Fitting Gamma Ranking Models to Soccer Data: Luce's Choice Axiom and Thurstone's Law in Practice

Abstract:

This thesis aims to construct a ranking system measuring the latent variable football ability with random utility methodologies. We consider discriminal processes that are based on gamma distributions. We model football match outcomes as normal, logistic and Laplace Thurstone processes and relate these to Luce´s choice axiom. Numerical optimization of the likelihoods computed through a general linear model produces parameter estimates. Thurstone´s Case V model, the normal model, is closest to the real ranking based on ranking correlation measured by Kendall´s tau. Dawkins´ model, the Laplace model, has the highest value for goodness of fit, measured by the Akaike Information Criterion, due to its ability to model fat tails and a sharp peak simultaneously. Luce's model, the logistic distribution, overestimates the thickness of the tails and is too flat at the peak. Thurstone´s Case V model cannot cope with the excess kurtosis.

Author: Yugesh Raghoenath (333768)
Supervised by: Dr. A.J. Koning

Erasmus School of Economics
Erasmus University

February 2017

1. Introduction

In recent years statistical analysis has become increasingly important in sports. In a market that expands and becomes more competitive year after year, an extra edge might come from such analyses. Trainers, players, media, bookmakers and gamblers use information, insights and numerous statistics for very different purposes. Specific examples are scouting future opponents and players or evaluating past performances. Financial applications include estimating realistic odds as a bookmaker. To do so, it is key to assess the strengths and weaknesses of teams and players adequately.

Within a competition, players and teams are comparable on many statistics as matches are played amongst each other. Still, translating the individual statistics to a combined measure of strength or football ability, if you will, is less trivial; winning matches against stronger opposing teams is a better indicator of high football ability than winning matches against weaker opponents. Moreover, losing multiple games in a row with minimal differences and winning one game with a major difference can result in a positive goal deficit. With a system that estimates aggregate team football abilities correctly, bookmakers quote their odds more efficiently.

The goal of this research is to create a ranking of teams that allows for a realistic measure of the football ability. We model the team's latent football ability pre-match as random variables and the matches as paired comparisons of these random variables. In this context the team with the highest football ability before the match has the highest probability of winning the match. This is no fixed value because the difference between random football abilities is again a random variable. To create an aggregate team football ability model, we adopt random utility methodologies in a paired comparison setting.

Random utility models are applicable to a range of problems in modelling discrete data. Applications include measuring chess players skill level, see (Henery, 1992), or measuring the strength of some latent stimuli, such as facial attractiveness, see (Bauml, 1994). The literature regarding ranking or ordering random objects according to preferences is extensive. Analyzing and Modelling Rank Data (Marden, 1996) is one of many books aiming to give a comprehensive overview of important ranking models.

Random utility models consist out of a deterministic part and a random part, the error. The distribution of the error obviously determines the distribution of the random utility. We model the random football abilities, X_i, as gamma random variables. Say we are to rank k teams, then the times until each team reaches a certain

number of points has a gamma distribution. We consider three different classes of gamma ranking models based on different values for the shape parameter: the class of models when the shape parameter goes to zero, the class of models when the shape parameter is equal to one and the class of models when the shape parameter goes to infinity. The error distribution is different for these three cases. These three cases correspond to the exponential distribution in the first case, the Gumbel distribution in the second case and the normal distribution in the third case. To facilitate linear models with non-normally distributed errors we work with general linear models (GLM).

We proceed to model the outcome of soccer matches as discriminal processes as described by the law of comparative judgment in (Thurstone, 1927). A discriminal process is any process in which comparisons are made between pairs of a collection of entities with respect to magnitudes of some attribute. The distribution of differences in football ability, $F(X_i - X_j)$ is thus what concerns the Thurstone models we consider. F determines the distribution of possible outcomes of football matches. Thurstone's original model, see (Thurstone, 1928), elaborates on the class of Thurstone models where F is the normal distribution. The gamma ranking models we use imply three distinct classes of Thurstone models; a Thurstone model with exponential, Gumbel and normally distributed football abilities, X_i. Evidently, this results in different probability distributions F for each of the Thurstone models.

Precisely these three classes of Thurstone models are analyzed in (Yelott, 1977), where the relationship between Thurstone's law of comparative judgment, the double exponential distribution and Luce´s choice axiom (Luce, 1959) is investigated. Luce's choice axiom is a set of axiomatic foundations. Selection according to Luce´s Choice Axiom is said to have "independence of irrelevant alternatives" (IIA). This means that the probability of selecting one item over another from a pool of many items is not affected by the alternatives in that particular pool. Moreover, that particular probability is independent of the presence or absence of other items in said pool. To be more precise, the ratio of quantified preferences of these two items remains the same, while the absolute quantifications of these preferences may differ after adding or removing items to the pool of alternatives. Section 3.2.3 shows all axioms before going into the practical applications of the model. Additionally, Appendix C contains notes on the proof of 'Luce's Lemma 3', independence from irrelevant alternatives.
According to (Yelott, 1977), Thurstone models with exponential random variables are equivalent to Dawkins' Threshold model for paired comparisons (Dawkins, 1969). In addition, in (Yelott, 1977), derivations show that every Thurstone model with the Laplace distribution as its difference distribution is equivalent to Dawkins' model for paired comparisons. Finally, according to (Yelott, 1977) assumptions underlying Dawkins' model are

too strong to be generally true and it seems fair to attribute the success of Dawkins' model for paired comparisons to the fact that it happens to be a Thurstone model in this special case.
In (Block & Marschak, 1960) as well as in the proof by Marley and Holman in (Luce & Suppes, 1965) it is showed that exponential discriminal Thurstone processes are equivalent to the logistic distribution. Therefore it abides by Luce´s choice axiom. Our third model, Thurstone´s original model, abides by the choice axiom as well according to (Yelott, 1977).

To model these processes, it is appropriate to draw gamma random variables from the same family but with a different location shift, see (Stern, 1990). Therefore, we limit our models to cases in which X_i and X_j have the same distribution, while the general model by Thurstone does not require such a restriction. For the exponential model, we model the outcome of football matches as Laplace random variables with a common scale equal to one; we mentioned earlier this model is equivalent to Dawkins' model for paired comparisons. For the Gumbel model, we apply Luce´s model for paired comparisons; this is the model discussed in (Bradley & Terry, 1952). Finally, for Thurstone's normal model, we assume that the continuous preferences are uncorrelated and have common mean; this model is known as Thurstone´s Case V model (Thurstone, 1928).

In modeling the deterministic part we need to take external settings into account. Soccer matches are matches between two teams, so we need to work in a paired comparison setting instead of in a setting with complete experiments, as is the case for horse races for example. The literature on paired comparisons is extensive and used in a wide array of research topics. In (Wickelmaier & Choisel, 2007) pairwise evaluations of sounds are analyzed through a standard Bradley-Terry model, while we mentioned earlier that (Bauml, 1994) presents applications involving facial attractiveness. Thurstone's original model is applied to analyze subjective health outcomes, see (Maydeu-Olivares & Bockenholt, 2008), and to rate the skill level of chess players, see (Henery, 1992). A detailed overview of the extensions to these two models in a sports context is given in (Cattelan, 2012).
Evidently, we need to adapt these random utility models to a football match setting. Earlier research shows us several things: Firstly, in (Agresti, 2002) the Bradley-Terry model is derived with a home advantage because this turns out to be an important factor empirically. In typical football settings we need to allow for draws. In (Davidson, 1970) the Bradley-Terry model which allows for draws is derived and in (Henery, 1992) Thurstone's original model with a home parameter is derived. We do not add the extension that allows for draws, since we should be able to create a sufficient ranking system without allowing for ties. We believe this is the case, because the empirical probability that two teams perform exactly the same on numerous metrics such as, the number of goals scored, the number of goals conceded and the number of points acquired, is extremely slim.

This reasoning especially holds when we consider that a football season has more than 200 matches; to our knowledge, we historically never found such an event. We take this as sufficient advice that we can exclude the possibility of equal latent football abilities in our model.
We use data on the final scores rather than data on the number of points every team acquired against each opponent. By not analyzing the number of points received for a certain match instead we prevent that a lot of information on the difference (or there lack of) is omitted. We work in a regression setting to facilitate our models. With numerical optimization procedures we optimize the appropriate likelihood functions to find parameter estimates.

As mentioned we aim to construct a ranking system that adequately assess aggregated team football abilities. A whole different approach to ranking is minimizing some distance based metric, see (Mallows, 1957). Distance-based ranking models choose the rankings in such a way that the optimal ranking minimizes the distance *d* with respect to some arbitrary metric on the set of all possible rankings. In (Diaconis, 1988) the most popular distance based metrics are considered, such as the following: Kendall's tau (Kendall, 1948), Spearman's rho (Spearman, 1904), Spearman's footrule (Spearman, 1904) the Hamming metric (Hamming, 1950), and Cayley's metric (Cayley, 1859). In order to objectively judge our models we adopt two of those distance based metrics to quantify the distance between our model and the real ranking. We adopt Kendall's tau and Spearman's rho as distance based ranking evaluation criterion. Our main criteria is Kendall's tau. This correlation coefficient recently gained more interest because it is used in the subsection of finance, risk management; more specific it is used to describe dependences using copulas. As secondary distance based metric we use Spearman's rho. Spearman's rho is nothing else than Pearson's linear correlation for rankings. This means testing for independence of rankings is straightforward.
To compare non-nested models we measure their goodness of fit with the Akaike Information Criterion, see (Akaike, 1973). The Akaike Information Criterion (AIC) is used to measure the goodness of fit, while punishing the less parsimonious models for their lack of simplicity.

Our main research question becomes: "*Among the models by Bradley-Terry, Thurstone´s Case V model and the model by Dawkins, which random utility regression model fits soccer data best in terms of minimizing the distance to real rankings according to the Kendall's tau metric?"* We investigate whether the widely used Bradley-Terry model is indeed most fit to model football data. Theoretically, estimates should be similar: A Thurstone model with logistic distribution is equivalent to Luce's Choice axiom. Moreover, the difference distribution of the exponential distribution and the double exponential distribution explicitly show similarities.

Furthermore we look into the possible addition of a home advantage effect and how adding this parameter changes our model in terms of uncertainty and bias.

The data we use to answer this question consists out of all matches played during the season 2013/2014 in the highest professional Dutch soccer league, the Dutch Eredivisie. This competition can be viewed as a balanced design in the sense that every team plays each opponent exactly twice. It is important to note that every team only plays each team once at home (and thus only once away).

This research is a valuable extension to the existing literature in the sense that the comparison of these three gamma ranking models has not been performed in a football setting, to our knowledge. Moreover we present a complete framework to model the latent football abilities, which is easily extendable with for example other external effects, such as team experience or team synergy.

We find that Thurstone's Case V model outperforms the model by Dawkins' and Bradley-Terry in terms of mimicking the real ranking measured in ranking correlation by Kendall´s tau. Moreover all three models are capable of measuring the latent football ability adequately. This is underlined by the fact that they have a significant amount of rank correlation with the real ranking as well as among themselves, measured by Kendall´s tau and Spearman´s rho. This agrees with derivations regarding the similarity between Dawkins' model and the Bradley-Terry model, see (Yelott, 1977) and earlier findings in (Luce, 1959).
The goodness of fit, measured by AIC, suggests Dawkins´ model, the laplace model, is the best fit because of its ability to model fat tails as well as a sharp peak. Thurstone's Case V model, the normal model, lacks the first of these abilities and the model by Bradley and Terry, the logistic model, lacks the second of these abilities. We estimate a positive home effect which is showed to be an important extension to model the latent football abilities.
Further research could investigate how the skill of the individual team players and the number of matches they have played together, or synergy if you will, affect the team's skill pre-game. Other extensions include adding parameters for a team's experience or even dependence among consecutive games. In (Cattelan, 2009) it is shown how to deal with the dependence among the performances of a team. To incorporate these effects one should collect (or construct) the appropriate data and reformulate the likelihoods we specify in Section 3.3. The distributions we consider as difference distributions have in common that they are all symmetrical. This essentially means that positive deviations and negative deviations from the point estimate of the match outcome imply the same probability if the absolute deviation is the same. A straight forward way to extend this research is by evaluating the performance of asymmetric distributions.

The proceeding sections of this thesis are structured in the following way. In Section 2 we cover aspects of our data. Section 3 extensively elaborates on the theory on GLM, gamma models and our methodology regarding creating and evaluating ranking systems. In Section 4 we show parameter estimates for the football abilities and evaluation metrics for our models. Finally, Section 5 contains the answer to our research question in the summary and some possible ways to follow up this research.

Table of content

2. Data

The data in this research contains all matches played during the season 2013/2014 in the highest professional Dutch soccer league. In this league 18 teams compete in a competition in which every team plays each opponent twice, once at home and once in an away match. This means we have a balanced design; we have data on 306 (18 times 17) matches. More specifically, we have data on the final score of every match. We construct Table 1 with the number of goals scored against every opponent:

Table 1 - Matrix with goals scored against each opponent

	ADO Den Haag	AZ	Ajax	Groningen	FC Twente	FC Utrecht	Feyenoord	Go Ahead Eagles	Heracles Almelo	NAC Breda	NEC	PEC Zwolle	PSV	RKC Waalwijk	Roda JC Kerkrade	SC Cambuur	SC Hereveen	Vitesse
ADO Den Haag	0	2	2	4	4	4	5	4	0	3	2	2	2	2	1	5	1	2
AZ	3	0	3	3	2	1	3	4	4	3	3	4	2	6	4	3	3	3
Ajax	7	6	0	3	4	4	4	7	4	4	5	3	1	2	5	3	6	1
Groningen	2	2	2	0	1	2	0	4	6	4	9	1	4	5	5	2	3	5
FC Twente	3	4	1	6	0	7	6	3	6	7	7	3	4	1	5	4	3	2
FC Utrecht	4	3	1	1	1	0	2	2	4	6	4	2	1	4	3	2	3	3
Feyenoord	6	3	2	3	3	6	0	7	3	4	8	4	5	2	6	7	4	3
Go Ahead Eagles	4	2	0	3	2	3	2	0	3	2	5	4	2	6	4	0	1	2
Heracles Almelo	4	2	1	1	1	2	3	2	0	2	3	2	2	6	5	1	5	3
NAC Breda	2	3	0	3	2	4	2	6	4	0	2	1	2	1	7	1	0	3
NEC	4	4	2	3	4	2	4	4	2	2	0	4	0	5	5	2	4	3
PEC Zwolle	7	1	2	0	3	3	2	3	4	2	8	0	2	2	3	3	1	1
PSV	5	2	4	2	5	6	1	6	3	3	7	3	0	2	4	2	1	4
RKC Waalwijk	3	1	0	2	1	6	1	3	2	4	3	2	3	0	2	4	2	5
Roda JC Kerkrade	4	4	1	4	1	4	1	2	2	1	6	1	3	2	0	2	5	1
SC Cambuur	1	1	2	5	1	3	1	2	3	0	3	3	1	5	1	0	4	4
SC Hereveen	4	9	3	7	1	4	1	5	4	2	3	5	4	8	5	3	0	4
Vitesse	1	1	2	3	1	4	2	5	5	5	4	5	7	5	4	6	5	0

Please note that Table 1 consists out of the cumulative sum of the final scores between two teams. The observation in (2,4) represents the cumulative number of goals AZ scored in two matches against Groningen, which is equal to 3.

In classical paired comparison settings, it is common to construct the input data in a different sense; namely that every win, every draw and every loss has a similar weight. A win might lead to 2 point, a draw might lead to 1 point and a loss to 0 points. By doing so possible information about the difference in strength might be

omitted; a larger difference in goals typically indicates a larger difference in football ability. We circumvent this by working with the number of goals rather than the number of points a team received against opponents.

We investigate the addition of a home parameter which should capture the latent effect of a team being stronger at home matches. To back up this claim we calculate the percentage of matches won by the home and away team. Table 2 shows we empirically find a home advantage that is in line with earlier findings in (Agresti, 2002):

Table 2 - Home advantage

	Home	Away
% Matches won	39.54	33.01

There is an increase of approximately 6.5% in terms of matches won by a team playing at home.

Please note we cannot use Table 1 for further analyses as this Table does not discriminate among matches played at home or away. We turned to our raw dataset with the matches in order to perform our further analysis.

Table 3 shows the final ranking which we use to compare our models to in terms of ranking correlation.

Table 3 - End ranking Eredivisie 2013/2014

Eredivisie	Played	w	d	l	Points	Goals		
Ajax	34	20	11	3	71	+69	-28	(+41)
Feyenoord	34	20	7	7	67	+76	-40	(+36)
FC Twente	34	17	12	5	63	+72	-37	(+35)
PSV	34	18	5	11	59	+60	-45	(+15)
SC Hereveen	34	16	9	9	57	+72	-51	(+21)
Vitesse	34	15	10	9	55	+65	-49	(+16)
Groningen	34	14	9	11	51	+57	-53	(+4)
AZ	34	13	8	13	47	+54	-50	(+4)
ADO Den Haag	34	12	7	15	43	+45	-64	(-19)
FC Utrecht	34	11	8	15	41	+46	-65	(-19)
PEC Zwolle	34	9	13	12	40	+47	-49	(-2)
SC Cambuur	34	10	9	15	39	+40	-50	(-10)
Go Ahead Eagles	34	10	8	16	38	+45	-69	(-24)
Heracles Almelo	34	10	7	17	37	+45	-59	(-14)
NAC Breda	34	8	11	15	35	+43	-54	(-11)
RKC Waalwijk	34	7	11	16	32	+44	-64	(-20)
NEC	34	5	15	14	30	+54	-82	(-28)
Roda JC Kerkrade	34	7	8	19	29	+44	-69	(-25)

3. Methodology

The methodology to arrive at parameter estimates for the football abilities consists out of three main parts: A generalized linear model (GLM), the gamma ranking models and numerical optimization. The GLM framework allows us to model the football abilities X_i as a random variable with gamma distribution while maintaining a regression setting. We discuss the GLM framework in Section 3.1. In Section 3.2.1 we first introduce some notation in order to derive the distribution of the different discriminal Thurstone processes and link them to Thurstone's Case V model, Luce's model and Dawkins' model in Sections 3.2.2 – 3.2.4 respectively. We do so for different values of the shape parameter of the gamma distribution and link these to the GLM framework through the appropriate link functions. Section 3.3 contains notes on the estimation process; we numerically optimize the likelihood functions based on the GLM models. Finally, in Section 3.4 we discuss the evaluation of the gamma ranking models with ranking correlation measures, likelihood ratio tests and Akaike information criterion (AIC) values.

3.1 Generalized Linear Models

In our research we formulate the gamma ranking models as a linear regression. A typical linear regression fixes the error distribution of response variables as a normal distribution. A Generalized Linear Model (GLM) generalizes that assumption by allowing the linear model to be related to the response variable via a link function. Moreover, a GLM allows the variance of each measurement to be a function of the predicted value of said measurement.

For a GLM framework we require our dependent variable to be generated from a distribution from the exponential family. Examples of such distributions include widely used distributions such as the normal, exponential and gamma distribution. Obviously this assumption is satisfied for our models.

The mean, μ, of the distribution then depends on the independent variables, X, through:

$$E(Y) = \mu = g^{-1}(X\beta),$$

where g is the link function and the other terms denote the same quantities as in a linear regression.

That is; $E(Y)$ is the expected value of the random variable Y and $X\beta$ is a linear combination of unknown parameters β. This is our linear predictor.

It is common to choose a function, V, that describes how the variance depends on the mean:

$$V(Y) = V(\mu) = V(g^{-1}(X\beta)).$$

V may be a function of the predicted value or conveniently, it may follow from the exponential distribution.

In this research we estimate the unknown parameters, β, with numerical optimization based on maximizing the appropriate likelihood function. More details on this numerical estimation follow in Section 3.3. Other popular estimation methods include least squares fits to variance stabilized responses or methods using Bayesian inference.

Via the linear predictor η we incorporate information about the independent variables into the model. The link function transforms the linear predictor into the expected value of the data corresponding to a particular model. η can be expressed as $\eta = X\beta$, because η is a linear combination of the unknown parameters β.

Three components make up a GLM:

1. Probability distribution: A response variable generated from a particular exponential distribution.
2. Linear predictor: $\eta = X\beta$.
3. Link function: g such that $E(Y) = \mu = g^{-1}(\eta)$.

The link function links the mean of the distribution of the dependent variable to the linear predictor. There is a wide range of commonly used link functions and choosing one can be quite arbitrary. A particular link function can be chosen in a way to match the domain of the link function to the range of the mean of the distribution function.

In this thesis we distinguish two variations for η:

1. A simple model in which $\eta = \lambda$.
2. A model with a home parameter in which $\eta = \lambda + Home$.

In the both cases we simply estimate the football ability X_i to have a single parameter football ability. λ is the vector of these abilities. The second case however adds a common home parameter to the model.
This parameter is equal for all teams. This additive model results in a fixed increase in football ability whenever a team is playing at home. So we have a simple model:

$$X_i = \lambda_i + \varepsilon \tag{1}$$

and a model with a home advantage effect:

$$X_i = \lambda_i + Home + \varepsilon. \tag{2}$$

In Equations (1) and (2) ε is the error term. We assume X_i to have gamma distribution and we subsequently model the outcome of football matches with the discriminal process $X_i - X_j$.

In Sections 3.2.1-3.2.4 we derive the distribution of $X_i - X_j$ for different values of the shape parameter of the gamma distribution. We then show that for those distributions, the corresponding models are the normal model by Thurstone, the model by Bradley-Terry, which is Luce's model for paired comparisons and the model by Dawkins for paired comparisons respectively. Finally we choose the appropriate link functions, g, based on the distributions $X_i - X_j$ and its domain. We compute likelihood functions based on the appropriate GLM models and numerically optimize those likelihood functions to find estimates for λ_i and $Home$.

3.2 Ranking Theory

Before we get into the gamma ranking models we need to introduce some notation.
Let $\pi = \pi_1, \dots, \pi_k$ denote a permutation of k teams where team π_p is ranked $(p = 1, \dots, k)$.
π_p^{-1} is the rank of team p. The vector $\pi^{-1} = [\pi_1^{-1}, \ \dots, \ \pi_k^{-1}]$ is thus the vector of ranks. For example; say $k = 3, \ \pi = (3\ 1\ 2)$ and say $\pi^{-1} = (2\ 3\ 1)$ is the permutation. In this case team 3 has the first ranking, team 1 has the second ranking and team 2 has the third ranking.
We define $p(\pi)$ as the distribution of permutations. $P(A)$ is thus the chance of observing event A out of all outcomes out of $p(\pi)$. For example; say we wish to compute the probability that team s finishes first:

$$p(s) = p(\pi_s^{-1}) = \sum_{\pi:\pi_1=s} p(\pi).$$

We generate random permutations $\Pi_l = (\Pi_{l1},\ \Pi_{l2}, \dots,\ \Pi_{lk})(l = 1, \dots,\ n)$ from the distribution of permutations (π). These are essentially all different team rankings.

Obviously these rankings can occur in different frequencies. The empirical probability mass function is thus:

$$p_N(\pi) = \frac{1}{N}\sum_{m=1}^{N} I(\Pi_m = \pi),$$

where I is the indicator function which is equal to 1 if $I(A)$ occurs and 0 otherwise.

3.2.1 Gamma Ranking models

The setting of gamma ranking models is as follows; say k teams compete in a competition where each team is to score r points. Let $X_1, \dots, X_k$ denote the times until k independent teams score r points. If we assume the scored points by team i to have a Poisson process with scoring intensity λ_i, then X_i has the gamma distribution with shape parameter equal to r and scale parameter equal to $\lambda_i > 0$. Let r be the shape parameter for all players and let $\lambda = [\lambda_1, \dots,\ \lambda_k]$ be the vector of scoring rates. The probability of permutation π, $X_{\pi 1} < \dots < X_{\pi k}$, is $p_\lambda^{(r)}(\pi)$ or, if we suppress the dependence on λ, is then $p^{(r)}(\pi)$. Please note that we assume r to be a known constant and integer for now. The probability, $p^{(r)}(\pi)$, is:

$$p^{(r)}(\pi) = P(X_{\pi 1} < \dots < X_{\pi k})$$

$$= \int_0^\infty \int_0^{x_{\pi_k}} \dots \int_0^{x_{\pi_2}} \left\{ \prod_{i=1}^{k} \frac{1}{\Gamma(r)} \lambda_{\pi_i}^r x_{\pi_i}^{r-1} exp(-\lambda_{\pi_i} x_{\pi_i}) \right\} dx_{\pi_1} \dots dx_{\pi_{k-1}} dx_{\pi_k} \tag{3}$$

$$= \int_0^\infty \int_0^{z_{\pi_k}\frac{\lambda_{\pi_{k-1}}}{\lambda_{\pi_k}}} \dots \int_0^{z_{\pi_2}\frac{\lambda_{\pi_1}}{\lambda_{\pi_2}}} \left\{ \prod_{i=1}^{k} \frac{1}{\Gamma(r)} \lambda_{\pi_i}^r x_{\pi_i}^{r-1} exp(-\lambda_{\pi_i} x_{\pi_i}) \right\} dx_{\pi_1} \dots dx_{\pi_{k-1}} dx_{\pi_k} \tag{4}$$

$$= g_r(z_i) = \frac{1}{\Gamma(r)} exp[-r(z_i - v_i)] exp\{-exp[-(z_i - v_i)]\}, z_i \in (-\infty, \infty). \quad (5)$$

Equation (4) follows from a change of variables $z_{\pi_i} = \lambda_{\pi_i} x_{\pi_i}$. We arrive at (5) by a logarithmic transformation: $Z_i = -\ln X_i$, where $X_i \sim \Gamma(r, \lambda_i)$. This transforms the gamma ranking model into a ranking model based on a location family of random variables with $v_i = \ln \lambda_i$ as the location parameter. You can recognize (5) as the generalized extreme value density in (Mihram, 1975). We normalize $\sum_{i=1}^{k} v_i = 0$, so that we arrive at (6) for the probability of the permutation π:

$$p^{(r)}(\pi) = \int_{-\infty}^{\infty} \int_{-\infty}^{z_{\pi_1}} \dots \int_{-\infty}^{z_{\pi_{k-1}}} \left\{ \prod_{i=1}^{k} g_r(z_{\pi_i} - v_{\pi_j}) \right\} dz_{\pi_k} \dots dz_{\pi_2} dz_{\pi_1}. \quad (6)$$

Our initial interpretation of r is that r is the total number of points a team is to score. The ranking model based on a location family of random variables gives us an alternative intuition behind this framework. We interpret the score for the mth object drawn from density $g_r(0)$ with location parameter $v_m = \ln \lambda_m$. In our setting this means that we draw the mth team's football ability X_m from the location family of random variables with $v_m = \ln \lambda_m$. This means comparisons are between team abilities with the same distribution except for a location shift. This limits us to Thurstone models with the discriminal processes in which both teams in a match are of the same distribution.

In our paired comparison football setting we are interested in the probability that team i beats team j $P(X_{\pi_i} > X_{\pi_j})$:

$$P(X_{\pi_i} > X_{\pi_j}) = P(X_{\pi_i} - X_{\pi_j} > 0),$$

which we model as

$$P\left(X_{\pi_i} - X_{\pi_j} > 0\right) = F\left(X_{\pi_i} - X_{\pi_j}\right). \quad (7)$$

Equation (7) holds as this is the basis of Thurstone processes and thus how we derive the three main choice models. In Subsections 3.2.2-3.2.4 we distinguish the three different cases: $r \to \infty$, $r = 1$, and $r \to 0$ and we find the appropriate distribution for the Thurstone processes $F(X_{\pi_i} - X_{\pi_j})$. After finding these distributions we choose the appropriate links functions g for the GLM framework. The three cases for r correspond to the

three preference models (by Thurstone, Luce and Dawkins respectively). Please note that we may drop π from our notation, so that we have $F(X_i - X_j)$.

3.2.2 Thurstone's Case V model

Thurstone's model scales a collection of stimuli based on paired comparisons. In our case this means that by evaluating all football matches between two teams, the law of comparative judgment estimates scaled values of the latent football ability such that we can make comparisons between all teams.
In his original model Thurstone assumes items can be compared on one metric, in our case X_i. Moreover, a discriminal process $F(X_i - X_j)$, which is normally distributed, determines the outcome of the comparison.

In the context of gamma ranking models, we consider the model in which the shape parameter goes to infinity, $r \to \infty$. We can approximate the distribution of permutations under the gamma model by considering permutations of normal random variables. We show in Appendix A that when r is large, $X \sim \Gamma(r, \lambda)$ is approximately Gaussian with mean $r\lambda^{-1}$ and variance $r\lambda^{-2}$. We compute the probabilities of the permutations $\pi = (1,\ 2, \dots,\ k)$ from a k-dimensional integral of normal densities.

We compute $F(X_{\pi_i} - X_{\pi_j})$ because we are interested in the probability of team i having a greater football ability than team j modelled by a Thurstone process with normally distributed football abilities. This ranking model is called the Thurstone-Mosteller-Daniels model (Daniels, 1950), which is the analogue of the paired Thurstone-Mosteller model. We obviously need Thurstone-Mosteller's paired comparison model. We know that the difference of these two independent normal random football abilities is again normally distributed:

$$X_i \sim N(r\lambda_i^{-1}, r\lambda_i^{-2})$$

$$X_j \sim N(r\lambda_j^{-1}, r\lambda_j^{-2})$$

$$X_i - X_j \sim N\big(r(\lambda_i^{-1} - \lambda_j^{-1}), r(\lambda_i^{-2} - \lambda_j^{-2} + \rho_{i,j})\big)$$

$$X_i - X_j \sim N(\mu, \sigma^2),$$

where $\mu = r(\lambda_i^{-1} - \lambda_j^{-1})$ and $\sigma^2 = r(\lambda_i^{-2} - \lambda_j^{-2} + \rho_{i,j})$ and $\rho_{i,j}$ denotes the covariance among the X_i. As discussed elsewhere, in (Engledrum, 2000) or (Torgerson, 1958), the full-blown model has too many parameters (such as, means, variances, and covariances,) that have to be estimated. We apply simplifying assumptions: We assume that the distributions are uncorrelated and normally distributed with different means but the same variance. The variance is equal to 1. This model is known as Thurstone's Case V model. The probability of team i beating team j is:

$$P(X_i - X_j > 0) = \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

$$= \int_{-\mu}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} dx.$$

By symmetry of the Gaussian,

$$= \int_{-\mu}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} dx = \int_{-\infty}^{\mu} \frac{1}{\sigma} \phi(\tfrac{x}{\sigma}) dx = \int_{-\infty}^{\frac{\mu}{\sigma}} \frac{1}{\sigma} \phi(t) dt$$

$$= \Phi(\mu)\,, \tag{8}$$

where $\Phi(z)$ is the standard normal cumulative distribution functions (CDF). This means Thurstone's Case V model corresponds to a discriminal process in which the teams competing in a match follow a gamma distribution with a shape parameter going to infinity.

For normally distributed Thurstone models we do not need to change any formulations to work in a GLM framework because these Thurstone models imply normally distributed errors. This means we have the straightforward identity link function:

$$X\beta = \mu.$$

In Section 3.3.1 we show how to formulate the appropriate likelihood function based on Equation (8).

3.2.3 The Bradley-Terry model

We model the gamma ranking model with $r = 1$ as Luce's model. If this assumption regarding the shape parameter holds, (6) is equal to the extreme value distribution. Again, we model the outcome of football matches as Thurstone processes: $F(X_{\pi_i} - X_{\pi_j})$. In Appendix B we show that the difference $X_{\pi_i} - X_{\pi_j}$ is logistic if X_{π_i} and X_{π_j} are Gumbel distributed quantities. In the introduction we refer to (Block & Marschak, 1960) and Holman and Marley cited by (Luce & Suppes, 1965) as alternative proof of this fact. Moreover, if this is the case, the Thurstone process $F(X_{\pi_i} - X_{\pi_j})$ is equivalent to Luce's model.

Luce's model was introduced to study behavior. Before going into detail on the practical model and its estimation, we first show the assumptions underlying Luce's model:

D1. Let there be four sets $R, S, T \; and \; U$, such that $R \subset S \subset T \subset U$.

D2. Let $x, y, z \in T$.

D3. Let $P(x, y)$ be the probability of choosing x instead of y, where $0 < P(x, y) < 1$.

D4. $P_s(R)$ is the probability of choosing R given choice from among alternatives in S.

The Choice Axiom then states:

(i) $P_T(R) = P_s(R)P_T(S)$.

(ii) $If \; P(x, y) = 0$ for some $x, y \in T$, $P_T(S) = P_{T-\{X\}}(S - \{X\})$.

The choice axiom defines the relationship of how choices within subsets are related in the context of an individual making choices under uncertainty. The well-known implication of the choice axiom is Lemma 3: Independence of irrelevant alternatives (IIA). This Lemma states that the relative probability of choosing alternatives is invariant to the composition of the larger set of alternatives. Again, the ratio is invariant, not the probabilities themselves. Another way of stating the same fact is that the log-odds of two choices are constant: $log\big(P_s(X) - P_s(Y)\big) = c$. We give the proof for Lemma 3 in Appendix C.

We compute the probability of the permutation π, $p^{(r)}(\pi) = P(X_{\pi 1} < \; \ldots < \; X_{\pi k}),$ as follows:

$$p^{(1)}(\pi) = \frac{\lambda_{\pi_1}}{\sum_{p=1}^{k} \lambda_{\pi_s}} \dots \frac{\lambda_{\pi_{k-1}}}{\sum_{p=2}^{k} \lambda_{\pi_p}} \frac{\lambda_{\pi_k}}{\lambda_{\pi_k}},$$

where $p^{(1)}(\pi)$ is the probability that k independent exponential random variables with means $\lambda_1^{-1}, \dots, \lambda_k^{-1}$ are ranked according to π. This probability has an intuitive interpretation of a sequence of rankings. The first fraction is exactly the probability that X_{π_1} is the minimum of k exponentials if their means are $\lambda_{\pi_1}, \dots, \lambda_{\pi_k}$. This means we rank π_1 as the first player. The second fraction is the probability that X_{π_2} is the smallest exponential of $k-1$ exponential random variables; that is X_{π_2} is the smallest in the set, excluding X_{π_1}.

In our case, we are working in a soccer setting, which means a match between two competing teams: $k = 2$. Once again, the probability of team i beating team j is defined as the probability of team i having a greater random football ability than team j. This is $P(X_{\pi_i} > X_{\pi_j})$ or $P(X_i > X_j)$ if we suppress the dependence on π.

Bradley and Terry derived the equivalent model for paired comparisons:

$$P(X_i > X_j) = \frac{\lambda_i}{\lambda_i + \lambda_j},$$

where λ_i and λ_j are the mean football ability of team X_i and team X_j respectively.

Bradley and Terry then improved their model by assuming exponential score functions:

$$P(X_i > X_j) = \frac{v_i}{v_i + v_j}, \tag{9}$$

so that $v_m = exp(\lambda_m)$ for $m = 1, 2$. Please note that this essentially comes down to estimating a logit model in a GLM framework, with the logit link function:

$$Log \frac{P(X_i > X_j)}{P(X_j > X_i)} = v_i - v_j.$$

This once more underlines the fact that the difference of two Gumbel distributed random football abilities is logistic.

Based on Equation (9) we formulate the likelihood function in Section 3.3.2.

3.2.4 Dawkins' model

Dawkins´ threshold model assumes that objects in a choice experiment have a latent ´threshold´. The more an object is preferred, the lower its threshold. More formally, $t_1, t_2, \dots, t_k$ are the thresholds corresponding to the choice objects $X_1, X_2, \dots, X_k$ and if $t_1 < t_2 < \cdots < t_k$, X_1 is most preferred and X_k the least. Furthermore, V denotes an "excitation" random variable, with $P(V \leq v) = H(v)$. We assume that $H(v)$ is monotonically increasing for v all and bounded: $0 < H(v) < 1$.
In our setting, when we have a match between team X_i and X_j with $t_i \leq t_j$, the outcome of the match depends on where V is relative to these thresholds: If $V < t_i \leq t_j$, simply draw another sample of V.
If $t_i < V \leq t_j$, team i beats team j. If $t_i \leq t_j < V$, team i and team j have a 50 percent chance of winning the match. Dawkins showed that the probability of team i beating team k, according to this model, in a complete system is equal to:

$$p_{ik} = 1 - 2p_{kj}p_{ji}. \tag{10}$$

Evidently, we define p_{ij} as the probability of team i beating team j in the case that V does not fall below both thresholds. Moreover Equation (10) only holds when $p_{ij} \geq 0.5,\ p_{jk} \geq 0.5$. The remaining cases can be derived from (10).

According to (Yelott, 1977), Dawkins' threshold model for paired comparisons is equivalent to a Thurstone model with a discriminal process of exponential variables.

The gamma ranking model with $r \to 0$ results in football abilities, X_i, that are exponential random with mean ε^{-1} and location parameter $\ln \lambda_i$. In (Stern, 1987) this is shown by considering the L_1 distance between the densities. This means we can consider the gamma ranking model with r sufficiently small by considering permutations of k independent exponential random variables with the same scale parameter ϵ and different location parameters. The equation for the probability of the permutation $\pi = (1,\ 2, \dots,\ k)$ is then again a multiple integral of the form of (3).

As mentioned before, we need the distribution of the Thurstone model with a discriminal process $X_i - X_j$ based on exponential random variables. We will show that $X_i - X_j$ has the double exponential distribution, or more commonly called the Laplace distribution, if $X_m \sim Exp(\varepsilon^{-1}, \ln \lambda_m)$ for $m \in \{1, \dots, k\}$.

Say h denotes the PDF of the standard exponential distribution:

$$h(v) = \begin{cases} e^{-v}, & v \geq 0 \\ 0, & v < 0 \end{cases}.$$

The PDF of $X_i - X_j$ is then:

$$g(u) = \int_{-\infty}^{\infty} h(v)h(v-u)dv,$$

by convolution. We then separate two cases. In the first case $v \geq 0$:

$$g(u) = \int_{u}^{\infty} e^{-v}e^{-(v-u)}dv = e^{u}\int_{-\infty}^{\infty} e^{-2v}dv = \frac{1}{2}e^{-u}.$$

And if $v < 0$:

$$g(u) = \int_{0}^{\infty} e^{-v}e^{-(v-u)}dv = e^{u}\int_{-\infty}^{\infty} e^{-2v}dv = \frac{1}{2}e^{u}.$$

It's trivial to show that the difference in location shifts of $X_i - X_j$ will show in a similar fashion for Y. So we ultimately find:

$$X_i - X_j \sim Laplace\big(\ln\lambda_i - \ln\lambda_j, \varepsilon\big). \tag{11}$$

We estimate the model for a fixed $\varepsilon = 1$. According to (Yelott, 1977), any Thurstone model with the Laplace distribution as its difference distribution also implies (10) and will consequently be equivalent to Dawkins' model for paired-comparison experiments.

Going back to the GLM framework, we need to choose the link function, g, as the Inverse link function for exponential variables:

$$X\beta = \mu^{-1}.$$

Section 3.3.3 contains notes on how to construct the likelihood based on (11).

3.3 Estimation

This Section contains notes on the estimation of the parameters. Per match, the gamma ranking models imply some distribution of possible match outcomes. Say we denote the CDF of that distribution of outcomes by $F(X_i - X_j)$ for a match between X_i and X_j. Based on the distribution F we compute a likelihood function. We once more remind you that that the probability that team i beats team j is $P(X_i > X_j) = F(X_i - X_j)$. We can construct the likelihood as follows:

$$Likelihood = \prod_{i<j} \prod_{j} \binom{n_{ij}}{a_{ij}} \left[F(X_i - X_j)\right]^{a_{ij}} \left[1 - F(X_i - X_j)\right]^{n_{ij} - a_{ij}}. \tag{12}$$

In (12) n_{ij} is the number of goals scored in a match between team i and j and a_{ij} is the number of times team i scored against team j. Note that this likelihood takes the form of a binomial likelihood with the probability of a success equal to $F(X_i - X_j)$, the probability that team i beats team j.

We stress the fact that the distribution $F(X_i - X_j)$ depends on the assumption with respect to the shape parameter of the gamma distribution. This means $F(X_i - X_j)$ separates the different models, while the binomial form of (12) as a whole holds for all models.

We need to maximize (12) over the unknown parameters. Maximizing the likelihood is analogous to maximizing the natural logarithm of the likelihood function:

$$\sum_{i<j} \sum_{j} \log \binom{n_{ij}}{a_{ij}} + a_{ij} \log\left[F(X_i - X_j)\right] + (n_{ij} - a_{ij}) \log\left[1 - F(X_i - X_j)\right].$$

When optimizing likelihoods we generally drop the term $\log \binom{n_{ij}}{a_{ij}}$ as this does not depend on the parameters we maximize over.

3.3.1 Thurstone's Case V model

For Thurstone's Case V model we decided in Section 3.2.2 on the CDF of the standard normal distribution as a choice for F, see (8). $F(X_i - X_j)$ is thus $\Phi(\mu_i - \mu_j)$. We substitute $F(X_i - X_j)$ for $\Phi(\mu_i - \mu_j)$ in (12) and when we drop the term $\binom{n_{ij}}{a_{ij}}$, we find:

$$Max \prod_{i<j} \prod_j [\Phi(\mu_i - \mu_j)]^{a_{ij}} [1 - \Phi(\mu_i - \mu_j)]^{n_{ij} - a_{ij}}. \quad (13)$$

As mentioned, the maximization problem described in (13) has the same solution as maximizing the natural logarithm of the likelihood. We do so with respect to one constraint:

$$Max \sum_{i<j} \sum_j a_{ij} Log[\Phi(\mu_i - \mu_j)] - (n_{ij} - a_{ij}) Log[1 - \Phi(\mu_i - \mu_j)]$$

$$\text{w.r.t.} \sum_{i=1}^{k} \mu_i = 0,$$

The constraint should prevent overfitting. Because the underlying distribution for μ_i and μ_j is standard normal, we theoretically allow for all values for μ_i and μ_j; we do not require further constraints on the parameters.

3.3.2 The Bradley-Terry model

The Bradley-Terry model has a different likelihood function as $F(X_i - X_j)$ has a different distribution. As mentioned in Section 3.2.3, we assumed the following model:

$$F(X_i - X_j) = \frac{v_i}{v_i + v_j}, \quad (9)$$

where $v_m = exp(\lambda_m)$. When we substitute $\frac{v_i}{v_i + v_j}$ for $F(X_i - X_j)$ in (12), we find the likelihood for the Bradley-Terry model:

$$Likelihood \propto \prod_{i<j} \prod_{j} \left(\frac{v_i}{v_i + v_j}\right)^{a_{ij}} \left(\frac{v_j}{v_i + v_j}\right)^{(n_{ij} - a_{ij})}. \quad (14)$$

Note the proportionality symbol; we dropped terms that did not depend on the parameters we optimize over. We then maximize the natural logarithm of (14):

$$Max \sum_{i<j} \sum_{j} a_{ij} Log\left(\frac{v_i}{v_i + v_j}\right) - (n_{ij} - a_{ij}) Log\left(\frac{v_j}{v_i + v_j}\right),$$

$$\text{w. r. t.} \sum_{i=1}^{k} v_i = 1,$$

$$v_i \geq 0 \; \forall \; i.$$

The constraint on the sum of the parameters is in place to prevent overfitting. Please note that we did not incorporate this constraint literally for the Bradley-Terry model. We fixed one of the team abilities to achieve the same effect. Moreover we require v_i to be positive as we chose the exponential score function $v_m = exp(\lambda_m)$.

3.3.3 Dawkins' model

For Dawkins' model we found that $F(X_i - X_j) = Laplace(\ln\lambda_i - \ln\lambda_j, \varepsilon)$, where we chose $\varepsilon = 1$ which means:

$$F(X_i - X_j) = \frac{1}{2} e^{-|a_{ij} - (n_{ij} - a_{ij}) - (\ln\lambda_i - \ln\lambda_j)|}. \quad (15)$$

Please note that (15) depends on $a_{ij} - (n_{ij} - a_{ij}) - (\ln\lambda_i - \ln\lambda_j)$, which essentially is the difference between the observed score differential and the differential of estimated natural logarithms of the football abilities λ.

To arrive at the likelihood we take the product over all matches between all teams:

$$Likelihood \propto \prod_i \prod_j e^{-|a_{ij}-(n_{ij}-a_{ij})-(\ln\lambda_i-\ln\lambda_j)|}. \tag{16}$$

In (16) we dropped the factor $\frac{1}{2}$ as well, because this does not depend on the parameters we maximize over. We then compute the natural logarithm of the likelihood and maximize that with respect to one constraint on the parameters:

$$Max \sum_{i<j} \sum_j -|a_{ij} - (n_{ij} - a_{ij}) - (\ln\lambda_i - \ln\lambda_j)|$$

$$\sum_{i=1}^{k} \lambda_i = 0.$$

The constraint again should prevent overfitting.

We use the global optimization tools from *Matlab* in combination with the built in functions *'fmincon' and 'fminunc'* to deal with constrained numerical optimization.

3.4 Evaluating rankings

Based on the described methodologies in Section 3.1 – 3.3, we estimate team football ability parameters which we can thereafter rank. Obviously not all models produce the same ranking, so we should evaluate their performance in some way. We do so in three distinct ways: We look at rank correlation, we analyze standard errors and we compare the models' fit. To compare non-nested models we use the Akaike Information Criterion (AIC) as a measure of the goodness of fit. We compare nested models with the likelihood Ratio tests.
This Section is made up in the following way: First we introduce the concept of concordance in Section 3.5.1. We elaborate on Kendall's tau and Spearman's rho in Section 3.4.2 and 3.4.3 respectively. They measure the extent to which rankings correlate. In Section 3.5.4 we go into more detail on how we calculate and apply the AIC and finally in Section 3.5.5 we discuss the likelihood ratio test.

3.4.1 Concordance

To explain how we compute Kendall's tau and Spearman's rho, we first have to elaborate on the concept of concordance and discordance. Say we have a set of n observations of the joint random variables X and Y. Moreover, we assume uniqueness of the x_i and y_i. We call pairs of observations concordant if the ranks for both elements agree and we call pairs of observations discordant if the ranks do not agree:

Concordant: $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$

Discordant: $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$.

Obviously there is a third category. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant.

3.4.2 Kendall's tau

Kendall's rank correlation coefficient is developed by Maurice Kendall. With Kendall's tau (τ) we measure the ordinal association between two measured quantities. This is a non-parametric coefficient of correlation on ranks, or a rank statistic. We use the term 'correlation' as a measure of the linear relationship between covariates. We should rather categorize Kendall's tau as a measure of association because this refers to a monotone relationship between covariates. We use Kendall's τ to distinguish the results of the models from each other.

We define Kendall's τ as:

$$\tau = \frac{(\#Corcordant\ Pairs) - (\#Discordant\ Pairs)}{\frac{1}{2}n(n-1)}. \tag{17}$$

In (17), $\frac{1}{2}n(n-1)$ is the denominator. This the total number of pair combinations, which results in a tau in the range of $(-1,1)$. The more the ranking of the measured quantities X and Y are in agreement, the closer τ

is to 1. The opposite is true as well; the more the ranking of the measured quantities X and Y are in disagreement, the closer τ is to -1. In the case of independence between the rankings we find $\tau = 0$.

For the population τ we have a similar equation for random variables X and Y. Let $(X_\mathrm{i}, Y_\mathrm{i})$ and $(X_\mathrm{j}, Y_\mathrm{j})$ be independent vectors with the same distribution(X, Y), then

$$\tau = P\big[(X_i - X_j)(Y_i - Y_j) > 0\big] + P\big[(X_i - X_j)(Y_i - Y_j) < 0\big]$$

and

$$P\big[(X_i - X_j)(Y_i - Y_j) > 0\big] + P\big[(X_i - X_j)(Y_i - Y_j) < 0\big] = corr\big[sign(X_i - X_j), sign(Y_i - Y_j)\big].$$

τ is the Pearsen product-moment correlation coefficient of the random variable sign$\big(X_\mathrm{i} - X_\mathrm{j}\big)$ and sign$\big(Y_\mathrm{i} - Y_\mathrm{j}\big)$. Therefore, τ is sometimes called the difference sign correlation coefficient.

We utilize Kendall's tau as test statistic to establish whether two rankings may be regarded as statistically independent.

If we define τ as we did in (42), we find the following quantity to be standard normal when the variables are statistically independent:

$$z_a = \frac{3[(\#Corcordant\ Pairs) - (\#Discordant\ Pairs)]}{\sqrt{\frac{1}{2}n(n-1)(2n+5)}}.$$

After computing z_a we can test in the usual way whether the two rankings are statistically independent; we use the cumulative probability of a standard normal distribution to create confidence intervals.

3.4.3 Spearman's rho

Spearman rho (ρ) is a correlation coefficient developed by the psychologist C. Spearman in 1904. This is a non-parametric coefficient of correlation on ranks (or a rank statistic) as well. Based on a similar reasoning as in Section 3.4.2 on Kendall's tau we should also categorize Spearman's rho as a measure of association. Spearman's rho is calculated by performing the Pearson product-moment correlation coefficient computations to the ranks associated with a sample $\{(x_i, y_i)\}_{i=1}^{n}$. Say $R_i = rank(x_i)$ and $S_i = rank(y_i)$; We then find the calculate (Pearson) correlation coefficient rs for $\{(R_i, S_i)\}_{i=1}^{n}$ as follows:

$$rs = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2 \sum_{i=1}^{n}(S_i - \bar{S})^2}},$$

$$= 1 - \frac{6\sum_{i=1}^{n}(R_i - S_i)^2}{n(n^2-1)}, \tag{18}$$

where $\bar{R} = \sum_{i=1}^{n} R_i/n = \frac{n+1}{2} = \sum_{i=1}^{n} S_i/n = \bar{S}$. (18) holds solely because of the lack of ties.

Say we have two random variables, X and Y with corresponding distribution functions $F_X(X)$ and $F_y(Y)$. We define the population parameter, denoted by ρ_s, as the pearson product-moment correlation coefficient of $F_X(X)$ and $F_y(Y)$:

$$\rho_s = corr[F_x(X), F_y(Y)]$$

$$= 12E[Fx(X)Fy(Y)] - 3.$$

$F_X(X)$ and $F_y(Y)$ are sometimes referred to as the "grades" of X and Y. Spearman's ρ_s is therefore occationally referred to as the grade correlation coefficient.

Spearman's rho is a measure of association based on the concept of concordance, just as Kendall's tau is. In Subsection 3.5.1 we touched on the subject of concordance and discordance. If $(X_\mathrm{i}, Y_\mathrm{i})$, $(X_\mathrm{j}, Y_\mathrm{j})$ and $(X_\mathrm{h}, Y_\mathrm{h})$ are independent random vectors with the same distribution as (X, Y), then

$$\rho s = 3P\left[\left(X_{\mathrm{i}} - X_{\mathrm{j}}\right)\left(Y_{\mathrm{i}} - Y_{\mathrm{h}}\right) > 0\right] - 3P\left[\left(X_{\mathrm{i}} - X_{\mathrm{j}}\right)\left(Y_{\mathrm{i}} - Y_{\mathrm{h}}\right) < 0\right]. \quad (19)$$

That is; the difference between the probabilities of concordance and discordance between the random vectors (X_{i}, Y_i) and $\left(X_j, Y_h\right)$ is proportional to ρs. Obviously we can swap $\left(X_j, Y_h\right)$ for $\left(X_h, Y_j\right)$ in (19).

As Spearman's rho is nothing else than Pearson's linear correlation we can use a t-test for its significance. We should however perform a transformation on r in the following way:

$$t = r\sqrt{\frac{n-2}{1-r^2}}.$$

This statistic is approximately distributed as a Student's t-distribution with n-2 degrees of freedom under the null hypothesis.

3.4.4 Akaike information criterion

We measure the goodness of fit with the Akaike information criterion (AIC). When adding parameters to models, the model never gets worse in terms of fitting the training sample. With every extra parameter we introduce the model loses generalization power. Out of sample estimation might perform worse with the introduction of extra parameters. Moreover, more complex models are more difficult to interpret. AIC offers a relative estimate of the information lost when a given model is used to represent the process that generates the data. In doing so, it deals with the trade-off between the goodness of fit of the model and the complexity of the model. AIC does so by introducing a penalty term for the number of parameters in the model.

To compare nested models we use the likelihood ratio test as well. For non-nested models we cannot do likelihood ratio tests as these are restricted to nested models. Therefore, we use AIC to compare non-nested models.

AIC measures the fit of a model relative to other models for a given dataset. While this does not quantify the absolute fit for a given model, it does provide a means for model selection. This means there is no concrete test in which we reject or accept some null hypothesis.

Let L denote the maximum likelihood value under some model and let k denote the number of estimated parameters in the model. AIC then has the following equation:

$$AIC = 2k - 2\log(L).$$

The model with the highest absolute AIC value is favored.

3.4.5 Likelihood ratio test

The likelihood ratio test is used to compare nested models. In our case the models without a home parameter are nested within models with a home parameter. This is quite straightforward as one can simply impose the restriction that the home parameter is equal to zero. The intuition is that we calculate the ratio between the model with the restriction and without the restriction. Based on this ratio we can perform tests on whether or not the difference in likelihood is significant.

Let L_1 be the maximum value of the likelihood of the data without the additional restriction. In other words, L_1 is the likelihood of the data with all the parameters unrestricted and maximum likelihood estimates substituted for these parameters. Let L_0 be the maximum value of the likelihood when the parameters are restricted (and reduced in number) based on the assumption. Assume k parameters were lost (so L_0 has k less parameters than L_1).

We compute the ratio $\frac{L_0}{L_1}$, which should be between 0 and 1. The less likely the assumption is, the smaller $\frac{L_0}{L_1}$ will be. We calculate $-2\ Ln(\frac{L_0}{L_1})$, which is distributed as $\chi^2(k)$ where k is the number of restrictions we impose.

In our case, L_1 is the likelihood of the data with the model with the home parameter and L_0 is the likelihood of the data without the home parameter. This means we calculate $-2\ Ln(\frac{L_0}{L_1})$ and compare its value to a $\chi^2(1)$ cumulative distribution function.

4. Results

This Section contains the findings of this research. Section 4.1 is aimed at the comparison between the models by Thurstone, Bradley-Terry and Dawkins. In Section 4.2 we discuss the model with a home parameter. We use Kendall's tau, Spearman's rho and Akaike Information Criterion (AIC) as evaluation criteria and perform the appropriate tests. Finally, 4.3 contains the results of the comparison of the simple model to the model with a home parameter through the likelihood ratio tests.

4.1 Simple Model

The first model we consider is the model where we estimate the team's football ability with the simple model: $X_i = \lambda_i + \varepsilon$. Table 4 shows the estimates for λ_i based on the different models. Please note that SC Heerenveen is the reference category for the Bradley-Terry model.

As expected, the three gamma ranking models show a similar ranking in terms of their point estimate. The final three rows in Table 4 show Kendall's tau, Spearman's rho and the AIC values for these three models. Kendall's tau and Spearman's rho paint the same picture: Thurstone's model, the gamma ranking model with the shape parameter going to infinity, is closest to the real ranking in terms of the distance measured by Kendall's tau. Dawkins' model, with a shape parameter going to zero, takes a second place and the model with a shape parameter of 1, the Bradley-Terry is model third. Bradley-Terry's model and Dawkins' model are not relatively just far apart from the real ranking when considering their Kendall's tau values are 0.76 and 0.79 respectively. This is in contrast to the value corresponding to the Thurstone model, which is 0.97. This is once more underlined by the Spearman's rho values that also find the gamma ranking model with a relatively large shape parameter ($r \to \infty$), Thurstone's model, to be closest to the real ranking. Moreover, the distance measured by Spearman´s rho between Bradley and Terry's model and Dawkins' model is similar with a value of 0.90.

The parameter estimates are all significantly different from zero (marked bold in Table 4). The reference category (SC Heerenveen) in the Bradley-Terry model does not have an estimated standard error.
The goal deficit is only positive from Ajax to AZ as is shown in the final column in Table 3 in the data Section. We estimated these models on the goal deficit data instead of on the match outcome (win, draw, loss) data.

In the Thurstone model and Dawkins' model we see this aspect of the data in the sense that after AZ all teams have a negative parameter estimate.

Table 4 - Results for the simple gamma ranking models

Final ranking	Thurstone		Bradley-Terry		Dawkins	
Ajax	**0.76**	(0.028)	**1.94**	(0.145)	**1.14**	(0.0277)
Feyenoord	**0.53**	(0.029)	**1.45**	(0.069)	**1.00**	(0.0288)
FC Twente	**0.51**	(0.034)	**1.51**	(0.08)	**0.97**	(0.0336)
PSV	**0.27**	(0.028)	**1.06**	(0.041)	**0.42**	(0.0277)
SC Heerenveen	**0.28**	(0.03)	1.00	()	**0.58**	(0.0301)
Vitesse	**0.24**	(0.028)	**1.11**	(0.038)	**0.44**	(0.028)
Groningen	**0.12**	(0.029)	**0.83**	(0.024)	**0.11**	(0.029)
AZ	0.00	(0.029)	**0.89**	(0.029)	**0.11**	(0.0288)
ADO Den Haag	**-0.11**	(0.029)	**0.59**	(0.013)	**-0.53**	(0.0292)
FC Utrecht	**-0.16**	(0.032)	**0.56**	(0.012)	**-0.53**	(0.0315)
PEC Zwolle	**-0.12**	(0.026)	**0.75**	(0.023)	**-0.06**	(0.0264)
SC Cambuur	**-0.21**	(0.03)	**0.67**	(0.02)	**-0.28**	(0.0302)
Go Ahead Eagles	**-0.24**	(0.028)	**0.53**	(0.01)	**-0.67**	(0.0278)
Heracles Almelo	**-0.28**	(0.029)	**0.62**	(0.015)	**-0.39**	(0.0286)
NAC Breda	**-0.30**	(0.029)	**0.63**	(0.016)	**-0.31**	(0.029)
RKC Waalwijk	**-0.37**	(0.031)	**0.52**	(0.01)	**-0.56**	(0.0313)
NEC	**-0.40**	(0.021)	**0.55**	(0.009)	**-0.78**	(0.0205)
Roda JC Kerkrade	**-0.51**	(0.016)	**0.52**	(0.01)	**-0.69**	(0.0159)
Kendall´s tau	0.97		0.76		0.79	
Spearman´s rho	0.996		0.90		0.90	
Akaike Information Criterion	-998		-1220		-1740	

The absolute uncertainty around the parameter estimates is quite similar for the gamma ranking models by Thurstone and Dawkins. The model by Bradley and Terry seems to have larger standard errors for the teams for which a higher football ability is estimated.
It's notable that we do not find that the ranking based on the Bradley-Terry model is 'the middle way' between the Thurstone model and Dawkins' model in terms of its estimates and in terms of standard errors.

To calculate win probabilities is quite straightforward. For the Thurstone model we assumed the football abilities to be normally distributed with different means but common variance of 1. To calculate the expected probability that Ajax beats Feyenoord we use the normal CDF as follows: $\Phi(0.76 - 0.53) = 0.591$ which

means the expected probability Ajax beats Feyenoord is 59.1 %. Similarly for the Bradley-Terry model we compute the expected probability that Ajax beats Feyenoord as $\frac{e^{1.94}}{(e^{1.94}+e^{1.45})} = 0.6201$ which is 62.01%. For the Dawkins model we use the Laplace CDF: $\frac{1}{2} + \frac{1}{2}sgn(1.14-1)\left(1-e^{(-|1.14-1|)}\right) = 0.5653$ or 56.53%. Note that the match played in Amsterdam ended in 2-1 just as the match played in Rotterdam ended in 2-1. This means the goal deficit over these matches is 0. Clearly these estimates are highly similar.

Note that the underlying distributions for the football ability in Thurstone's Case V model is the normal distribution, while Luce's choice axiom implies Gumbel football abilities and Dawkins' model corresponds to exponential football abilities. These models imply normally distributed, logistically distributed and Laplace distributed match outcomes.

Figure 1 shows the simulated distributions for the outcome of the match Ajax – Feyenoord according to these three distributions. It's evident that these distributions are located at $0.76 - 0.53 = 0.23$, $1.94 - 1.45 = 0.49$ and $1.14 - 1 = 0.14$ for the normal, logistic and exponential distributions respectively. Moreover, all three distributions are symmetrical. In our context this means negative and positive deviations from the point estimate have the same implication for the probabilities of such outcomes.

Moreover, we see that the logistic distribution and the Laplace distribution have fatter tails. These distributions are better equipped to deal with data with more outliers than in the normal case. Based on the simulated densities we calculate the corresponding kurtoses to show this: The kurtoses in our simulations are 3.0020, 4.197 and 5.9943 for the normal, logistic and Laplace model respectively. In this sense the logistic model is in between the normal model and the Laplace model. This finding extends to cases in which we compare more than two items at a time.

As additional proof of the fact that the normal distribution cannot cope with the fat tails in the data in comparison to the logistic and Laplace distribution, we calculate binomial cumulative probabilities and perform the appropriate tests.

We empirically observe 123 out of 314 matches ending with a goal deficit of at least 2. This is a proportion of 0.3917. We calculate the probability of Ajax beating Roda JC Kerkrade with at least 2 goals, because Ajax and Roda JC Kerkrade have the highest and lowest estimated football abilities respectively. Any binomial cumulative probability based on this probability should be considered an upper limit. The outcome of the match Ajax – Roda is distributed with a variance of 1 around the location, $0.76 - -0.51 = 1.27$. The probability is then: $1 - \Phi(2 - 1.27) \approx 0.2327$. The probability of observing more than 123 of such match results is:

$$1 - \sum_{i=0}^{123} \binom{314}{i} 0.2327^{i} 0.7673^{314-i} \approx (1-1) = 0.$$

On a 5% significance level this result indicates that the estimate of 0.2327 is too high. Especially when considering only the matches between Ajax and Roda have that particular estimate. For more evenly matched teams this chance would be even smaller.

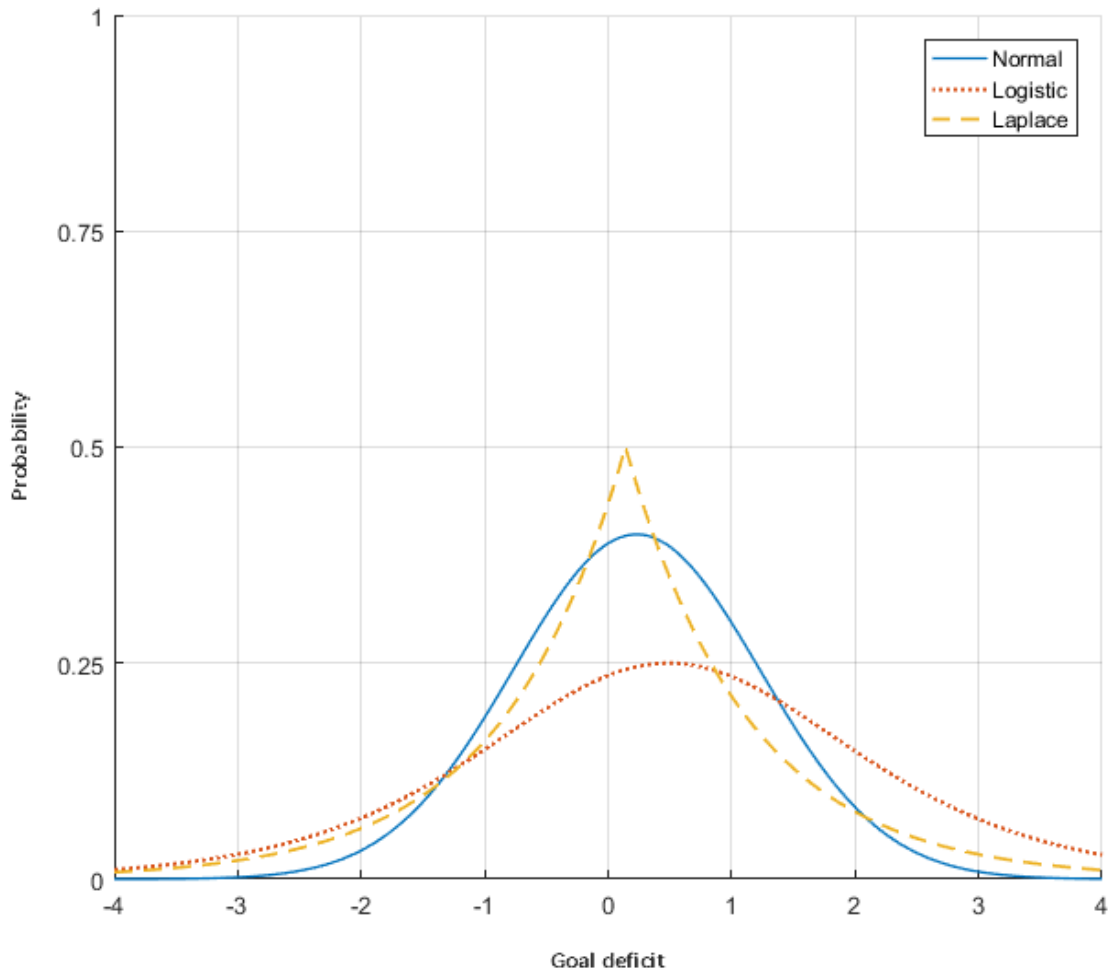

Figure 1: Probability Distribution Functions of the Thurstone models for Ajax - Feyenoord

Following a similar reasoning for the other models we find the probability of Ajax beating Roda with at least a 2 goal deficit to be $\frac{e^{1.94}}{(e^{1.94}-e^{0.52})} = 0.8053$ and $\frac{1}{2}+\frac{1}{2}sgn(2-1.14-0.69)\left(1-e^{(-|2-1.14-0.69|)}\right) = 0.5782$ for the logistic and Laplace models.

The AIC values back up the claim that the logistic distribution and Laplace distribution are better able to deal with the fatter tails in the data. The model by Dawkins has the highest absolute value for AIC, the measure of goodness of fit. Bradley-Terry is second in this regard, while Thurstone is third. The corresponding values are -998, -1220 and -1740, respectively. Based on these findings, Dawkins' model seems to be the best fit.
We believe this is due to the fact that the Laplace distribution combines fat tails with a narrow peak.

In Table 5 we see that all rankings are positively correlated in terms of Kendall's tau as we may expect from Thurstone models that all abide by Luce's choice axiom. Moreover all these models have the ability to realistically evaluate team football abilities, as is shown above. The real ranking is the highest correlated to the Thurstone model and quite similarly correlated to the other two models. Note that the Gamma ranking model by Dawkins has a higher correlation to both the Thurstone model and Dawkins' model than these models among each other. In this sense we might say Dawkins' estimates are in between the other two.

Table 5 - Kendall's tau ranking correlation coefficient for the simple model

Kendall's tau	Real	TH	BT	DA
Real	X	0.97	0.77	0.79
TH	X	X	0.77	0.82
BT	X	X	X	0.89
DA	X	X	X	X

To test whether we find significant proof for our claims, we compute Z-statistics and P-values. We reject the hypothesis of independent rankings for all rankings on all commonly used significance levels (5%, 2.5%, 1%). This means we should reject the hypothesis that the real ranking and the rankings proposed by any one of the gamma ranking models are independent. We reject the hypothesis that the rankings based on the gamma ranking models are independent among their selves as well. Please note that Table 6 contains these Z-statistics and within the brackets their corresponding P-values.

Our main result is once more underlined by Spearman's rho; Table 7 shows that the Thurstone model has the highest correlated ranking to the real ranking. Furthermore, we find that the tests for independence for all gamma ranking models and the real ranking among each other are rejected at all the commonly used significance levels. Table 8 shows these results.

Table 6 - Kendall's tau testing for independence for the simple model

tau-Z(P)	Real	TH	BT	DA
Real	X	5.64(1.7-08)	4.47(7.8-e-06)	4.58(4.6e-06)
TH	X	X	4.47(7.8-e-06)	4.73(2.2e-06)
BT	X	X	X	5.15(2.6e-07)
DA	X	X	X	X

Table 7 - Spearman's rho ranking correlation coefficient

Spearman's rho	Real	TH	BT	DA
Real	X	0.996	0.90	0.90
TH	X	X	0.90	0.91
BT	X	X	X	0.97
DA	X	X	X	X

Table 8 - Spearman's rho testing for independence

spearman's rho	Real	TH	BT	DA
Real	X	43.89(0)	8.06(5.0e-07)	8.11(4.7e-07)
TH	X	X	8.46(2.7e-07)	8.73(1.7e-07)
BT	`X	X	X	15.43(5.0e-11)
DA	X	X	X	X

Concluding remarks regarding the simple model are that Thurstone's model is best equipped to reproduce the ranking, though that's not the case for its fit as a whole. Dawkins' model especially seems to outperform

Thurston's model and Bradley-Terry's model in that regard because of its ability to model fat tails as well as sharp peaks.

4.2 Model with home parameter

While our measures of association agreed on the fact that the rankings based on the simple gamma ranking models were relatively close to the real ranking, we believe a home parameter would be a great addition. This was supported by the fact that we empirically found a home advantage in Section 2. We estimated the model: $X_i = \lambda_i + Home + \varepsilon$.

Table 9 shows the results for the gamma ranking models based on the model with a home parameter. Table 9 is built up in the same fashion as Table 4, except for the fact that the first row contains the home parameter estimates.

We used an additive model, in the sense that the team football ability at home is simply calculated by adding the home parameter to the strength parameter. A fairly obvious but important observation is that the home parameter is positive in all models, which means we find support for a home advantage in our model. More specifically, the advantage of playing at home is larger than the football ability of quite a lot of teams. For Thurstone's model and the Bradley-Terry model only Ajax' estimated football ability is larger than the home advantage. Bradley-Terry's model has a relatively small home parameter.

The model by Dawkins has the highest absolute value for AIC, the measure of goodness of fit. Bradley-Terry is second in this regard, while Thurstone is third. This is in agreement with our findings from the simple model. The corresponding values are -1010, -1242, and -1742 respectively. Based on these findings Dawkins' model seems to be the best fit.

We estimate the probability of Ajax to beat Feyenoord when playing at home to be $\Phi(0.74 + 0.65 - 0.43) = 0.8365$, which is 83.15%. For the Bradley-Terry model we find that same probability as follows: $\frac{e^{1.87+1.39}}{(e^{1.87+1.39}+e^{1.39})} = 0.8665$ or 86.65%.
With the Dawkins model we use the Laplace CDF: $\frac{1}{2} + \frac{1}{2} sgn(1.25 + 0.86 - 1.05)\left(1 - e^{(-|1.25+0.86-1.05|)}\right) = 0.8268$ or 82.68%.

Table 9 - Results for the gamma ranking models with a home parameter

Final ranking	Thurstone		Bradley-Terry		Dawkins	
Home Parameter	**0.65**	(0.002)	**1.39**	(0.009)	**0.86**	(0.003)
Ajax	**0.74**	(0.018)	**1.87**	(0.138)	**1.25**	(0.015)
Feyenoord	**0.43**	(0.019)	**1.39**	(0.065)	**1.05**	(0.015)
FC Twente	**0.48**	(0.026)	**1.44**	(0.073)	**1.04**	(0.017)
PSV	**0.21**	(0.020)	**1.04**	(0.039)	**0.45**	(0.015)
SC Hereveen	**0.19**	(0.020)	1.00	()	**0.59**	(0.016)
Vitesse	**0.14**	(0.019)	**1.02**	(0.033)	**0.46**	(0.015)
Groningen	0.00	(0.019)	**0.81**	(0.024)	**0.12**	(0.016)
AZ	**0.07**	(0.019)	**0.84**	(0.026)	**0.13**	(0.016)
ADO Den Haag	**-0.22**	(0.019)	**0.55**	(0.011)	**-0.57**	(0.015)
FC Utrecht	**-0.27**	(0.021)	**0.54**	(0.011)	**-0.57**	(0.015)
PEC Zwolle	**-0.03**	(0.015)	**0.72**	(0.022)	**-0.03**	(0.017)
SC Cambuur	**-0.09**	(0.021)	**0.62**	(0.018)	**-0.25**	(0.012)
Go Ahead Eagles	**-0.31**	(0.019)	**0.51**	(0.010)	**-0.73**	(0.015)
Heracles Almelo	**-0.19**	(0.019)	**0.61**	(0.015)	**-0.41**	(0.015)
NAC Breda	**-0.16**	(0.018)	**0.62**	(0.016)	**-0.30**	(0.015)
RKC Waalwijk	**-0.32**	(0.022)	**0.48**	(0.009)	**-0.59**	(0.015)
NEC	**-0.34**	(0.017)	**0.51**	(0.008)	**-0.89**	(0.015)
Roda JC Kerkrade	**-0.33**	(0.017)	**0.51**	(0.010)	**-0.75**	(0.015)
Kendall´s tau	0.82		0.76		0.78	
Spearman´s rho	0.93		0.82		0.92	
Akaike Information Criterion	-1010		-1242		-1742	

Looking at Table 9 we see that the Thurstone Case V model once more outperforms Dawkins' and the Bradley-Terry model in terms of Kendall's tau. This is also the case in terms of the other association measure, Spearman's rho. The main result in Table 5 is that Thurstone's Case V model actually moved away from the real ranking towards the other gamma ranking models in terms of distance measured by rank correlation in terms of Kendall's tau.

Table 10 and 11 contain the Kendall's tau coefficient and corresponding Z-statistics and P-values. The results agree to an extent with the ones in Section 4.1; that is, we reject the hypothesis that these rankings are independent among each other. Adding the home parameter moved Thurstone's Case V model closer to the Bradley-Terry model and Dawkins´ model, but consequently further away from the real ranking. This is reflected in the Kendall's tau values. The rank correlation between the Thurstone model and the alternative

gamma ranking models increased. In terms of Spearman's rho, the measured association with the real ranking is still the highest for the Thurstone Case V model. After adding a home parameter, all gamma ranking models are more in agreement on the ranking of parameter estimates in the sense that they have a higher rank correlation among each other, which is in line with earlier findings regarding Kendall´s tau.

Table 9 - Kendall's tau ranking correlation coefficient for the home model

Kendall's tau	Real	TH	BT	DA
Real	X	0.82	0.76	0.78
TH	X	X	0.93	0.94
BT	X	X	X	0.88
DA	X	X	X	X

Table 10 - Kendall's tau testing for independence for the home model

tau-Z(P)	Real	TH	BT	DA
Real	X	4.73(2.2e-06)	4.43(9.3e-06)	4.55(5.5e-06)
TH	X	X	5.42(6.1e-07)	5.45(4.9e-08)
BT	X	X	X	5.08(3.9e-07)
DA	X	X	X	X

Table 11 - Spearman's rho ranking correlation coefficient for the home model

Spearman's rho	Real	TH	BT	DA
Real	X	0.93	0.82	0.92
TH	X	X	0.88	0.98
BT	X	X	X	0.93
DA	X	X	X	X

Table 12 - Testing for independence for Spearman's rho for the home model

Spearman's rho	Real	TH	BT	DA
Real	X	40.88(0)	7.35(1.6e-07)	8.29(3.5e-07)
TH	X	X	8.23(3.8e-07)	9.38(6.7e-8)
BT	X	X	X	14.82(9.1e-11)
DA	X	X	X	X

4.3 Nested model comparisons

In Section 4.1 we compared the non-nested models. To compare the models based on their fit we compare the models´ standard errors and perform the likelihood ratio test in this Section.

Generally speaking, the gamma ranking models with a home parameter have smaller standard errors than the ones without a home parameter (please note that we are comparing Thurstone Case V models among each other, Bradley-Terry models among each other and Dawkins' models among each other). This means we estimate our parameters, the football abilities, more accurately after we add a home parameter. Moreover the home parameter was significantly different from 0. Based on these findings we conclude that the real ranking obviously is influenced by a home effect, which causes Thurstone's model to have a lower ranking correlation after introducing a home effect in our model. This separates a home advantage effect from football ability, which clearly is not the case for the real rankings. When attempting to mimic the real ranking one should not add a home effect, however when attempting to model the latent football abilities, one definitely should add a home advantage effect.

To test whether or not we find similar evidence in terms of likelihood, we show results of the likelihood ratio tests in Table 14. All three gamma ranking models agree on the results. We reject the hypothesis that the model with additional restriction perform just as well as the model without the restriction on all commonly used significance levels. This means we conclude that the increase in likelihood based on our data is significant on the 5% significance level and that we reject the hypothesis that both models are just as likely based on the data.

Table 13 - Likelihood Ratio test results

	Thurstone	Bradley-Terry	Dawkins
Simple model	517	628	888
Home model	524	640	890
Chi-squared(1)	14	24	4
P-value	1.8e-04	9.6e-07	0.0455

Please note that these findings are supported by the example calculations we performed in Section 4.1 and 4.2. The estimated probability that Ajax beats Feyenoord at home increased approximately 30%, which is quite a lot speaking in a soccer setting.

Conclusion

In this research we investigate which of the gamma ranking models, that are closely related to Luce´s choice axiom, measure latent football abilities most accurately when modelling football matches as discriminal Thurstone processes as described by the law of comparative judgment; we considered Thurstone's Case V model, the Bradley-Terry model and Dawkins' model and measured their distance to the real ranking by Kendall's tau. Furthermore, we compared their goodness of fit by computing AIC values. We show that the parameters estimates by Thurstone's model, the normal model, are closest to the real Eredivisie ranking in terms of ranking correlation. Moreover, all three gamma ranking models are well equipped to model football abilities, because of similarities in terms of their difference distribution. In terms of the goodness of fit Dawkins' model, the Laplace model, seems to outperform the other models because of its ability to model fat tails as well as a sharp peak. Thurstone's model lacks the first of those abilities and the model by Bradley and Terry, the logistic model, lacks the second. Additionally, our findings suggest that adding a home advantage effect is a valuable extension to the gamma ranking models in the sense that the football abilities are estimated more efficiently and the likelihood increases significantly when introducing a home effect.

Future work

Further research could investigate how the skill of the individual team players or the number of matches they have played together, or synergy if you will, affect the team skill pre-game. Other extensions include team experience or even dependence among consecutive games. In (Cattelan, 2009) it is shown how to deal with the dependence among the performances of a team. To incorporate these effects one should collect (or construct) the appropriate data and reformulate the likelihoods we specified in Section 3.3. Finally team dependent home parameters can also be investigated. This essentially comes down to modelling the same team, playing at home and playing away, as two separate teams. In our research this has no use as we will get two different parameters per team and therefore ranking a team football ability will not make sense intuitively.

Appendix

A. Proof of the normal-gamma link

We need to prove that $X \sim \Gamma(r, \lambda) \to N(r\lambda^{-1}, r\lambda^{-2})$ if $r \to \infty$. This proof consists out of 2 parts:

1. Proof with moment generating functions that the gamma distribution is infinitely divisible.
2. Use the central limit theorem to show that $X \sim \Gamma(r, \lambda) \to N(r\lambda^{-1}, r\lambda^{-2})$ if $r \to \infty$.

1. Suppose that $X_1, \dots, X_n$ are independent random variables and that X_i has the gamma distribution with shape parameter r_i and scale parameter λ for $i \in \{1, \dots, n\}$. Then $\sum_{i=1}^{n} X_i$ has the gamma distribution with shape parameter $\sum_{i=1}^{n} r_i$ and scale parameter λ.

   Proof:
   Recall that the moment generating function(MGF) of $\sum_{i=1}^{n} X_i = X_1 + \cdots + X_n$ is the product of the MGF's of $X_1, \dots, X_n$, so

$$E(e^{tX}) = \frac{1}{(1-\frac{t}{\lambda})^{r_i}} \frac{1}{(1-\frac{t}{\lambda})^{r_2}} \dots \frac{1}{\left(1-\frac{t}{\lambda}\right)^{r_n}} = \frac{1}{\left(1-\frac{t}{\lambda}\right)^{\sum_{i=1}^{n} r_i}}, \qquad t < \frac{1}{b}.$$

   From this result it follows that the gamma distribution is infinitely divisible.

2. The central limit theorem states that the sum of independent and identically distributed random variables X_i with expected value $E[X_i] = \mu < \infty$ and variance $0 < Var(X_i) = \sigma^2 < \infty$. Then the random variable

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

   converges in distribution to the standard normal distribution as n goes to infinity, that is

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x), \quad for\ all\ x \in \mathrm{R}$$

   Where $\Phi(x)$ is the standard normal CDF.

If we use the fact that the gamma distribution is infinitely divisible, we can decompose $X \sim \Gamma(r, \lambda)$ into a sequence of r independent random variables $(X_1, \dots, X_r)$ such that the sum $S_r = \sum_{i=1}^{r} X_i \sim \Gamma(r, \lambda)$. More specifically we choose $X_i \sim \Gamma(\frac{1}{r}, \lambda)$. When $r \to \infty$ we can apply the central limit theorem. We have:

$$Z_r = \frac{S_r - r\lambda^{-1}}{\sqrt{r}\lambda^{-1}},$$

where Z_r converges in distribution to the standard normal distribution. This means S_r converges to a normal distribution with mean $r\lambda^{-1}$ and variance $r\lambda^{-2}$. X should then also converge to a normally distributed variable with mean$r\lambda^{-1}$ and variance $r\lambda^{-2}$ because S_r has the same distribution as X.

B. Proof of the Gumbel-logistic link

We need to prove that if X and Y are independent and each has the standard Gumbel distribution, then $Z = X - Y$ has the standard logistic distribution.

Proof:

The distribution function of Y is $G(y) = \exp(-e^{-y})$ for $y \in \mathrm{R}$ and the density function of X is $g(x) = e^{-x}\exp(-e^{-x})$ for $x \in \mathrm{R}$. For $z \in \mathrm{R}$, conditioning on X gives:

$$P(Z \le z) = P(P \le X + z) = E[P(Y \le X + z \mid X)]$$

$$= \int_{-\infty}^{\infty} \exp(-e^{-(x+z)})\, e^{-x} \exp(-e^{-x})\, dx.$$

Substituting $u = -e^{-(x+z)}$ gives

$$P(Z \le z) = \int_{-\infty}^{0} e^{u}\exp(e^{z}u)\, e^{z}\mathrm{du} = e^{z}\int_{-\infty}^{0} \exp[u(1 + e^{z})]\, du$$

$$= \frac{e^{z}}{1 + e^{z}}, \qquad z \in \mathrm{R}\,.$$

As a function of z, this is the standard logistic distribution function.

C. Proof of Lemma 3 (IIA)

Lemma 3 (independence from irrelevant alternatives):

$For\ x, y \in S,$

$$\frac{P(x,y)}{P(y,x)} = \frac{P_S(x)}{P_S(y)}$$

Proof:

By Axiom we have

$$P_S(x) = P(x,y)[P_S(x) + P_S(y)]$$

So

$$P_S(x) = P(x,y)[P_S(x) + P_S(y)]$$

$$P_S(x) = P(x,y)P_S(x) + P(x,y)P_S(y)$$

$$(1 - P(x,y))P_S(x) = P(x,y)P_S(y)$$

$$P(y,x)P_S(x) = P(x,y)P_S(y)$$

$$\frac{P(x,y)}{P(y,x)} = \frac{P_S(x)}{P_S(y)}$$

References

Agresti, A. (2002). *Categorical Data Analysis, Second Edition.* Wiley Series in Probability and Statistics.

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.* Budapest: Akadémiai Kiadó.

Bauml, K. H. (1994). Upright versus upside-down faces: How interface attractiveness varies with orientation. . *Percept. Psychophys. 56*, 163–172.

Block, D. H., & Marschak, J. (1960). *Random Orderings and Stochastic Theories of Responses.* Stanford, California: Stanford University Press.

Bradley, R., & Terry, M. (1952). Rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika, 39*, 324–345.

Cattelan. (2009). *Correlation models for paired comparison data. Ph.D. thesis.* Dept. Statistical Sciences, Univ. Padua.

Cattelan. (2012). Models for Paired Comparison Data: A Review with Emphasis on Dependent Data. *Statist. Sci. 27, Number 3*, 412-433.

Cayley, A. (1859). Sixth Memoir on Conics. §209-229.

Daniels, H. E. (1950). Rank Correlation and Population Models. *Journal of the Royal Statistical Society, Ser. B, 12*, 171-181.

Davidson, R. (1970). "On Extending the Bradley-Terry Model to Accommodate Ties in. *Journal of the American Statistical Association, 65*, 317–328.

Dawkins, R. (1969). Animal Behavior. *Animal Behavior 17*, 134-141.

Diaconis, P. (1988). *Group representations in probability and statistics.* Institute of Mathematical Statistics.

Engledrum, P. G. (2000). *Psychometric Scaling: A Toolkit for.* Winchester: Imoteck Press.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal, 29 (2)*, 147–160.

Henery, R. J. (1992). An Extension to the Thurstone-Mosteller Model for Chess. *Journal of the Royal Statistical Society. Series D (The Statistician) Vol. 41, No. 5* , 559-567.

Kendall. (1948). *Rank Correlation Methods.* Charles Griffin & Company Limited.

Luce. (1959). *Individual choice behavior: a theoretical analysis.* New York: Wiley.

Luce, R., & Suppes, P. (1965). *Preference, Utility, and Subjective Probability.* New York: Wiley.

Mallows, C. (1957). Non-null ranking models. . *Biometrika, 44(1)*, 114-130.

Marden, J. I. (1996). *Analyzing and Modeling Rank Data.* CRC Press.

Maydeu-Olivares, & Bockenholt. (2008). Modeling subjective health outcomes: Top 10 reasons to use Thurstone's method. *Medical Care, 46*, 346-348.

Mihram, G. (1975). A generalized extreme-value density. *South African Statistical Journal 9*, 153-162.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology 15*, 72–101.

Stern. (1987). Gamma Processes, Paired Comparisons and Ranking "A Continuum of Paired Comparison Models". *Biometrika*.

Stern. (1990). Models for Distributions on Permutations. *Journal of the American Statistical Association 85(410)*, 558-564.

Thurstone. (1927). A Law of Comparative Judgment. *Psychological Review. 34 (4)*, 273–286.

Thurstone. (1928). "Attitudes Can Be Measured.". *American Journal of Sociology 33*, 529-554.

Torgerson, W. (1958). *Theory and Methods of Scaling.* New York: Wiley.

Wickelmaier, & Choisel. (2007). Evaluation of multichannel reproduced sound: scaling auditory attributes. *Journal of the Acoustical Society of America 121*, 388-400.

Yelott, J. I. (1977). The Relationship between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment, and the Double Exponential Distribution. *Journal of Mathematical Psychology*, 109-144.