on which there have been a surging interest in recent years. A comprehensive review of published models of this kind is provided in [5]. Here we just review briefly some of the key papers and ideas.

The first plus/minus models were invented for ice hockey in the late 1950's. The basic idea is to estimate of a player's ability by comparing the performance of his team when he is playing compared to when he is not. In the most basic form one simply counts the points scored by a player's team when he is playing and subtracts the points scored by the opposing team. The cumulative score for a player over some set of matches is then taken as his plus/minus rating, which is an estimate of his playing strength.

A more advanced form is the so-called adjusted plus/minus model. Here one divides each match into segments where no changes of personnel occur, and then fit a regression model (usually multiple linear regression) to the observations made up by the segments. The model in [13], considered the first published example of an adjusted plus/minus model, uses the equation

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i$$

to model basketball matches. Here Y_i is the difference between home team points per possession and away points per possession in observed segment i, $X_{ij} = 1$ if player j played for the home team in that segment and $X_{ij} = -1$ if he played for the away team. The estimated parameter β_0 models the average home team advantage, while the parameter β_j estimates the ability of player j, and ϵ_i is an error term.

Since the same players tend to often play together in many matches there may be difficulties with separating the contributions of individual players from that of their team mates. In technical regression terms one may run into issues with multicollinearity, as noted in [10]. The solution to that problem suggested there is to include match data from many different seasons. This helps because players often transfer to different teams between seasons, thus separating them from each other.

Another problem, noted in [15], is that in addition to multicollinearity there tends to be an issue with overfitting when estimating the parameters. In part that may be due to the large number of parameters that need to be fitted, since there is at least one parameter for each individual player (two if offensive/defensive contribution is estimated separately). They suggest using ridge regression to deal with these problems, as is standard in modern machine learning and other settings with multicollinearity and overfitting issues. Ridge regression is performed by putting a quadratic penalty term on the size of the parameters, thus shrinking them towards zero. In a linear regression setting it can be interpreted from a Bayesian point of view as putting a Gaussian prior distribution on the parameters.

Although not as well-established as plus/minus models for basketball and ice hockey, these types of models have recently been applied to football. Compared to these sports, one can argue that such models are less suitable for football, as there are few substitutions per match, increasing multicollinearity, as well as few goals per observation, i.e. a sparse output variable. Nevertheless, as this thesis will hopefully demonstrate, a successful application is possible.

2.3 Modeling results of football matches

There are many different ways to model the score of a football match. In plus-minus models, it is common to define the output variable to be the difference between the goals scored by the home team and away team, so that one match has a single observed output variable per match/match segment. However, in this paper we consider the bivariate output variable (H, A) consisting of the number of home goals H and away goals A.

Models for the number of goals scored by a team in a football match have been studied since at least the early 1950's. The most commonly used distribution used to fit such models is the Poisson model, which can be considered as a standard model for this application. While it was noted early on that the Poisson distribution does not tend to provide a very accurate fit to observed match scores (which tend to be closer to a negative binomial distribution, i.e. a mixture of Poisson variables), the inclusion of relevant covariates (modeling differing playing strength levels between teams), improves the fit of the Poisson model significantly. An early example of a model which can be considered as a standard model for football matches can be found in [12]. There, the number of goals scored by the home team and away team are modeled as Poisson variables whose means are conditional on the attacking and defensive strengths of the teams.

The Poisson model can be motivated by considering goals to arrive according to a point process. A strong team, with a high offensive strength parameter, will tend to attack more intensely, leading to a higher probability of scoring goals and thus a higher number of expected goals over the match. If the opposing team is strong defensively, the probability of a goal will on the other hand be decreased accordingly, as they will be capable of mitigating the attacks. The Poisson assumption means that goals will arrive in a uniform manner over the match, independent of other goals. This is unlikely to be more than approximately true, as for example a team being one goal down could be expected to start to attack more aggressively to get even. An example of a point process model of football matches which is not Poisson is found in [1].

2.4 Modeling dependence in football matches

In the previous section we discussed the bivariate variable (H, A) consisting of the number of goals by the home team and by the away team, respectively. A question one has to consider is if H and A can be considered to be independent, or if there is some kind of dependence that has to be taken into account. Given that the teams involved in a match interact with each other, it is certainly reasonable to suspect that some kind of dependence could be involved.

If one takes a sample of real-world match results and simply calculates some kind of dependence measure for the sample, such as a the sample correlation coefficient, ρ, or Kendall's tau, τ, one usually finds a small negative correlation. Indeed, for the sample of 67,949 matches in the data set used in this paper, the sample correlation is $\rho = -0.062$ while $\tau = -0.0374$. The result makes some sense intuitively, because if a team scores a relatively high number of goals in a match, there is an increased probability that it is facing a relatively weaker opposing team, that will in turn tend to score relatively fewer goals. A key assumption behind that reasoning is that the offensive and defensive strength of teams will generally tend to be correlated, i.e. teams that have stronger attacking ability than average are also more likely to have a stronger average defense and vice versa. In [8], it is found that the estimated offensive and defensive strength of teams are indeed correlated in their model. When appylying the model to predict matches, it leads to the predicted outcomes (H, A) being negatively correlated in a way that corresponds closely to observed data, despite an assumption of independent Poisson margins.

However, an interesting question is if there is still some dependence left to model after the relevant covariates (i.e. team strengths) have been taken into account, as described in the previous paragraph. Instead of the negative dependence described there, one usually anticipates some positive dependence that would for example increase the probability of a match ending in a draw.

In [11], the authors study the question of dependence using a bivariate Poisson regression model. In their basic bivariate Poisson model, it is assumed that match results $(H,\ A)$ are distributed as $(P_1 + P_0,\ P_2 + P_0)$, where $P_i \sim Poi(\lambda_i)$, $i = 0, 1, 2$, are independent. According to a well-known elementary result, the marginal distributions are Poisson, with $H \sim Poi(\lambda_1 + \lambda_0)$ and $A \sim Poi(\lambda_2 + \lambda_0)$. The model allows for a positive covariance, since $\mathrm{Cov}(H, A) = \mathrm{Cov}(P_1 + P_0, P_2 + P_0) = \mathrm{Var}(P_0) = \lambda_0$. After fitting their model, it is found by means of a hypothesis test that $\lambda_0 > 0$ with probability very close to 1. Furthermore their bivariate Poisson model achieves a higher log-likelihood and lower AIC and BIC values than a model they fit assuming independence. The difference seems rather slight, but by putting covariates on λ_0 and inflating the diagonal of the distribution (i.e. increasing the probability that $H = A$, meaning a match ending in a draw), they are able to achieve even better model fits.The level of dependence seems to be corresponding to a correlation around 0.1, according to their results.

Another alternative to postulating a specific kind of bivariate model, such as the bivariate Poisson model described above, is to use a copula to model the dependence structure separately from the margins. This approach is taken in [9], in which different copulas are applied to a Poisson regression model in order to forecast match results in FIFA World Cup tournaments. The quality of the forecasts are evaluated based on the basis of different scoring rules, such as Rank Probability Score, RPS. Their copula-based models improve the predictions compared to a model where independence is assumed, but the difference is quite small. Again, small correlations on the order 0.1 are suggested.

3 Copulas and dependence modeling

This section gives some background information on copula theory and specifications of some particular classes of copulas that are employed in this paper. A useful introduction to copulas is provided in [7].

3.1 Definition and Sklar's theorem

A copula is a multivariate cumulative distribution function whose marginal distributions are uniformly distributed on the interval $[0, 1]$. We will mostly consider bivariate copulas in this paper. A function $C\,[0, 1] \times [0, 1] \to [0, 1]$ is a bivariate copula if

- $C(0, v) = C(u, 0) = 0$
- $C(1, v) = v,\ C(u, 1) = u$

- $C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0,\ 0 \leq u_1 \leq u_2 \leq 1,$ $0 \leq v_1 \leq v_2 \leq 1.$

The usefulness of copulas in modeling stochastic dependence is largely due to a result known as Sklar's theorem. The theorem states that every multivariate cumulative distribution function

$$H(x_1, \ldots, x_d) = P(X_1 \leq x_1, \ldots, X_d \leq x_d)$$

can be written as

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$$

where C is a copula and $F_i(x_i) = P(X_i \leq x_i)$ is the marginal cumulative distribution function of the ith marginal variable. Hence, the copula C completely characterizes the dependence between the marginal variables. The theorem further states that C is uniquely defined on

$$\text{Ran}(F_1) \times \ldots \times \text{Ran}(F_d),$$

i.e. the Cartesian product of the ranges of the marginal cumulative distribution functions. Hence the copula C is uniquely determined in the case that all marginal distributions are continuous. However, if some marginal is not continuous, then C is not unique. This unidentifiability issue may or may not be a problem in practice, depending on the application, but it is a first example of many on how copula theory is more complicated in the case of non-continuous marginal distributions.

3.2 Dependence measures

The most commonly used measure of the dependence between two random variable is the Pearson correlation coefficient. For the random variables X and Y it is defined as

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}\, X}\sqrt{\text{Var}\, Y}}.$$

In spite of its popularity, it has certain drawbacks. It is only defined when the variables have finite second moments, it can only measure linear dependence and is not invariant under strictly increasing transforms.

3.2.1 Kendall's tau

A dependence measure with some theoretical advantages over the Pearson correlation coefficient is Kenndal's tau. Let (X_1, Y_1) and (X_2, Y_2) be independent copies of the random variable (X, Y). The probabilistic definition of Kendall's tau for (X, Y) is

$$\tau_{X,Y} = \text{P}((X_1 - X_2)(Y_1 - Y_2) > 0) - \text{P}((X_1 - X_2)(Y_1 - Y_2) < 0).$$

Given a sample $(x_1, y_1), \ldots, (x_n, y_n)$ from (X, Y) the definition suggests the following estimator of $\tau_{X,Y}$

$$\hat{\tau}_{X,Y} = \frac{c-d}{c+d} = \frac{c-d}{\binom{n}{2}},$$

where c is the number of pairs (x_i, y_i), (x_j, y_j) with $i \neq j$ and $(x_i - x_j)(y_i - y_j) > 0$. Similarly d is the number of such pairs where instead $(x_i - x_j)(y_i - y_j) < 0$.

Kendall's tau can also be defined analytically for a copula C as

$$\tau_C = 4 \int_0^1 \int_0^1 C(u,v) dC(u,v) - 1.$$

If the marginal distributions of X and Y are continuous, and the distribution function H of (X, Y) has the copula representation $H(x, y) = C(F(x), F(y))$, one can show that $\tau_{X,Y} = \tau_C$. Hence Kendall's tau is completely determined by the copula and not influenced by the particular marginal distribution. In the case of discrete margins, the value of Kendall's tau will depend on the marginal distributions, and the probabilistic and analytical definitions will not coincide [6].

3.2.2 Spearman's rho

A similar measure to Kendall's tau is Spearman's rho. If (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) are independent copies of (X, Y), it is defined as

$$\rho_{X,Y} = \mathrm{P}\left((X_1 - X_2)(Y_1 - Y_3) > 0\right) - \mathrm{P}\left((X_1 - X_2)(Y_1 - Y_3) < 0\right).$$

The analytical copula definition is

$$\rho_C = 12 \int_0^1 \int_0^1 C(u,v) du dv - 3.$$

If the marginal distributions are continuous, $\rho_{X,Y} = \rho_C$, similar to the case with Kendall's tau.

3.3 Families of copulas

In this section we present some types of copulas that are used in this master's thesis.

3.3.1 Parametric families of copulas

In parametric modeling using copulas, one assumes that the copula belongs to some parametric family C_θ, where $\theta \in \Theta,\ \Theta \subseteq \mathbb{R}$ is the parameter to estimate. Often θ can be interpreted as governing the level of dependence.

3.3.2 Archimedean copulas

A wide and often used class of copulas is the family of Archimedean copulas. A bivariate archimidean copula is characterized by a generator functions Ψ, in terms of which the copula can be written as

$$C(u, v) = \Psi\left(\Psi^{-1}(u) + \Psi^{-1}(v)\right).$$

Necessary conditions that $\Psi : [0, \infty] \to [0, 1]$ has to satisfy for C to be a copula are

1. $\Psi(0) = 1$, $\lim_{x\to\infty} \Psi(x) = 0$
2. Ψ is continuous
3. Ψ is decreasing in $[0, 1]$ and strictly decreasing in $[0, \inf\{x;\ \Psi(x) = 0\}]$.

A sufficient condition in the bivariate case, provided the necessary conditions listed above are fulfilled, is that Ψ is convex.

The table below presents the Archimedean copulas that are employed in this paper.

Table 1: Archimedean copulas

Copula	Definition
Clayton	$C_\theta(u, v) = [\max(u^{-\theta} + v^{-\theta} - 1, 0)]^{-1/\theta}$, $\theta \in [-1, \infty)$, $\theta \neq 0$
Frank	$C_\theta(u, v) = -\frac{1}{\theta}\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$, $\theta \in [-\infty, \infty]$, $\theta \neq 0$
Gumbel	$C_\theta(u, v) = \exp\left(-[(-\log u)^\theta + (-\log v)^\theta]^{1/\theta}\right)$, $\theta \in [1, \infty)$
AMH	$C_\theta(u, v) = \frac{uv}{1-\theta(1-u)(1-v)}$, $\theta \in [-1, 1]$
Joe	$C_\theta(u, v) = 1 - [(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta]^{1/\theta}$, $\theta \in [1, \infty)$
FGM	$C_\theta(u, v) = uv + \theta uv(1-u)(1-v)$, $\theta \in [-1, 1]$

3.3.3 Elliptical copulas

An elliptical copula C is defined in terms of the distribution function of a so-called elliptical distribution. The d-dimensional random vector $\boldsymbol{X}$ is said to have an elliptical distribution if its characteristic function φ has the form

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = e^{i\boldsymbol{t}^T\boldsymbol{\mu}}\Psi\left(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right),$$

where $\boldsymbol{\mu}$ is a vector, $\boldsymbol{\Sigma}$ a non-negative definite matrix and Ψ some function. The definition of an elliptical copula C in terms of the distribution function H of an elliptic distribution is

$$C(\boldsymbol{u}) = H(H_1^{-1}(u_1), \dots, H_d^{-1}(u_d))$$

where $\boldsymbol{u} = (u_1, \ldots, u_d)$ and H_i is the marginal distribution function of the ith margin of H.

In this paper, the two elliptical copulas we use are the bivariate normal copula and the Student's t copula with υ degrees of freedom, where υ is some positive number. The bivariate normal copula is defined as

$$C_\theta(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left(-\frac{x^2 - 2\theta xy + y^2}{2(1-\theta^2)}\right) dxdy,$$

where Φ^{-1} is the quantile function of the standard normal distribution. The bivariate Student's t copula with $\upsilon > 0$ degrees of freedom is defined as

$$C_\theta(u, v) = \int_{-\infty}^{t_\upsilon^{-1}(u)} \int_{-\infty}^{t_\upsilon^{-1}(v)} \frac{1}{2\pi\sqrt{1-\theta^2}} \left(1 + \frac{x^2 - 2\theta xy + y^2}{2(1-\theta^2)}\right)^{-(\upsilon+2)/2} dxdy$$

where t_υ^{-1} is the quantile function of Student's t distribution with υ degrees of freedom. For both types of copulas, $-1 < \theta < 1$.

3.4 Tail dependence

When two random variables exhibit concordance at extreme levels, we say that they have tail dependence. If (X, Y) are random variables, their upper tail dependence coefficient $\lambda_U(X, Y)$ is defined as

$$\lambda_U(X, Y) = \lim_{t \to 1^-} \mathrm{P}(Y > F_Y^{-1}(t) | X > F_X^{-1}(t))$$

and their lower tail dependence coefficient $\lambda_L(X, Y)$ as

$$\lambda_L(X, Y) = \lim_{t \to 0^+} \mathrm{P}(Y \leq F_Y^{-1}(t) | X \leq F_X^{-1}(t)).$$

If the margins are continuous, their tail dependence coefficients can be expressed in terms of their copula $C_{X,Y}$ as

$$\lambda_U(X, Y) = \lim_{t \to 1^-} \frac{1 - 2t - C_{X,Y}(t, t)}{1 - t}$$

and

$$\lambda_L(X, Y) = \lim_{t \to 0^+} \frac{C_{X,Y}(t, t)}{t}.$$

The table below presents tail dependence coefficients for the copulas used in this thesis.

Table 2: Tail dependence coefficients of different copulas

Copula	λ_U	λ_L
Normal	0	0
Student's t_υ	$2t_\upsilon(\frac{\sqrt{\upsilon+1}\sqrt{1-\theta}}{\theta(1+\theta)})$	$2t_\upsilon(\frac{\sqrt{\upsilon+1}\sqrt{1-\theta}}{\theta(1+\theta)})$
Clayton	0	$2^{-1/\theta}$
Frank	0	0
Gumbel	$2-2^{1/\theta}$	0
AMH	0	$= \begin{cases} 0.5 & \text{if } \theta = 1 \\ 0.5 & \text{if } \theta < 0 \\ 0 & \text{otherwise.} \end{cases}$
Joe	$2-2^{1/\theta}$	0
FGM	0	0

3.5 Statistical inference for copulas

In this section, we describe four different methods for estimating parameters of a bivariate distribution, where both parameters δ_1 and δ_2 of the marginal distributions and a parameter θ for the copula need to be estimated. In the following, we assume that we have a sample of N independent bivariate observations $\{\boldsymbol{x_n}\}_{n=1}^N$, $\boldsymbol{x_n} = (x_{1,n}, x_{2,n})$, from the bivariate random variable $\boldsymbol{X} = (X_1, X_2)^T$. We assume that the ith margin of $\boldsymbol{X}$ has distribution function $F_i(\cdot;\delta_i)$. Provided the marginal distributions are continuous, we let $f_i(\cdot;\delta_i)$ denote the probability density function for $i = 1, 2$. If they are discrete, it instead denotes the probability mass function. The distribution function is

$$F_{\boldsymbol{X}}(x, y) = C_\theta\{F_1(x;\ \delta_1),\ F_2(y;\ \delta_2)\}.$$

If the margins are continuous, $\boldsymbol{X}$ has the density function

$$f_{\boldsymbol{X}}(x, y) = c_\theta\{F_1(x;\ \delta_1),\ F_2(y;\ \delta_2)\}f_1(x;\ \delta_1)f_2(y;\ \delta_2)$$

where $c_\theta(u, v) = \frac{\partial}{\partial u}\frac{\partial}{\partial v}C_\theta(u, v)$ is the copula density. If the margins are discrete we instead have the probability mass function

$$\begin{aligned} f_{\boldsymbol{X}}(x, y) &= C_\theta\{F_1(x;\ \delta_1),\ F_2(y;\ \delta_2)\} - C_\theta\{F_1(x-1;\ \delta_1),\ F_2(y;\ \delta_2)\} \\ &- C_\theta\{F_1(x;\ \delta_1),\ F_2(y-1;\ \delta_2)\} + C_\theta\{F_1(x-1;\ \delta_1),\ F_2(y-1;\ \delta_2)\}. \end{aligned}$$

3.5.1 Full maximum likelihood

In full maximum likelihood, one simultaneously estimates all the parameters $\upsilon = (\delta_1, \delta_2, \theta)^T$. The likelihood function is

$$\mathcal{L}(\upsilon; x_1, \ldots, x_N) = \prod_{i=1}^{N} f_{\boldsymbol{X}}(x_{1,i}, x_{2,i})$$

and the log-likelihood is

$$\ell(\upsilon) = \ell(\upsilon; x_1, \ldots, x_N) = \sum_{i=1}^{N} \log f_{\boldsymbol{X}}(x_{1,i}, x_{2,i}).$$

The parameters are estimated as

$$\hat{\upsilon}_{FML} = \arg\max_{\upsilon} \ell(\upsilon).$$

The main drawback with the method is that it is computationally heavy, especially for distributions of high dimension. High dimensionality in that sense is not a problem with the models in this paper, as we only consider bivariate distributions. However, we use covariates with dimensions in the tens of thousands for the marginal parameters δ_1 and δ_2, and full maximum likelihood is thus not a very attractive choice.

3.5.2 Method of moments

In the method of moments method for copula fitting, the parameters of the marginal distributions are fitted using standard method of moments estimators. Hence one equates empirical moments from the observations with theoretical moments that are functions of the parameters δ_1 and δ_2, and then solves for the parameter values.

For the estimation of the copula parameter, we assume that Kendall's tau, τ_K, or Spearman's rho, ρ_S, is a one-to-one function of the copula parameter θ, i.e. $\tau_K = \phi(\theta)$ or $\rho_S = \psi(\theta)$. The estimate of θ is calculated as $\hat{\theta} = \phi^{-1}(\hat{\tau}_K)$ or $\hat{\theta} = \psi^{-1}(\hat{\rho}_K)$ where $\hat{\tau}_K$ and $\hat{\rho}_S$ are sample values of τ_K and ρ_S.

3.5.3 Inference for margins

The inference for margins method is conducted in two steps. The first step is to find parameter estimates for the marginal distributions, $\hat{\delta}_1$ and $\hat{\delta}_2$. Taking these estimates as given, one then estimates the copula parameter θ by maximizing a pseudo log-likelihood function. Hence, the approach can be summarized as

1. Estimate parameters δ_j, $j = 1, 2$, from the marginal distributions

$$\hat{\delta}_j = \arg\max_{\delta_j} \ell_j(\delta_j) = \arg\max_{\delta_j} \sum_{i=1}^{N} \log f_j(x_{j,i};\quad \delta_j).$$

2. Estimate the copula parameter θ as

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \log \hat{f}_{\boldsymbol{X}}(x_{1,i}, x_{2,i}),$$

where

$$\hat{f}_{\boldsymbol{X}}(x,y) = c_\theta\{F_1(x;\ \hat{\delta}_1),\ F_2(y;\ \hat{\delta}_2)\}f_1(x;\ \hat{\delta}_1)f_2(y;\ \hat{\delta}_2)$$

if the margins are continuous, and

$$\begin{aligned}\hat{f}_{\boldsymbol{X}}(x,y) = {} & C_\theta\{F_1(x;\ \hat{\delta}_1),\ F_2(y;\ \hat{\delta}_2)\} - C_\theta\{F_1(x-1;\ \hat{\delta}_1),\ F_2(y;\ \hat{\delta}_2)\} \\ & -C_\theta\{F_1(x;\ \hat{\delta}_1),\ F_2(y-1;\ \hat{\delta}_2)\} + C_\theta\{F_1(x-1;\ \hat{\delta}_1),\ F_2(y-1;\ \hat{\delta}_2)\}.\end{aligned}$$

if they are discrete.

3.5.4 Canonical maximum likelihood

Canonical maximum likelihood is similar to inference for margins, except one uses pseudo data consisting of normalized ranks

$$\hat{u}_i = (\text{rank}(x_{1,i})/(n+1),\ \text{rank}(x_{2,i})/(n+1))$$

instead of

$$(F_1(x_{1,i};\ \hat{\delta}_1),\ F_2(x_{2,i};\ \hat{\delta}_2))$$

as arguments to the pseudo-likelihood function. Another way to see it, is that one uses the empirical distribution function of the margins. An advantage with this approach is that one does not have to assume a specific parametric distribution for the marginals.

3.6 Copulas, discrete marginal distributions and regression

We have noted multiple times in this section that copula theory is in many ways more complicated in the case when at least one of the marginal distributions has a discrete distribution. Concordance measures like Kendall's tau and Spearman's rho will then depend on the marginal distributions and not only on the copula. However, dependence properties will often be inherited from the copula, and the parameter θ can still often be interpreted as a dependence parameter [6].

Since copulas are only determined on $\text{Ran}\,(F_1) \times \ldots \times \text{Ran}\,(F_d)$, there is, at least in theory, an issue with the identifiability of the copula in the case of discrete margins. Even more worrisome, it is shown in [17] that estimation of copula parameters often fails badly in the presence of discrete margins. However, they claim that including covariates in the model helps with identifying the copula parameter. The reason for that may be that the expected means, obtained from the covariates, are continuous.

4 The model

Our model for the final score of a football match is that the number of goals scored by each team is Poisson distributed, with expected mean depending on how strong the teams are. The team strengths are assumed to depend linearly on the individual strengths of the participating players. These individual player strengths are the main parameters that are to be estimated when fitting the model to a data set. We then use a copula to model the dependence between the goals scored by each team.

4.1 Marginal distributions

The model for the marginal distributions is very similiar to the model used in the author's Bachelor's thesis [4], and is inspired by [18]. We assume that the number of goals scored by the home team, Y_H, and by the away team, Y_A, are Poisson distributed with means λ_H and λ_A, respectively. The means satisfy the equations

$$\log(\lambda_H) = \mathbf{x}_H^T \boldsymbol{\beta} \tag{1}$$

and

$$log(\lambda_A) = \mathbf{x}_A^T \boldsymbol{\beta} \tag{2}$$

where $\boldsymbol{\beta}$ is the parameter vector to be estimated and the coefficient vectors $\mathbf{x}_H$ and $\mathbf{x}_A$ depend on which players participate in the match. To explain how these vectors look, we can expand $\log(\lambda_H)$ as

$$\log(\lambda_H) = \beta_0 + \beta_{HOME} + \sum_{i \in H} x_i O_i - \sum_{j \in A} x_j D_j \tag{3}$$

and $\log(\lambda_A)$ as

$$\log(\lambda_A) = \beta_0 + \sum_{j \in A} x_j O_j - \sum_{i \in H} x_i D_i \tag{4}$$

where, for $k = i, j$,

$$O_k = \beta_{1k} - c\,(a_k - 27)^2\,\beta_{AGE} + \sum_{d \in DIV} p_{kd}\beta_d. \tag{5}$$

The condition $i \in H$ is satisfied if player i played for the home team in the match, and $j \in A$ is satisfied if player j played for the away team. x_k is the playing time (minutes played divided by 90) of player k in the match, while O_k is his estimated offensive strength. The offensive strength depends on different parameters as shown in (5).

The parameter β_{1k} is the main part of the estimate of the offensive strength of player k, based on the number of goals scored by his team in matches he has played.

The division parameters β_d, $d \in DIV$, which are communal for all players, estimate the relative strength of different divisions. The set DIV consists of all the represented divisions in the data set, and the coefficients p_{kd} denote the proportion of recorded minutes that player k has played in division d. These division components help to better differentiate between leagues with marked difference in average playing strength.

The parameter β_{AGE}, which is also common for all players, allows a player's strength to vary through time. Here a_i is the age in years of player i in the given match. The constant c is an arbitrary scaling constant. If it has the same sign as β_{AGE}, the player will be assumed to achieve peak performance level at age 27. When he is younger, his performances can reasonably be assumed to be weaker due to insufficient accumulation of experience. On the other hand, when he gets older, his physical capabilities will decline.

In a similar way we have for the defensive strength D_k of player k

$$D_k = \beta_{2k} - c\,(a_k - 27)^2\,\beta_{AGE} + \sum_{d \in DIV} p_{kd}\beta_d, \tag{6}$$

where β_{2k} measure defensive performances.

In summary

$$\boldsymbol{\beta} = (\beta_0,\ \beta_{HOME},\ \beta_{11},\ \beta_{21},\ \ldots\ \beta_{1p},\ \beta_{2p},\ \beta_{d1},\ \ldots\ \beta_{dt},\ \beta_{AGE})^T$$

is the parameter vector to estimate, where p is the number of players and t the number of division. It is easy to derive from (1), (2), (3), (4) (5) and (6) explicit formulas for the coefficient vectors $\mathbf{x}_H^T$ and $\mathbf{x}_A^T$, but they are awkward to write down.

It should be noted that we assume a priori that certain players have offensive strength 0 or defensive strength 0. We classify players as either defenders (including goalkeepers), midfielders or forwards. Offensive strength is not estimated for defenders, while defensive strength is not estimated for forwards. In practice this is achieved by setting the pertinent playing times equal to 0. For example, if player i played is a forward and played for the home team, x_i would be set to his playing time in (3) but equal to 0 in (4).

4.2 Joint distribution and copula

According to Sklar's theorem, the cumulative distribution function H of any bivariate distribution with given marginal distribution functions F and G can be written as

$$H(y_1, y_2) = C(F(y_1), G(y_2)),$$

where C is a copula. We will now apply this to the bivariate variable (Y_H, Y_A), where Y_H and Y_A are as presented in the previous section. In the following we assume that C is from some parametric family of copulas, and write it as $C = C_\theta$ where θ is the copula parameter to estimate. We also assume that C_θ does not depend on the covariate vectors $\mathbf{x}_H^T$ and $\mathbf{x}_A^T$.

The marginal distribution functions are

$$F_{Y_H}(y_1|\boldsymbol{\beta}, \mathbf{x}_H) = e^{-e^{\mathbf{x}_H^T\boldsymbol{\beta}}} \sum_{k=0}^{\lfloor y_1 \rfloor} \frac{(e^{\mathbf{x}_H^T\boldsymbol{\beta}})^k}{k!}$$

and

$$F_{Y_A}(y_2|\boldsymbol{\beta}, \mathbf{x}_A) = e^{-e^{\mathbf{x}_A^T\boldsymbol{\beta}}} \sum_{k=0}^{\lfloor y_2 \rfloor} \frac{(e^{\mathbf{x}_A^T\boldsymbol{\beta}})^k}{k!}.$$

The distribution function H_θ of (Y_H, Y_A) is

$$H_\theta(y_1, y_2) = C_\theta(F_{Y_H}(y_1|\boldsymbol{\beta}, \mathbf{x}_H),\ F_{Y_A}(y_2|\boldsymbol{\beta}, \mathbf{x}_A)),$$

and the probability mass function $h_\theta(y_1, y_2)$, which gives the probability of a specific result $(Y_H, Y_A) = (y_1, y_2)$, is

$$\begin{aligned} h_\theta(y_1, y_2) = {} & C_\theta(F_{Y_H}(y_1),\ F_{Y_A}(y_2)) - C_\theta(F_{Y_H}(y_1 - 1),\ F_{Y_A}(y_2)) \\ & - C_\theta(F_{Y_H}(y_1),\ F_{Y_A}(y_2 - 1)) + C_\theta(F_{Y_H}(y_1 - 1),\ F_{Y_A}(y_2 - 1)), \end{aligned} \tag{7}$$

where the conditioning on $\boldsymbol{\beta}$, $\mathbf{x}_H$ and $\mathbf{x}_A$ has been dropped from the notation for convenience.

4.2.1 Calculating probabilities of match outcomes

Often we are not interested in the exact result of a match, but rather the coarser outcomes of a home win, a draw, or an away win. The probability of a home win can be calculated as

$$\mathrm{P}(Y_H > Y_A) = \sum_{y_1=1}^{\infty} \sum_{y_2=0}^{y_1-1} h_\theta(y_1, y_2). \tag{8}$$

In a practical setting, the series has to be truncated. The number of terms that has to be included depends on the expected number of goals scored and the desired accuracy. In general, a relatively low upper bound on the number of goals, e.g. between 10 and 25, should be enough in most settings with somewhat evenly matched teams. The probability of an away win or a draw can in the same way be written as

$$\mathrm{P}(Y_H < Y_A) = \sum_{y_2=1}^{\infty} \sum_{y_1=0}^{y_2-1} h_\theta(y_1, y_2) \tag{9}$$

and

$$\mathrm{P}(Y_H = Y_A) = \sum_{y=0}^{\infty} h_\theta(y, y). \tag{10}$$

4.3 Estimation and regularization

Assume that we have a set of n observed match results

$$(y_{11}, y_{12}),\ (y_{21}, y_{22}),\ \dots\ (y_{n1}, y_{n2}),$$

where y_{i1} and y_{i2} are the number of goals scored by the home team and away team respectively, in match i, and that we also have corresponding coefficient vectors $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ for each match. The log-likelihood of the model is

$$\ell(\boldsymbol{\beta}, \theta) = \sum_{i=1}^{n} \log h_\theta(y_1, y_2),$$

where

$$\begin{aligned} h_\theta(y_{i1}, y_{i2}) = {} & C_\theta(F_1(y_{i1}),\ F_2(y_{i2})) - C_\theta(F_1(y_{i1}-1),\ F_2(y_{i2})) \\ & - C_\theta(F_1(y_{i1}),\ F_2(y_{i2}-1)) + C_\theta(F_1(y_{i1}-1),\ F_2(y_{i2}-1)), \end{aligned} \tag{11}$$

with

$$F_j(y_{ij}) = e^{-e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}} \sum_{k=0}^{\lfloor y_{ij} \rfloor} \frac{(e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}})^k}{k!}$$

for $j = 1, 2$. In principle, parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\theta}$ can now be found by full maximum likelihood estimation as

$$\hat{\boldsymbol{\beta}}, \hat{\theta} = \arg\max_{\boldsymbol{\beta}, \theta} \ell(\boldsymbol{\beta}, \theta).$$

In practice, the numerical computations required to find the maximum might be infeasible. For the data set used in this paper, $\boldsymbol{\beta}$ has dimension in the tens of thousands, making the computations very heavy, especially for the elliptic copulas (normal and Student's t) which have distribution functions that are computationally heavy in themselves. Instead we use inference for margins to estimate the marginal parameters and copula parameters separately.

4.3.1 Fitting the marginal model

When fitting the marginal distributions, we first note that we can fit both distributions (i.e. the one for home goals and the one for away goals) simultaneously since they are both Poisson distributed with the same covariates and parameter vector. With the notation $\mathbf{x}_{2*i-1}^T = \mathbf{x}_{i1}^T$, $\mathbf{x}_{2*i}^T = \mathbf{x}_{i2}^T$, $y_{2*i-1} = y_{i1}$, $y_{2*i-1} = y_{i2}$, for $i = 1, 2, \dots, n$ we have the likelihood function

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{2n} e^{-e^{\mathbf{x}_i^T\boldsymbol{\beta}}} \frac{(e^{\mathbf{x}_i^T\boldsymbol{\beta}})^{y_i}}{y_i!}.$$

By taking logarithms and simplifying, we get the log-likelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{2n} y_i \mathbf{x}_i^T \boldsymbol{\beta} - e^{\mathbf{x}_i^T \boldsymbol{\beta}}.$$

Again, numerical optimization will be needed to find $\hat{\boldsymbol{\beta}}$, however in this case standard computational packages can be used. The package *glmnet* for the programming language R can very efficiently find an estimate through the regularized optimization

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} -\frac{1}{2n}\ell(\boldsymbol{\beta}) + \lambda\left((1-\alpha)\sum_{i=1}^{p-1}\beta_i^2/2 + \alpha\sum_{i=1}^{p-1}|\beta_i|\right),$$

where the vector $(\beta_1, \ldots, \beta_{p-1})$ contains the components of $\boldsymbol{\beta}$ that are to be penalized. Generally the intercept β_0 is excluded from the penalization, and other parameters can be excluded as well. More generally, different parameters can be penalized using different weights, but we only use the weights 1 for full penalization or 0 for no penalization in this paper. The penalty term, which is weighted by the tuning parameter λ, penalizes large values of the parameters. The size of λ determines how heavily deviations from 0 are suppressed, and is subject to a bias-variance trade-off. If λ is very large, the parameter estimates are pushed close to 0 giving low variance but high bias. If, on the other hand, it is chosen to be very small, then bias is minimized but variance in the estimates will tend to be high and overfitting the parameters might become an issue. The value of α, which is chosen between 0 and 1, determines how much the penalization will come from a L_2 penalty term and a L_1 penalty term, respectively. Using a linear combination of L_2 and L_1 penalization in this way is known as elastic net regularization. Pure L_2 regularization is known as ridge regression or Tikhanov regularization, while pure L_1 regression is known as the LASSO.

4.3.2 Fitting the copula

Given the estimated parameter vector $\hat{\boldsymbol{\beta}}$, we define two pseudo-samples as

$$(\hat{u}_{i1},\ \hat{u}_{2i}) = (F_1(y_{1i};\ \hat{\boldsymbol{\beta}}),\ F_2(y_{2i};\ \hat{\boldsymbol{\beta}}))$$

and

$$(\hat{v}_{i1},\ \hat{v}_{2i}) = (F_1(y_{1i}-1;\ \hat{\boldsymbol{\beta}}),\ F_2(y_{2i}-1;\ \hat{\boldsymbol{\beta}}))$$

where

$$F_j(y_{ij};\hat{\boldsymbol{\beta}}) = e^{-e^{\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}}}\sum_{k=0}^{y_{ij}}\frac{(e^{\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}})^k}{k!},$$

$j = 1, 2$. We then fit the copula parameter θ as

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{2n} \log\left(C_\theta(\hat{u}_{i1}, \hat{u}_{2i}) - C_\theta(\hat{v}_{i1}, \hat{u}_{2i}) - C_\theta(\hat{u}_{i1}, \hat{v}_{2i}) + C_\theta(\hat{v}_{i1}, \hat{v}_{2i})\right).$$

The approach is essentially inference for margins, except that the marginal parameter vector $\boldsymbol{\beta}$ is fitted using an elastic net approach rather than standard maximum likelihood estimation.

5 The data set

Our data set contains 67,949 matches, including matches played by clubs as well as national teams, in the years 2012-2019. The club matches include matches played in 31 different league divisions, the UEFA Champions League (including qualification matches), and UEFA Europa League. The national team matches are from World Cups and Euro Cups (including qualification matches) and from UEFA Nations League. The data recorded for each match is the final result and the playing time of all players that participated in the match.

The number of players in the data set who have at least one minute played recorded, and for whom parameters are estimated, is 34,869. For each player, their birth year and playing position is included in the data set. The birth year is used for calculating each players age in every given game, to fit the age component as in (5). The positional data is used to classify players as either defenders, midfielders or forwards. As mentioned in 4.1, defenders are a priori assumed to have offensive strength equal to 0 while forwards in the same way are assumed to have defensive strength 0. Both offensive and defensive strength is thus estimated only for midfielders. A player's position is defined as his assigned position on the Transfermarkt, where the detailed positions ('left winger', 'defensive midfielder', etc.) were classified as either defender, midfielder or forward according to following scheme:

Table 3: Position classification

Model position	Transfermarkt positions
Defender	Goalkeeper, centre back, sweeper
Midfielder	Left/right back, defensive/central/right/left midfield
Forward	Attacking midfield, left/right winger, second striker, centre forward

6 Results

6.1 Parameter estimates and interpretations

The model was fitted with $\alpha = 0$, i.e. pure ridge regression, since trials including LASSO penalties indicated significantly worse prediction quality. The tuning parameter $\lambda = 1.9788$ was used for the penalization term.

The parameters β_0, β_{HOME} and β_{AGE} were excluded from the penalization. Excluding the intercept β_0 is completely standard. There are many good reasons for excluding β_{HOME}. It shouldn't be prone to overfitting, and penalizing it would give it a downward bias with potentially large impact on forecasts derived from the model. Excluding β_{AGE} from the penalization gave it a larger size, which seemed to give a more realistic expected development curve for players. When penalized its effect was very minor.

The two parameters with the highest influence on the overall number of goals predicted by the model are the intercept β_0 and the home advantage parameter β_{HOME}. The fitted values are $\hat{\beta}_0 = 0.1128$ and $\hat{\beta}_{HOME} = 0.2727$. Hence, if the offensive strength of each team exactly matches the defensive strength of the other, the home team can be expected to score $e^{0.1128+0.2727} \approx 1.47$ goals and the away team $e^{0.1128} \approx 1.12$ goals. As comparison, the observed means in the data set are 1.52 goals for the home team and 1.15 for the away team.

6.1.1 Estimated division parameters

The table below presents division parameters for the European first division leagues included in the data set. Recall that the division parameters are part of the final player ratings and model different average player strengths in different leagues, as described in section 4.1. As comparison, the Uefa rankings from 2019 are included. The leagues known as the big five (England, Spain, Italy, Germany and France) have the highest parameters by far, which seems correct. In part this is due to these nations being the only ones from which we also include second divisions in the data set. The established first division teams generally tend to defeat teams promoted from the second division, whose player's division components are more weighted towards the second division, which drives up the estimate of the first division component. However, the high estimates are probably mostly an effect of teams from these nations doing well in the Champions League and the Europa League. Also included are estimated mean market values of the clubs in each division, collected from Transfermarkt. The market value of a club is here defined as the sum of the market value of all its players.

Table 4: Division parameter estimates, European first divisions

Nation	$\hat{\beta}_d$	Uefa rank (2019)	Mean market value per club (m€)
England	0.0234	2	393.62
Spain	0.0211	1	254.08
Italy	0.0189	3	223.83
Germany	0.0188	4	231.22
France	0.0136	5	158.27
Turkey	0.00437	10	31.03
Russia	0.00432	6	57.69
Switzerland	0.00120	17	23.72
Belgium	0.000415	8	45.60
Portugal	0.000309	7	50.78
The Netherlands	-0.000678	11	52.06
Austria	-0.00153	12	20.27
Czech Republic	-0.00316	13	11.50
Denmark	-0.00331	16	13.30
Ukraine	-0.00367	9	24.13
Sweden	-0.00473	22	8.02
Norway	-0.00503	23	6.75
Poland	-0.00830	25	10.10
Hungary	-0.00918	33	8.04
Croatia	-0.0107	15	18.94
Scotland	-0.0108	20	17.64
Finland	-0.0125	38	2.80
Greece	-0.0128	14	17.63
Iceland	-0.0186	39	1.65

The corresponding parameters for the European second divisions can be seen in the table below.

Table 5: Division parameter estimates, European second divisions

Nation	β_d	Mean market value per club (m€)
Germany	-0.00422	20.42
England	-0.00449	47.23
Spain	-0.00486	17.22
France	-0.0111	14.29
Italy	-0.0122	15.46

The data set also contains data from the first leagues of two American nations, USA and Brazil. Since teams from these nations are not competing much against teams from other countries in the data set, and are thus somewhat isolated, these division parameter estimates are probably a bit less reliable.

Table 6: Division parameter estimates, American first divisions

Nation	β_d	Mean market value per club (m€)
Brazil	0.00990	44.31
USA	0.00939	23.30

6.1.2 Some player ratings

The most interesting parameters to consider are of course these related to the estimated ratings of individual players. Since there are tens of thousand such parameters, it is only possible to present small subsets. The table below shows the best 15 players for each position, restricted to players who have at least 10,000 minutes recorded (the condition is satisfied by 4182 players). The rating is defined as their estimated peak rating, i.e. their estimated rating at age 27 so that the age component in (5) is equal to 0 (the division components are however taken into account for these presented ratings). For some of the younger players on the lists, the rating is a forecast of their future level, while for others it is based on performances after the age of 27 as well. Some of the players, such as Arjen Robben and Philipp Lahm, are nowadays retired.

The list include many expected names, such as Kylian Mbappé and Neymar, the two most expensive transfers of all time, and the two players generally considered to be the best in the world during the past 10-15 years, Cristiano Ronaldo and Lionel Messi. Many others are also widely considered to be or have been world-class players. However, the lists also include some quite obscure names, such as José Giménez, Paulo Miranda, Emir Dilaver, Óliver Torres and Lior Refaelov, who are all very unlikely to make lists of top ranked players in the world based on consensus opinion.

Table 7: Top ranked players with at least 10,000 minutes recorded in data set, by position

	Defenders		Midfielders		Forwards	
	Player	D_i	Player	$O_i + D_i$	Player	O_i
1	Giorgio Chiellini	0.1247	Philipp Lahm	0.2361	Arjen Robben	0.1729
2	Keylor Navas	0.1197	Javi Martínez	0.2315	Kylian Mbappé	0.1725
3	Andrea Barzagli	0.1148	Fernandinho	0.2238	Neymar	0.1575
4	Medhi Benatia	0.1148	Rafinha	0.2211	James Rodríguez	0.1434
5	Jan Oblak	0.1140	Óliver Torres	0.2189	Luis Suárez	0.1430
6	Alisson Becker	0.1135	Sami Khedira	0.2172	Karim Benzema	0.1396
7	José Giménez	0.1125	Benjamin Mendy	0.2170	Cristiano Ronaldo	0.1346
8	Vincent Kompany	0.1050	Mousa Dembélé	0.2109	Mohamed Salah	0.1259
9	Felipe	0.1037	Adrien Rabiot	0.2065	Ángel Di María	0.1247
10	Diego Godín	0.1027	David Alaba	0.2059	Lionel Messi	0.1234
11	Ederson	0.1007	Thiago Alcântara	0.2057	Kevin De Bruyne	0.1224
12	Paulo Miranda	0.0973	Marco Verratti	0.2056	Sadio Mané	0.1210
13	Manuel Neuer	0.0968	Ivan Rakitić	0.2018	Julian Brandt	0.1208
14	Nicolás Otamendi	0.0951	Thomas Meunier	0.2014	Donny Van De Beek	0.1154
15	Emir Dilaver	0.0927	Daniel Carvajal	0.1978	Lior Refaelov	0.1152

6.1.3 Estimated copula parameters

The table below presents the estimated parameter for each copula family we use in this paper, after making the inferences presented in section 4.3.2, together with the maximized log-likelihood and the values of the Akaike information criterion (AIC). AIC is a well-known estimator of out-of-sample prediction error derived from information theory. It is defined as

$$\text{AIC} = 2k - 2\ell(\hat{\theta})$$

where k is the number of parameters and $\ell(\hat{\theta})$ is the maximized log-likelihood. We have 48605 fitted parameters for the marginal distributions, and one additional parameter for each copula except the Student's t copula which has two (the second being the degrees of freedom υ, which we set to $\upsilon = 5$ as it was found to be the integer value providing the best fit). The t copula has the lowest AIC value, indicating it may be the most appropriate copula to use.

It is interesting to note that the best fits, as measured by AIC, is obtained from the Student's t-, Gumbel- and Joe copulas. As indicated by table 2, these are the only copulas out of the ones considered that are able to model an upper tail dependence. Hence a reasonable hypothesis may be that dependence of that kind can be useful when modeling football matches.

Table 8: Estimated copula parameters

Copula	θ	$\ell(\hat{\theta})$	k	AIC
Independence copula	-	-182591.66	48605	462393.3129
Normal	0.0558	-182543.18	48606	462298.3636
Student's t_5	0.0600	**-182308.63**	48607	**461831.2199**
Clayton	0.0865	-182514.44	48606	462240.8722
Frank	0.319	-182535.23	48606	462282.4619
Gumbel	1.0482	-182490.76	48606	462193.5297
AMH	0.1549	-182534.66	48606	462281.3202
Joe	1.0694	-182496.74	48606	462205.4763
FGM	0.145	-182540.17	48606	462292.3314

6.2 Residual analysis of fitted model

An important method for assessing the overall fit of a regression model is residual analysis. A graphical analysis of model residuals can be used to identify discrepancies between the model and data.

A basic form of residual is the Pearson residual, which is defined as $r_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}_i^2}$, where y_i is observation number i, while $\hat{\mu}_i$ is its estimated expected value and $\hat{\sigma}_i^2$ its estimated variance. In the case of a normal linear regression model, the Pearson residuals will follow the standard normal distribution if the model is correct. Hence checking if they do provides a diagnostic of whether the model is correctly specified. The check can be made by means of a normal Q-Q plot.

If the model is not normal, the Pearson residual will generally not follow a normal distribution, and the method can not be used. However, if the model is continuous one can easily obtain standard normal residuals by first transforming the observations to a $U[0,1]$ random number via its probability integral transform (the cumulative distribution function of the fitted model) and then inputting the result to the quantile function of the standard normal distribution.

If the model is not continuous, the method does not work because the probability integral transform will not produce $U[0,1]$ numbers. In fact, an additional source of randomness is needed to transform the discrete observations to a continuous distribution. For a Poisson distribution, one can show that the numbers

$$r_{q,i}^* = \Phi^{-1}(F(y_i - 1, \hat{\lambda}_i) + u_i f(y_i, \hat{\lambda}_i)),$$

where the u_i are independent random variables uniform over $[0,1]$, are standard normal [14]. Here $F(y_i - 1, \hat{\lambda}_i)$ is the cumulative distribution function of a Poisson variable with rate $\hat{\lambda}_i$, evaluated in y_i, while $f(y_i, \hat{\lambda}_i)$ is the corresponding probability mass function. These numbers are called randomized normalized quantile residuals, and can be used for model validation. The plots below present a Q-Q plot and a histogram of these residuals for our fitted model. The plots generally seem quite reasonable, although the Q-Q plot suggests a left-skew. This might be related to teams scoring exactly 0 goals more often than a Poisson distribution would suggest. Indeed, zero-inflated Poisson distributions, which put higher probability on a 0 outcome than a regular Poisson distribution, have been suggested as appropriate for modeling football match scores, eg. in [11].

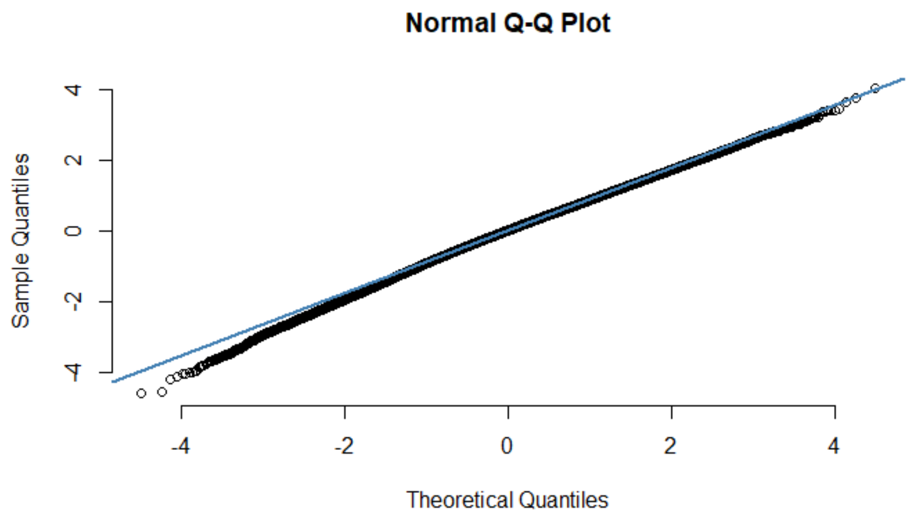

Figure 1: Normal Q-Q plot of randomized normalized quantile residuals

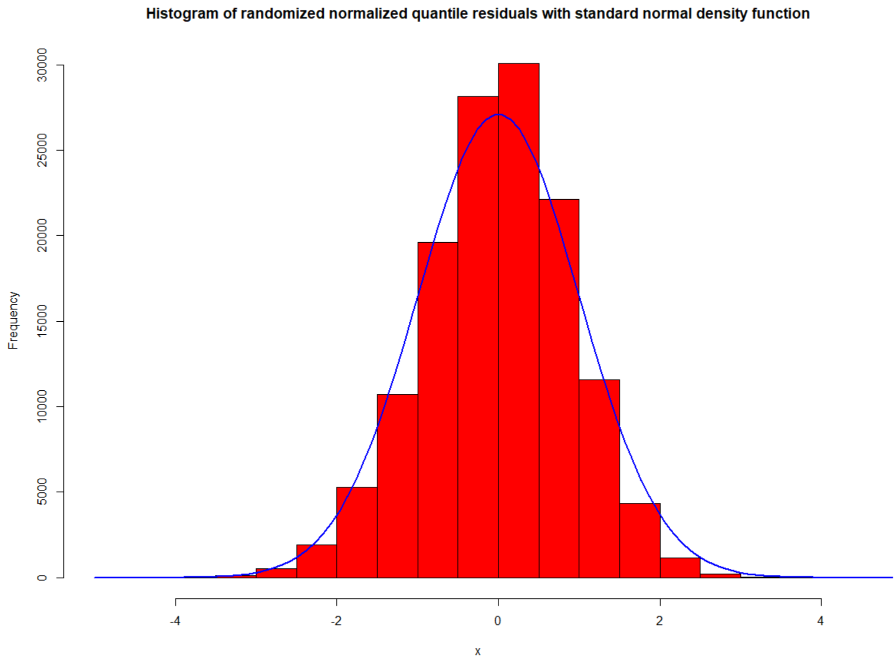

Figure 2: Histogram of randomized normalized quantile residuals with standard normal probability density function

7 Forecasting match results

Perhaps the most effective way to calibrate and test a model of this kind is to evaluate its ability to forecast results in out-of-sample football matches. Since the data set contains matches up to and including 2019, it is relevant and interesting to use it to forecast match results in 2020. We have collected starting lineups and results of 416 matches played in the 'big 5' leagues - the first divisions of England, Spain, Germany, Italy and France. The forecasts of our model are based on the estimated strength of the players in the starting lineups. For players with less than 2500 minutes recorded, we assume a playing strength of 0. The estimated strength of players with fewer minutes recorded is simply too volatile.

We make two different types of forecasts of the matches. The first type is to try to forecast if the matches will end in a home win, draw or away win. The probabilities of different outcomes are calculated according to (8), (9) and (10), truncated at 25 goals. We call this type of forecasts 1X2 forecasts. The second type of forecast we make is whether more or less than 2.5 goals will be scored in the match. We call that type O/U 2.5 forecasts. The probability of under 2.5 goals is of course easily calculated as the sum of the probabilites of a 0-0, 0-1, 0-2, 1-0, 1-1 or 2-0 result, and the probability of over 2.5 goals as its complement.

7.1 Measuring forecast quality

We use four different scoring rules to evaluate the quality of our forecasts. Below we present the definition of each scoring rule for a single match. The overall value is defined as the mean over all considered matches.

The first is the so-called Rank Probability Score (RPS), a well-established measure of prediction quality for football matches. The RPS value of a single problem instance is defined as

$$RPS_i = \frac{1}{r-1} \sum_{i=1}^{r-1} \left(\sum_{j=1}^{i} (p_i - e_i) \right)^2 ,$$

where r is the number of possible outcomes, p_i the forecasted probability of outcome i, and $e_i = 1$ if i is the observed outcome but $e_i = 0$ otherwise. For the 1X2 forecasts we have $r = 3$, while $r = 2$ for the O/U 2.5 forecasts. When $r = 2$, the RPS is also known as the Brier Score. The lower the RPS, the better the forecasts are. In the case of perfect predictions, i.e. when all the forecast probability is put on the observed outcome, RPS is equal to 0. RPS is quite a popular scoring rule for evaluating 1X2 forecasts, largely due to [3]. The RPS is a non-local scoring rule if $r \geq 3$ in that its value depends on the whole distribution of the forecast, and not just the probability of the observed outcome.

The second is the multinomial likelihood, which is defined as the probability of the observed outcome, as provided by the forecast.

The third is the classification rate, which is equal to 1 if the observed outcome is the highest-probability outcome according to the forecast and zero otherwise.

The fourth is the ignorance score, which is defined as $-\log_2(p_l)$, where p_l is the probability of the observed outcome. Similar to RPS, a lower value is better than a higher in this case. The ignorance score is proposed as an appropriate scoring rule for 1X2 forecasts in [19], which challenges the arguments for RPS put forward in [3].

7.2 Bookmaker's odds as benchmarks for forecast quality

The natural benchmark against which to test the quality of a model which forecasts results of football matches is forecasts derived from bookmaker's odds. Especially for high-profile, high-liquidity leagues such as the top 5 leagues of Europe we consider here, the betting market can be expected to be quite efficient in the sense that the odds reflect the true underlying probability distribution rather well.

Assuming one has collected odds from bookmakers (both historical 1X2 odds and O/U 2.5 odds are easily available on various websites.), the problem is then how to convert them to probabilities. We describe how it works for

1X2 forecasts. Let O_H, O_D and O_A be the bookmaker's odds for a home win, a draw, and an away win, respectively. As for odds in general, the way to convert them to probabilities is to take their multiplicative inverses. However, it will turn out that the sum of the inverted odds is not 1, but instead slightly larger. This is because the bookmakers have a margin on their provided odds. Let M denote the margin, so that

$$1 + M = \frac{1}{O_H} + \frac{1}{O_D} + \frac{1}{O_A} \tag{12}$$

where $M > 0$. Some different approaches to deal with the margin are described in [2].

The most obvious way to remove the effect of the margin is to divide the inverted odds by $\frac{1}{1+M}$, so that one gets the estimated probabilities $P_{HU} = \frac{1}{O_H(1+M)}$, $P_{DU} = \frac{1}{O_D(1+M)}$ and $P_{AU} = \frac{1}{O_A(1+M)}$ of the three different outcomes. The probabilities will then obviously sum to 1. This is probably the most common way to do it and is completely legitimate. However, it is not necessarily the best. The method assumes that the bookmaker has the same margin on each outcome, and that assumption is not necessarily accurate.

A more reasonable assumption is that the margin put on each outcome is proportional to the 'fair odds' of the outcome. Equivalently, it is assumed to be inversely proportional to its probability. The assumption is related to the so-called favorite-longshot bias, an observed phenomenon where bettors are willing to take unfavorable odds for low-probability outcomes, due either to risk preferences or to simply overestimating the probability of the rare event [16] According to this assumption, we can write the inverted odds as

$$\frac{1}{O_l} = P_{lP}\left(1 + \frac{c}{P_{lP}}\right) \tag{13}$$

for each outcome $l \in \{H, D, A\}$, where P_{lP} is the sought implied probability of the outcome, and c is a constant. The value of c can be determined as follows. Putting the values (13) into equation (12) gives $1 + M = 1 + 3c$, hence $c = \frac{M}{3}$. It is then seen from (13) that $P_{HP} = \frac{1}{O_H} - \frac{M}{3}$, $P_{DP} = \frac{1}{O_D} - \frac{M}{3}$, and $P_{AP} = \frac{1}{O_A} - \frac{M}{3}$. Hence, instead of dividing with a common number, as in case with the same margin on all outcomes, we subtract another common number.

7.3 Evaluation of forecasts

We evaluate the different copulas, based on the fitted copula parameters presented in table 8, by applying the RPS, multinomial likelihood, classification rate and ignorance score to obtain 1X2 and O/U 2.5 forecasts The odds used for deriving odds-based forecasts were collected from the website *Oddsportal.com*, and are aggregated odds from many bookmakers.

7.3.1 1X2 forecasts

The table below shows the performance of the different copulas, when used to provide 1X2 forecasts for the considered matches. Compared to the model assuming independence, there is a clear tendency towards better forecasts with the copula-based models, although the differences are quite small. Especially the Student's t copula with 5 degrees of freedom seems to work well. The odds-based predictions clearly outperform the model-based predictions overall. The 'naive' forecast which is provided for comparison always places the probabilities 0.45, 0.25 and 0.3 on a home win, a draw and an away win respectively.

In [9], the predictive power of models for predicting football matches based on different copulas are evaluated in a similar way. They also find a small benefit in using copulas rather than assuming independence.

Table 9: Forecasting performance of different models, 1X2

Model	RPS	Likelihood	Classification rate	Ignorance
Odds	0.20041	0.41683	0.50962	1.42442
Naive	0.22762	0.35373	0.44471	1.54511
Independence	0.20491	0.39714	0.50721	1.44544
Normal	0.20478	0.39774	0.50721	1.44387
Student's t_5	**0.20473**	0.39806	**0.50962**	1.44348
Clayton	0.20476	0.39706	0.50721	1.44357
Frank	0.20477	0.39758	0.50721	1.44380
Gumbel	0.20474	0.39799	0.50721	**1.44342**
AMH	0.20477	0.39756	0.50721	1.44378
Joe	0.20476	**0.39811**	0.50721	1.44376
FGM	0.20478	0.39754	0.50721	1.44393

7.3.2 O/U 2.5 forecasts

The table below presents the performance of the models when applied to provide O/U 2.5 forecasts. In contrast with the 1X2 forecasts, there does not seem to be an advantage at all to use copulas for these forecasts. The forecasts are very competitive compared to the ones obtained from bookmaker's odds, beating in on all scoring rules except the multinomial likelihood. Since the model is fitted to model the number of goals scored rather than 1X2 outcomes, it might not be so surprising that it performs relatively better on this type of forecasting.

Table 10: Forecasting performance of different models, O/U 2.5

Model	Brier score	Likelihood	Classification rate	Ignorance
Odds	0.11908	0.52476	0.59759	0.96509
Independence	**0.11868**	0.52098	0.59759	**0.96259**
Normal	0.11878	0.52043	**0.60000**	0.96319
Student's t_5	0.11869	**0.52118**	0.59759	0.96262
Clayton	0.11870	0.52057	**0.60000**	0.96274
Frank	0.11879	0.52033	0.59759	0.96322
Gumbel	0.11887	0.52029	0.59759	0.96369
AMH	0.11878	0.52034	0.59518	0.96318
Joe	0.11889	0.52040	0.59759	0.96380
FGM	0.11878	0.52038	0.59518	0.96317

7.4 Using the model to find profitable betting opportunities

While the forecast accuracy, as measured by e.g. RPS, is interesting in itself, it can be argued that the most interesting thing is often not the forecast probabilities themselves but rather functions of them, corresponding to real-world payoffs. In particular, it is interesting to see if the model can be used to find profitable betting opportunities in matches where the model's forecasts do not agree with the odds provided by bookmakers. To test this, we use the model to place fictional bets on historical matches. Our approach towards investigating this is again similar to investigations in [9]. We make fictional bets on both the 1X2 market and the O/U 2.5 market.

For deciding whether to place a 1X2 be on a match, we consider the odds O_H, O_D and O_A and the model-derived probabilities P_H, P_D and P_A for the three outcomes. The expected return R_l, $l \in \{H, D, A\}$ from betting on outcome l is

$$E[R_l] = P_l * O_l - 1$$

according to the model. For each match we place a bet on the outcome with highest expectation if the expected return is higher than a specified threhold $\epsilon > 0$, i.e. if

$$\max_{l \in \{H,D,A\}} E[R_l] > \epsilon.$$

The approach for finding bets on the O/U 2.5 market is completely analogous.

For simplicity we use a constant bet size of 1 unit in each case instead of using a more elaborate sizing method such as e.g. Kelly betting.

7.4.1 Betting on the 1X2 market

The table below presents the result of applying the method of betting described above to placing 1X2 bets on the same set of matches we evaluated our forecasts on. The bets are generally quite successful, however given the small sample size the results have to be interpreted a bit cautiously. The model is quite prone to bet on high odds, and hence the result is heavily influenced by whether or not a few low-probability outcomes occur or not.

Table 11: Betting performance of different models, 1X2

	$\epsilon = 0.15$		$\epsilon = 0.30$		$\epsilon = 0.45$		$\epsilon = 0.60$	
Model	#Bets	Profit	#Bets	Profit	#Bets	Profit	#Bets	Profit
Independence	242	1.28	157	14.20	90	23.08	51	10.42
Normal	227	-0.72	143	20.15	84	29.08	44	17.42
Student's t_5	197	**7.27**	120	**23.97**	65	15.34	36	**19.53**
Clayton	220	-6.10	139	19.64	80	27.15	43	18.42
Frank	226	-4.26	142	21.15	83	**30.08**	44	17.42
Gumbel	218	-4.10	136	17.76	79	28.15	43	18.42
AMH	226	-4.26	142	21.15	83	**30.08**	44	17.42
Joe	221	-3.4	138	20.64	82	25.15	44	17.42
FGM	227	-0.72	142	21.15	84	29.08	44	17.42

7.4.2 Betting on the O/U 2.5 market

The bets on the O/U 2.5 markets are mostly fairly close to break-even. Here, the odds are rarely very high on either outcome, so the variability in outcomes is a bit lower. The O/U forecasts performed well compared the odds-based forecasts when measured by scoring rules, but as the table indicates that may not be enough to guarantee profitable bets, since the bookmaker's margin implies that the edge must be sufficiently large.

Table 12: Betting performance of different models, O/U 2.5

	$\epsilon = 0.15$		$\epsilon = 0.30$		$\epsilon = 0.45$		$\epsilon = 0.60$	
Model	#Bets	Profit	#Bets	Profit	#Bets	Profit	#Bets	Profit
Independence	172	3.47	102	-4.05	69	-8.87	49	**-1.15**
Normal	169	2.11	101	-3.05	69	-8.87	49	**-1.15**
Student's t_5	173	4.57	103	-2.82	70	**-7.65**	49	**-1.15**
Clayton	174	1.47	103	-2.24	69	-8.87	49	**-1.15**
Frank	171	2.37	101	-3.05	68	-7.87	49	**-1.15**
Gumbel	162	**6.81**	100	**-2.05**	67	-9.12	47	-2.15
AMH	171	2.37	101	-3.05	68	-7.87	49	**-1.15**
Joe	158	6.12	98	-3.04	67	-9.12	46	-3.97
FGM	171	2.37	101	-3.05	69	-8.87	49	**-1.15**

8 Conclusions

This master's thesis presents a bivariate model for football matches based on individual player ratings obtained from an adjusted plus/minus model. Thus, one can say that it derives from two separate traditions in football modeling that are otherwise rarely combined. Bivariate models, often based on Poisson distributions, date back several decades but generally estimate team strengths directly at the team level rather than at player level. On the other hand, models for estimating individual player strength, which can generally be seen as variants of plus/minus models, are more recent. In general, they tend to consider the univariate difference between goals scored by each team as the output rather than the bivariate number of goals for each team. The bivariate approach has some advantages in the flexibility of how it can be applied. For example, 1X2 forecasts of the kind used in this thesis are easily derived from a bivariate model, whereas deriving them from a univariate model is significantly more involved. The difference is even clearer when considering O/U forecasts, which are again easily derived from a bivariate model, while it's not really clear at all how they would be derived from a univariate model. Whether univariate or bivariate output is ultimately more suitable for a plus/minus model does not seem clear, but both approaches are clearly viable.

A bivariate model requires considering a possible dependence. These considerations are a major part of this thesis. The copula-based approach we take has some advantages in flexibility compared to using a bivariate model based on common additive errors, such as exemplified by the so-called bivariate Poisson distribution.

Like some similar investigations, we find that using a copula provides a better fit and more accurate out-of-sample forecasts compared to assuming independence, but the difference is very slight. An interesting observation is that copulas with upper tail dependence generally seem to fit quite well. It could be interesting to investigate closer the importance of upper tail dependence in modeling scores of football matches.

References

[1] Boshnakov, G., Kharrat, T. & McHale, I. (2016) *A bivariate weibull count model for forecasting association football scores.*

[2] Buchdal, J. (2017?) *Using the wisdom of the crowd to find value in a football match betting market.*

[3] Constantinou, A. & Fenton, N. (2012) *Solving the Problem of Inadequate Scoring Rules for Assessing Probabilistic Football Forecast Models.*

[4] Ekman, R. (2019) *Forecasting football match results using individual player ratings: a Swedish case study.*

[5] Hvattum, L. (2019) *A comprehensive review of plus-minus ratings for evaluating individual players in team sports.*

[6] Genest, C. & Neslehova, J. (2007) *A primer on copulas for count data.*

[7] Genest, C. & Favre, A. (2007) *Everything you always wanted to know about copula modeling but were afraid to ask.*

[8] Groll, A., Schauberger, G. & Tutz, G. (2015) *Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014.*

[9] Groll, A., van der Wurp, H., Kneib, T., Marra, G. & Radice, R. (2019) *Generalised Joint Regression for Count Data with a Focus on Modelling Football Matches.*

[10] Ilardi, S. & Barzilai, A. (2008) *Adjusted plus-minus ratings: new and improved for 2007-2008.*
`http://www.82games.com/ilardi2.htm`

[11] Karlis, D. & Ntzoufras, I. (2003) *Analysis of sports data by using bivariate Poisson models.*

[12] Maher, M. (1982) *Modelling association football scores.*

[13] Rosenbaum, D. (2004) *Measuring how NBA players help their teams win.*
`http://www.82games.com/comm30.htm`

[14] Sadeghpour, A. (2016) *Empirical investigation of randomized quantile residuals for diagnosis of non-normal regression models.*

[15] Sill, J. (2010) *Improved NBA adjusted +/− using regularization and out-of-sample testing.*

[16] Snowberg, E. & Wolfers, J. (2010) *Explaining the favorite-longshot bias: is it risk-love or misperceptions?*

[17] Trivedi, P. & Zimmer, D. (2017) *A note on identification of bivariate copulas for discrete count data.*

[18] Vilain, J. & Kolkovsky, R. (2016) *Estimating individual productivity in football.*
http://econ.sciences-po.fr/sites/default/files/file/jbvilain.pdf

[19] Wheatcroft, E. (2019) *Evaluating probabilistic forecasts of football matches: The case against the Ranked Probability Score.*