भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

**WORKING PAPER NO: 466**

# Rank Consistent Bradley-Terry Models for Repeated Tournaments

**Akshay Kumar Singh**
*Indian Institute of Management Bangalore*
*Bannerghatta Road, Bangalore – 5600 76*
*akshay.singh@iimb.ernet.in*

**Shubhabrata Das**
*Professor*
*Quantitative Methods & Information Systems*
*Indian Institute of Management Bangalore*
*Bannerghatta Road, Bangalore – 5600 76*
*Ph: 080-2699 3150*
*shubho@iimb.ernet.in*

Year of Publication-July 2014

# Rank Consistent Bradley-Terry Models for Repeated Tournaments

Akshay Kumar Singh

Indian Institute of Management Bangalore

Bannerghatta Road, Bangalore 560076 India

email: akshay.singh@iimb.ernet.in

Shubhabrata Das

Indian Institute of Management Bangalore

Bannerghatta Road, Bangalore 560076 India

email: shubho@iimb.ernet.in

**Abstract**

The primary objective of this paper is to model the win-loss records of matches in a repeated tournament using the ranks of the teams. The work proposes modifications of Bradley-Terry (BT) model to make the estimation consistent with the ranks of the participating teams. The BT model with restricted maximum likelihood estimation involves too many parameters and the estimates typically lack strict monotonicity. A proposed class of rank-percentile BT models based on different parametric distribution resolves both the issues. Parameter estimation, goodness-of-fit using suitably framed test statistic and its null distribution, change point analysis in a nested model framework, as well as other estimation aspects are discussed in this article. Adaptive variations of the model that allow estimates to alter are also discussed. For demonstration, National Collegiate Athletic Association (NCAA) men and women basketball tournament data are considered. The discussed models provide excellent fit to the historical data using only a few parameters. The fit validates the ranking procedure implemented by the NCAA. The models can be extended in more general tournament structures, as shown through an analysis of results from the Indian Premiere League. The work done has potential for application in the wider domain of paired comparisons.

KEYWORDS: Knockout Tournament, Ranking, NCAA, Restricted Maximum Likelihood, Rank-percentile BT model, Round-Based Model, Chi-squared goodness-of-fit.

# 1 INTRODUCTION

Various national and international sports tournaments are played to choose the best team from a pool of participating teams or players. Formats such as round-robin, knockout or a mixture of both are usually the design of these tournaments. While the tournaments played over longer durations can have a round-robin design, a knockout format is usually played when there is a time-constraint or there is a large pool of participating teams. The pairing or fixtures in a knockout design can be random or standard. While more on standard and other variants of the knockout designs can be seen in Schwenk (2000), for the current work it is sufficient to note that a prior ranking of the teams is necessary for the design of a standard knockout tournament. The ranking of the teams for these knockout tournaments are usually given by a panel of experts or is based on a computerized rating system. There may be occasional controversies surrounding these rankings, but this work does not examine the process of ranking. Refer to Harville (2003) and Annis and Craig (2005) for the statistical treatment of ranking in specific example and general context of sports respectively. Many domestic tournaments like National Collegiate Athletic Association (NCAA) basketball tournament for men and women have typically been played in a *standard* knockout format.

The broad objective of this paper is to explain the historical results of a tournament repeated over the years using only the seeds or the ranks of the teams. The outcome of NCAA basketball tournament for men and women played over the years has been taken up as a prime case-study. These tournaments constitute of 64 teams which are subdivided into four groups or regions with each group having 16 (ranked) teams. A standard knockout tournament is played within each group. In the current work, outcomes of the games played every year within the four regions are compiled as repetitions of a knockout tournament. The manner in which these historical results are grouped together is detailed in Section 2.

The win-loss data from the games played in sports tournament like NCAA basketball can be seen as a special category of paired comparisons between objects where the comparisons yield

a winner and a loser. Accordingly, a broad framework of Bradley Terry (BT) model (Bradley and Terry, 1952) is adopted in this work. Parameters in a BT model are also referred to as the strengths of the objects (here teams). These strengths are not surrogates for the true abilities of the teams. Instead, they measure how much better or worse a team is, when compared with that of its opponents. Therefore, the modelling approach in this article assumes that the relative strengths of all the rank orders across the repetitions (years and regions) stays same. Hereafter, in this article, *strength* and *relative strength* will be used interchangeably.

The maximum likelihood estimates of strengths under the traditional BT model need not reflect a pre-decided rank order. Hence, various modifications of the general BT model are proposed by linking the strengths of the teams exclusively to their associated ranks. Given the rank order of the competing teams, it is sensible to consider the restricted maximum likelihood estimates, but these estimates typically lack strict monotonicity. Percentiles from various parametric distributions such as lognormal, beta, pareto, weibull, gamma etc. are taken as strength estimates to address this issue effectively. This distribution based modification also invokes model prudence and becomes useful in the context of sparse data, as in the case of knockout tournament. While the distribution based models have performed well with NCAA basketball data, a possible extension is proposed through round based adaptive models that may have greater applicability in other cases. All of these modifications have been discussed in detail in Sections 3.1 - 3.4.

In order to check the validity of the alternative models, a suitable test statistic is formulated by extending the chi-squared goodness-of-fit test. An intuitively appealing clubbing algorithm is implemented to apply the test on sparse data. A simulation study is undertaken to reflect on the null distribution of the test statistic. The details of the testing procedure, simulation and selection criteria are given in Section 3.5. It is observed that the model may change with different forms of replication. Hence, a test for possible change point in the strength structure is discussed in Section 3.6. The change point test can be used to detect a temporal change. However, in this article it is applied to see if identical model parameters are adequate for NCAA men's data vis-a-vis women's data.

The proposed modifications provide an adequate fit for NCAA data. Consequently, it is validated that the strengths of the teams are a function of their assigned rank only and this provides a testimony to the ranking process adopted by the NCAA. The result of the analysis done in the illustrative example of NCAA basketball context is detailed in Section 4. The methodology adopted here can be extended to other tournament designs. As an illustration, we consider results from different seasons of Indian Premiere League. A summary of the article alongwith other possible extensions and applications are discussed in Section 5.

### 1.1 Related Literature

The statistical literature does not talk about inferences related to the modelling of historical data from sports tournaments using the paired comparison models (e.g. BT model), but there is a considerable amount of work on knockout tournaments, and paired comparison model estimation from games of various structural forms. In a racing tournament (e.g. auto racing), the participants are not compared in pairs but a natural extension of the BT model is used by Graves et al. (2003) to draw inferences about driver's abilities. Games like tennis and racquetball involve multiple rounds in a match. Modifying the traditional models to explain the win-loss records in such games is done in Strauss and Arnold (1987). In the context of a knockout tournament, the possibility of a particular structure favoring a definite seed is dealt with in Schwenk (2000). A considerable section of the statistical literature focuses on optimal designs of the sports tournament (Glickman and Jensen, 2005; Graßhoff and Schwabe, 2008) and temporal evolution of the strengths of teams as an attempt towards prediction (Cattelan et al., 2013). Yet another section of the literature concentrates on variations in the BT model to incorporate various changes in the strength structure of the teams. Some notable contributions in this regard are mentioned in the subsequent section.

### 1.2 Bradley-Terry Models

Bradley and Terry (1952) proposed a simple probability model to explain the win-loss record of 'teams' or objects in a paired comparison design of the experiment. We refer to one run of such design as a tournament. The model assumes that a latent variable, $s_i$ expresses the worth or

'strength' of the $i^{th}$ team. Strength of all the teams are expressed as a vector and named as *strength-vector*. A typical strength-vector for $t$ teams is expressed as $S = \{s_1, s_2, s_3, \ldots, s_t\}$. The strength-vector measures the tendency of win of $i^{th}$ team against all the opponents. According to the BT Model, the probability of win of $i^{th}$ team against $j^{th}$ team is given as $\pi_{ij} = \frac{s_i}{s_i+s_j}\ \forall\ i, j \in \{1, 2, \ldots, t\}$. $S$ is estimated by maximizing the likelihood which is same as maximizing the log-likelihood. If the comparisons are independent, the log-likelihood can be written as

$$LL(S) = \sum_{i=1}^{t}\sum_{j=1}^{t} w_{ij}\left(\ln\frac{s_i}{s_i+s_j}\right), \tag{1}$$

where $w_{ij}$ denotes the number of wins of $i^{th}$ team against $j^{th}$ team. It is noted that $S$ is identified only up to a scalar multiple. Hence, it is generally assumed that $\sum\limits_{s_i \in S} s_i = 1$.

The BT models have been extended by making changes to functional form of the expression of $\pi_{ij}$. Agresti (2002) incorporated different winning probabilities for home and away matches. Rao and Kupper (1967) modified the model to incorporate the change in abilities when ties are allowed. The order in which pairs of objects are presented can also alter their winning probabilities. This order-effect has been accounted for in the extension of BT model given by Davidson and Beaver (1977). The present article discusses model-extensions to incorporate rank-order which are decided prior to the tournament.

## 2 DATA

Results of NCAA basketball tournament played over 29 years (1985 to 2013) for men and 20 years (1994 to 2013) for women have been compiled for the analysis. NCAA basketball tournament has a standard knockout design and the participating teams are chosen from a pool of collegiate teams. During the considered years, 64 teams [1] were chosen for the knockout round. The panel considers various intangible factors besides simple win-loss records. Perceived strength of the schedule in the regular season, statistical index (RPI) of the participating teams, AP poll and

[1] There are more than 64 teams in some of the latter years, with some teams playing an additional knockout match for making it into the final 64.

Coaches poll are some of the references that are used by the experts.

The overall ranking of the 64 teams is not publicly announced but the experts create four groups of 16 teams each and announce their ranking within each group. The 16 teams in a group play a standard knockout tournament among themselves. The win-loss records of various rank orders in such a sub-tournament played every year is used for the analysis. The complete rank based study of the win-loss records assumes that all the teams with identical ranks in different groups across years have the same strength. The knockout tournament played within each group selects a winner. The winners of the four groups thereby play the *Final Four*. The result of the final four is not considered in the current study as the relative ranking of the teams in different groups is not known. For each year, four replications of a knockout tournament among 16 rank orders are recorded for men and women. Thus a total of 80 ( four groups for 20 years) replications or brackets for women and 116 (four groups for 29 years) replications for men are analyzed.

The outcomes of all of these replications is reported in Table 1 for women and Table 2 for men. As an illustration of the table, one can see that of the 33(22+11) times rank-1 team has faced rank-2 team for women's tournament, rank-1 has won 22 encounters. It is interesting to note that a 16th-ranked team has never been able to defeat a 1st-ranked team in 116 replications for men. Similarly, no 15th-ranked team has been able to defeat a 2nd-ranked team and no 14th-ranked has been able to defeat the 3rd-ranked team in 80 replications for women.

## 3 MODELS AND THEIR SELECTION

Consider a standard knockout tournament with $t$ teams where $t = 2^k$ for some positive integer $k$. Here, $k$ represents the number of rounds of elimination required to select a winner. In the NCAA college basketball data, $t = 16$. Clearly, the number of rounds to select a winner is four in this context ($k = 4$). Let $\Pi = ((\pi_{i,j}))$ denote the win probability matrix where $\pi_{i,j}$ denote the probability that $i^{th}$ ranked team defeats a $j^{th}$ ranked team ($i \neq j$). The possibility of a tie is excluded in the analysis. Hence, the matrix $\Pi$ can be characterized by

Table 1: Win Matrix: NCAA women college basketball

| rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 22 | 17 | 37 | 21 | 3 | 2 | 39 | 37 | 0 | 1 | 1 | 3 | 0 | 0 | 79 |
| 2 | 11 | 0 | 20 | 3 | 1 | 14 | 43 | 0 | 0 | 25 | 8 | 0 | 0 | 0 | 80 | 0 |
| 3 | 7 | 20 | 0 | 4 | 0 | 37 | 5 | 0 | 0 | 2 | 14 | 0 | 0 | 80 | 0 | 0 |
| 4 | 10 | 3 | 2 | 0 | 35 | 0 | 0 | 0 | 2 | 0 | 0 | 15 | 74 | 0 | 0 | 0 |
| 5 | 3 | 1 | 0 | 23 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 63 | 3 | 0 | 0 | 0 |
| 6 | 1 | 4 | 20 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 57 | 0 | 0 | 0 | 0 | 0 |
| 7 | 1 | 10 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 53 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 10 | 0 | 2 | 0 | 0 | 0 | 0 | 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 9 | 0 | 0 | 23 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 | 1 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 0 | 0 | 0 | 6 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 2: Win Matrix: NCAA men college basketball

| rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 19 | 12 | 29 | 29 | 6 | 5 | 46 | 55 | 3 | 2 | 18 | 4 | 0 | 0 | 116 |
| 2 | 20 | 0 | 23 | 2 | 0 | 21 | 50 | 2 | 0 | 25 | 10 | 1 | 0 | 0 | 109 | 0 |
| 3 | 8 | 14 | 0 | 2 | 2 | 36 | 6 | 1 | 1 | 9 | 24 | 0 | 0 | 99 | 1 | 0 |
| 4 | 14 | 4 | 3 | 0 | 33 | 2 | 2 | 2 | 2 | 2 | 0 | 18 | 91 | 0 | 0 | 0 |
| 5 | 7 | 3 | 1 | 29 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 76 | 11 | 0 | 0 | 0 |
| 6 | 2 | 6 | 27 | 1 | 0 | 0 | 3 | 0 | 0 | 4 | 77 | 0 | 0 | 12 | 0 | 0 |
| 7 | 0 | 17 | 4 | 0 | 0 | 3 | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 1 | 3 | 0 |
| 8 | 10 | 2 | 0 | 4 | 2 | 1 | 1 | 0 | 56 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 9 | 5 | 1 | 0 | 0 | 1 | 0 | 0 | 60 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 10 | 0 | 17 | 3 | 0 | 0 | 2 | 45 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 0 |
| 11 | 3 | 1 | 12 | 0 | 0 | 39 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 |
| 12 | 0 | 0 | 0 | 11 | 40 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 |
| 13 | 0 | 0 | 0 | 25 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 17 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 0 | 7 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

a) $\pi_{ij} \in [0,1]$ and $\pi_{ii} = 0$; $\forall\ i, j \in \{1, 2, \ldots t\}$

b) $\pi_{j,i} = 1 - \pi_{i,j}$ where $i \neq j$

### 3.1 Straight-Rank Model and Traditional Bradley Terry Model

*Straight-Rank-model:* This model directly uses the rank order and does not involve any estimation. However, it is not within BT framework. The win probability matrix $\Pi$ is specified by defining $\pi_{i,j} = \frac{j}{i+j}$. Generally, win proportion of a first ranked team against a second ranked team is expected to be equal. However, in this model the first ranked team has a 66.67% chance of winning against the second-ranked team which may be too high. All the subsequent models are under the BT framework.

*BT models with strength as Reversed-Rank:* The possible limitations of the Straight-Rank-Model is addressed through Reverse-Rank-model. This model lies within the framework of BT where the strength-vector is taken to be the reverse of the rank-vector. More explicitly, $s_i = t+1-i$ where $t$ is no. of participating teams. In this case, $\Pi$ is specified by defining $\pi_{i,j} = \dfrac{t+1-i}{2t+2-i-j}$. Under this model, with 16 teams, the first-ranked team has $\frac{16}{31}$ chance of winning against the second-ranked team, while with 4 teams, the chance is $\frac{4}{7}$.

*BT models based on Pre-decided Strength:* The strength-vector for this BT Model, is decided by the modeler in advance. It incorporates the valued opinion of the modeler and all possible perceptions. In other words, it is an intuitive judgment about the relative strengths of all the teams. The strength-vector is fixed as $[100, 95, 90, 85, 80, 75, 70, 60, 50, 40, 30, 25, 20, 15, 10, 5]$ in the case of NCAA basketball data. There is less difference in the strengths of various rank orders in the tail than in the middle. Alternatively, a modeler can assume that there is more difference in the tail than in the middle and define pre-decided strengths accordingly. It is important that the modeler should propose such pre-decided strengths based on domain knowledge and not using the data. These pre-decided strength vectors are surrogates for a judgment about the spread of the relative strengths of each team.

*BT models with Maximum Likelihood Strengths:* This model requires estimation based on the data. Maximizer of the log-likelihood (1) is taken to be the strength vector for this model. The optimization problem of minimizing the negative of log-likelihood is subject to $s_j \geq 0,\ j = 1, 2, \ldots, t.\ , s_1 = 100$. Huang et al. (2006). Note that the identification of strength-vector,$S$ is done by fixing $s_1 = 100$. Thus, the optimization is reduced to a problem that can be solved using L-BFGS-B algorithms (Byrd et al., 1995). Yet another estimation issue arises when the entire group of objects can be classified into subgroups say A and B, such that none of the members of subgroup A have defeated any member in the subgroup B. The problem in such a case is that there is no maximizer of the likelihood (Hunter, 2004). Such estimation issues can be handled in the Maximum Likelihood based model by increasing the lower bound on the strengths to a fixed $\epsilon\ (> 0)$. As the optimization can be computationally taxing, Hunter (2004) devised a special class of the

EM algorithm to arrive at the maximum likelihood estimate. This specific iterative alforithms is referred to as MM algorithm as it involves Minorization followed by Maximization. The algorithm is implemented in R using BradleyTerry2 package (Turner and Firth, 2010). Accordingly, three different computational pathways have been chosen to arrive at maximum likelihood estimate (MLE) of the strengths in this article, namely:

(a) *Maximum Likelihood with Optim:* Using 'Optim' package in **R**, with L-BFGS-B algorithm and lower bound on the strength estimate as 0.001 .

(b) *Maximum Likelihood with Exponential Strength:* Expressing the strengths as exponential function and then using 'Optim' in **R** with L-BFGS-B algorithm and fixing lower bound on the exponent to be $\ln(0.001)$ .

(c) *Maximum Likelihood by MM:* Using the BradleyTerry2 Package in **R** to arrive at MM algorithm based estimate of the strength.

While the modeler may expect the higher ranked team to have greater strength, the MLE of $S$ may not reflect the same.

### 3.2 BT Models With Restricted MLE Strengths

To have an estimate consistent with the rank order, it is expected that the strengths obey monotonicity constraint, $s_1 \geq s_2 \geq \ldots \geq s_t$. Consequently, it may be prudent to build in this restriction while carrying out the maximum likelihood estimation. The theoretical derivation of the resultant estimated strength is neither straightforward nor may the solution have any closed-form. Therefore, a few solutions in the framework of isotonic Regression (corresponding to different norms) using Pool Adjacent Violators Algorithm (PAVA) algorithm (Mair et al., 2009) were tried out. Eventually, in any given data implementation, the solution having the maximum likelihood among the candidate for monotonised solutions is taken as the Restricted Maximum Likelihood (RML) estimate of strength. While restricted estimation can be modified to have efficient algorithms, it is known and verified that the optimal Restricted Maximum Likelihood (RML) will indeed be on an *extreme direction* defined by the feasible region. This would imply that many of

the inequalities in monotonicity constraint would actually be replaced by equalities in the optimal RML. Equality of strengths across teams with different ranks is a disincentive in using the RML approach.

### 3.3 Distribution based Rank-percentile BT models

A group of distribution based rank-percentile methods ensure a smoother impact of rank on the estimated strengths. In this framework, it is assumed that the strengths of the teams correspond to percentiles of some standard probability distribution. One can either look for uniform percentiles $(\frac{1}{t+1}, \frac{2}{t+1}, \ldots, \frac{t}{t+1})$ which are used in this article or look for some different pre-specified percentiles. These pre-specified percentiles have to decided by the modeler before looking at the data. One or more parameters of such models may not be specified to begin with and consequently these parameters have to be estimated by the maximum likelihood principle. The proposed methods have the additional advantage of model prudence, as they involve very few parameters. A small number of parameters are not a deterrent to a good fit. On the contrary, an excellent fit for these models will be demonstrated in Section 4. This model is also particularly useful in handling sparse data created from a knockout design; e.g. in a standard knockout tournament of 16 teams, rank-1 team plays rank-15 very rarely. Under this model, strengths are specified as

$$s_i = F^{-1}\left(\frac{i}{t+1}\right) \qquad \forall\ i \in \{1, 2, \ldots, t\}\,, \tag{2}$$

where $F$ is the cumulative density function of the concerned distribution and $t$ is the number of teams. Using (2) and (1), the log-likelihood for distribution based models is written as

$$LL(\theta) = \sum_{i=1}^{t}\sum_{j=1}^{t} w_{ij}\left(\ln \frac{F_\theta^{-1}\left(\frac{i}{t+1}\right)}{F_\theta^{-1}\left(\frac{i}{t+1}\right) + F_\theta^{-1}\left(\frac{j}{t+1}\right)}\right). \tag{3}$$

Similar to the approach in MLE, the minimizer of negative log-likelihood for distribution based rank-percentile is calculated using L-BFGS-B method. The change in scale parameters in various distributions multiplies the strength estimates by a scalar and hence need not be estimated. These models can be further categorized on the basis of number of estimated parameters. The models

considered in this article are:

1. No parameter estimated from the data:
    - Triangular distribution
    - Exponential distribution.
2. One parameter estimated from the data:
    - Normal distribution with fixed mean and standard deviation estimated from data.
    - Symmetric Beta distribution.
    - Lognormal distribution with only the shape parameter estimated from the data.
    - Gamma distribution with shape parameter estimated from the data.
    - Chi-Square Distribution.
    - Weibull distribution with shape parameter estimated from the data.
    - Pareto (I) distribution with shape parameter estimated from the data.
3. Two parameters estimated from the data: Asymmetric Beta Distribution.

### 3.4 Models with Round-Based strengths

Each NCAA basketball tournament from which the data has been compiled has four rounds of games. Hence, a round-based strength model which allows the strengths of the teams to vary in each round is postulated. In general, any of the estimated model in the previous sections can be used for this adaptive framework. However, distribution based rank-percentile strength models have a few estimated parameters and strict monotonicity of the strength estimates are maintained. This advantage ensures estimation of the models even in the case of sparse data. Accordingly, only rank-percentile models are considered for this modification. Besides round based variations in parameters, modifications are incorporated in this model to verify if in the round three and round four of NCAA basketball knockout tournament, all the teams start playing with equal strengths. Thus, depending on how the parameters of the considered rank-percentile models are allowed to

vary over different rounds, various alternate models can be proposed. The following round based variants of rank-percentile models are compared in this article:

- Parameter of the distribution is allowed to be different in each round.
- Parameter of the distribution is constrained to be the same in the first two and the last two rounds.
- Parameter of the distribution is constrained to be the same in the first three rounds.
- Parameter of the distribution is constrained to be the same in the first three rounds and in the last round all teams have equal strength $\left(s_i = \frac{1}{2};\ \forall i\ \in 1, 2, \ldots, t\right)$.
- Parameter of the distribution is constrained to be the same in the first two rounds and in the last two rounds all teams have equal strength.
- Parameter of the distribution is allowed to be different in first three (two) rounds and in the last (two) round(s) all teams have equal strength.

### 3.5 Model Selection and Goodness-of-Fit

*Building block of the Test Statistic (TS):* In this section, statistical validation of various models described in section 3.1 through 3.4. is taken up under the broad framework of goodness-of-fit. A simple null hypothesis is $H_0 : \Pi = \Pi^0$ where $\Pi = ((\pi_{ij}))$ and $\Pi^0 = \left(\left(\pi^0_{ij}\right)\right)$, with $\pi_{ij}$ being the conditional probability of win of rank-$i$ over rank-$j$, given that the $i$-ranked team play against the $j$-ranked team. Since this is essentially a test of multiple proportions ($\pi_{ij} = \pi^0_{ij}$ with $1 \leq i < j \leq t$), a natural option is to consider the test statistic, $T = \sum\sum_{1 \leq i < j \leq t} T_{ij}$ with

$$T_{ij} = \frac{(w_{ij} - E_{ij})^2}{E_{ij}} + \frac{(w_{ji} - E_{ji})^2}{E_{ji}} = N_{ij} \frac{\left(\frac{w_{ij}}{N_{ij}} - \pi^0_{ij}\right)^2}{\pi^0_{ij}\left(1 - \pi^0_{ij}\right)} \tag{4}$$

where $E_{ij} = N_{ij}\pi^0_{ij}$. Intuitively, the test statistic is formed on the basis of comparison between empirical probability $(\frac{w_{ij}}{E_{ij}})$ with hypothesized probability $(\pi^0_{ij})$ . For a round-robin tournament,

since $N_{ij}$ is non-random and is the same as $n$ (number of times tournament is played), it is easy to note that $T \to \chi^2_{\frac{t(t-1)}{2}}$ as $n \to \infty$.

This article deals with seemingly more terse problems encountered in a (standard) knockout structure of the tournament. The complexity arises because $N_{ij}$'s are not only typically random, but also $N_{ij} << n$ for some pairs of teams. The individual $T_{ij}$ may still approximately follow a $\chi^2_1$ distribution for sufficiently large $N_{ij}$, but a uniform guideline for largeness of $n$ may not be defined. Moreover, as the actual distribution of $T_{ij}$ is discrete, its approximation to $\chi^2_1$ is susceptible to the typical consideration of the continuity correction. However, in the aggregate form of $T$ (test statistic), the support is lot more dense and consequently the approximation to $\chi^2$ distribution is more effective.

*Clubbing:* At the core of the development of the postulated $\chi^2$ distribution from the binomial win-loss data (tournament outcome) is the issue of normal approximation to the binomial distribution, as in the case of simple test of proportion. For the validity of such an approximation, one needs the expected number in each 'cell' (expected entries in a win-matrix) to be at least five ($E_{ij} \geq 5$). In a standard knockout tournament, certain pairs of ranked teams have less chance of playing against each other and thus they have even lesser chance of recording a win. Consequently, even for large enough $n$, one observes very small $E_{ij}$ ($<< 5$). As is typically done for a chi-square goodness-of-fit test, this requires clubbing of cells.

Let $\Gamma = \{(i,j) : 1 \leq i < j \leq t\}$. Any clubbing operation can be mapped with a partition $\{\psi_1, \ldots, \psi_q\}$ of $\Gamma$ (i.e. $\psi_k \bigcap \psi_l$, $\forall k \neq l$, $\bigcup_{k=1}^{q} \psi_k = \Gamma$ ) such that $\forall$ pairs $(i,j)$ in the same $\psi_k$, the observed and expected data are clubbed together leading to a term $T_k^c$ of the form [analogous to (4)]:

$$\frac{\left(X_k^c - E_k^c\right)^2}{E_k^c} + \frac{\left(\tilde{X}_k^c - \tilde{E}_k^c\right)^2}{\tilde{E}_k^c}, \tag{5}$$

where $X_k^c = \sum_{(i,j)\in\psi_k} w_{i,j}$ ; $\tilde{X}_k^c = \sum_{(i,j)\in\psi_k} w_{j,i}$ ; $E_k^c = \sum_{(i,j)\in\psi_k} (w_{i,j} + w_{j,i})\pi_{i,j}$
and $\tilde{E}_k^c = \sum_{(i,j)\in\psi_k} (w_{i,j} + w_{j,i})\pi_{j,i}$.

Asymptotically (as $n \to \infty$ ) $T_k^c$ has a chi-square distribution with 1 degrees of freedom and eventually test statistic

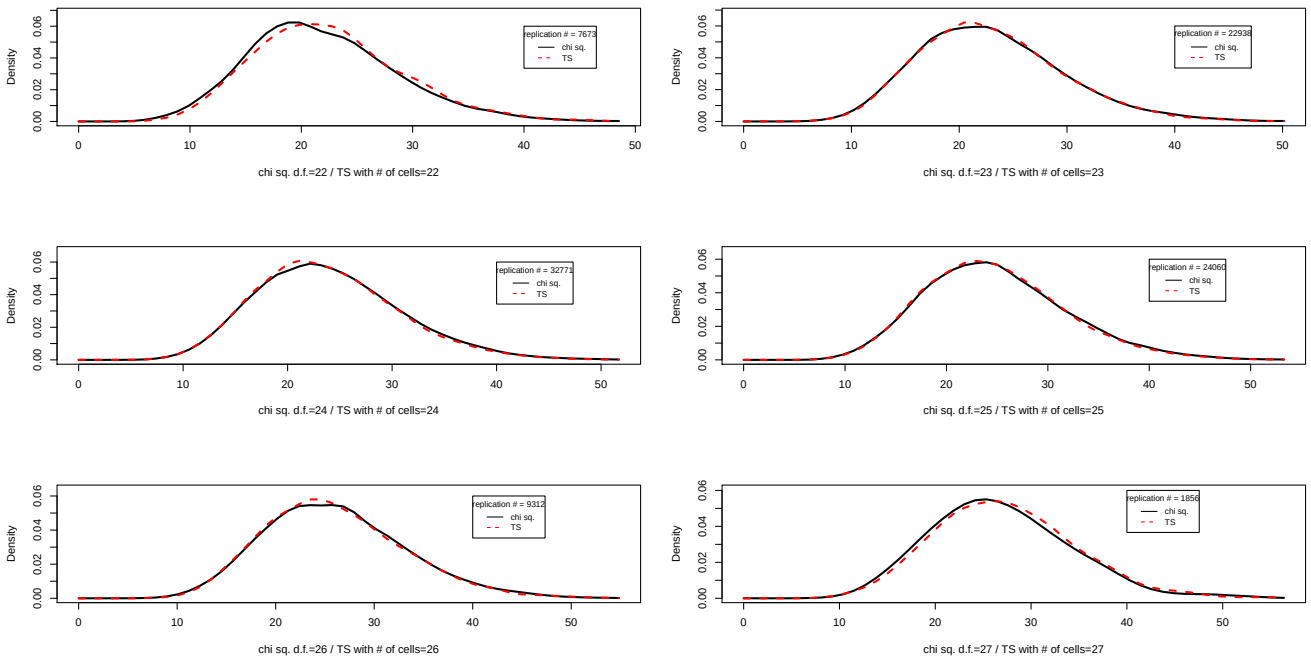
$$T^c \left( = \sum_{k=1}^{q} T_k^c \right) \tag{6}$$

has an asymptotic Chi-square distribution with $q$ degrees of freedom where $q$ the number of non-zero cells left after the clubbing.

The objective of clubbing is to ensure that the expected number of wins is at least five in each post-clubbed cell. The larger the clubbing, the less powerful the test becomes. Purely operational criterion requires minimum clubbing to maximize the number of cells left after clubbing. Such a procedure ensures a larger degrees of freedom but fails to have an intuitive appeal. E.g. $E_{1,15}$ and $E_{2,8}$ are typically very small in a standard knockout for 16 teams and can be clubbed as per an operational criteria, but such a clubbing can't be justified just on the basis of small value of expected wins. Hence, a more intuitive clubbing algorithm is considered in this article. Using the algorithm, only adjacent cells are clubbed and the procedure is symmetric. E.g. if needed, $E_{1,15}$ is clubbed with $E_{1,16}$ and symmetrically, $E_{15,1}$ is clubbed with $E_{16,1}$. The clubbing algorithm proposed is given in Appendix A. Hereafter, $T^c$ represents the clubbed test statistic according to the algorithm adopted here. In order to justify the inferences drawn from the clubbed test-statistic, $T^c$ its distribution is explored through a simulation study.

*Simulation:* It will be seen in Section 4 that strengths from uniform percentiles of a lognormal distribution provides a great fit for both men and women basketball tournament outcome. Hence, a simulation study is undertaken for $n = 100$ and $t = 16$ so that the results are approximately applicable in both of these contexts. Accordingly, the data generating process is chosen to be uniform percentiles, as in (2), of lognormal distribution. The shape parameter sigma in Figure 1 and Figure 2, refers to the standard deviation of log(strength). After each round of simulation, a parameter for the lognormal rank-percentile model is estimated. The estimated strength is then used to calculate an instance of Test-Statistic $T^c$ as in (6) and the corresponding degrees of freedom that corresponds to the number of cells left after clubbing. The degrees of freedom is also

Figure 1: Distribution of Test statistic $T^c$ vis-a-vis $\chi^2$

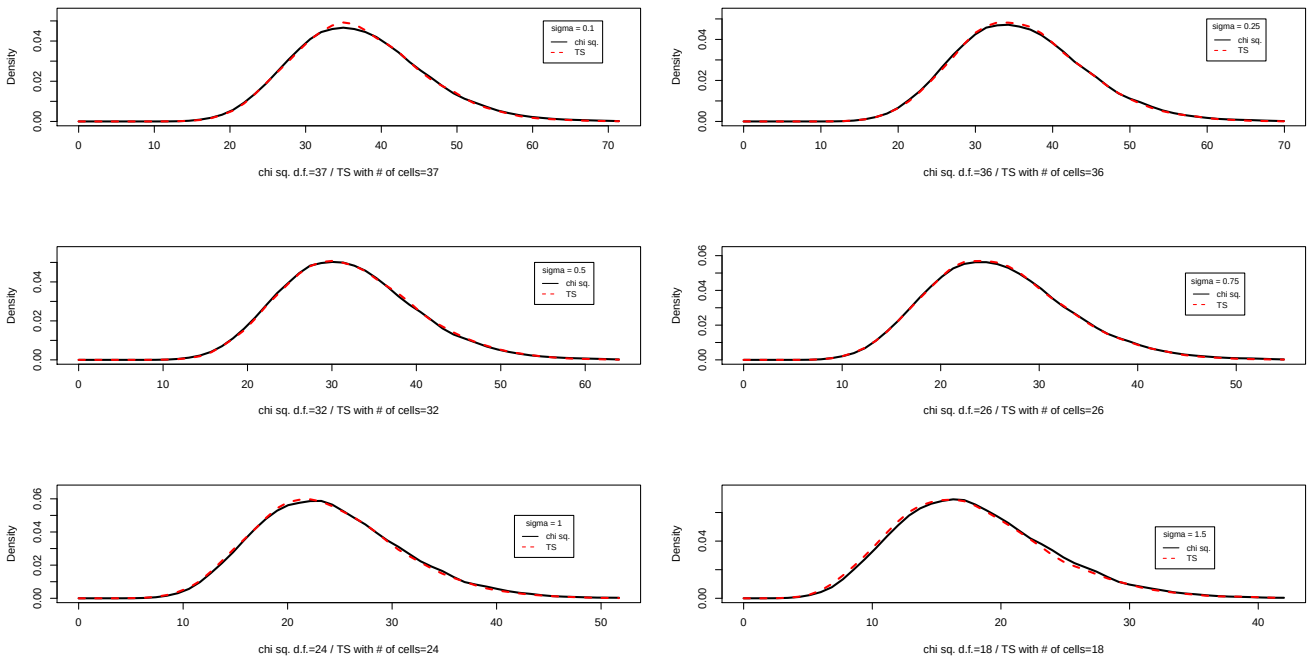



NOTE: Different plots are corresponding to number of clubbed cells which also matches the degrees of freedom. The plots are based on simulation from a rank-percentile BT model based on lognormal distribution

random because of the randomness of $N_{ij}$'s. The distribution of $T^c$ for various degrees of freedom are compared with the true density plot of chi-squared distribution with corresponding degrees of freedom. The near perfect visual fit of the distributions is demonstrated in Figure 1. This validates the p-values obtained in Section 4 to quantify the model fit for real data from NCAA.

For broader applicability, the distribution of Test Statistic, $T^c$, for various parameters of lognormal model is also explored through the simulation. Similar to the approach described in the previous paragraph, the distribution of $T^c$ is explored for varied degrees of freedom and is found to fit well to the chi-squared distribution. For brevity, only the test statistic corresponding to the modal degrees of freedom so obtained from the simulation is visually compared with the corresponding chi-squared distribution in Figure 2. All the considered comparisons pass the traditional goodness-of-fit test. The KS test statistic for the comparisons is found to be very small. Thus, the actual difference between the cumulative density function of the chi-squared distribution and the simulated distribution of TS is typically small. Moreover, the visual fits of the simulated data

Figure 2: Distribution of Test statistic for different shape parameter



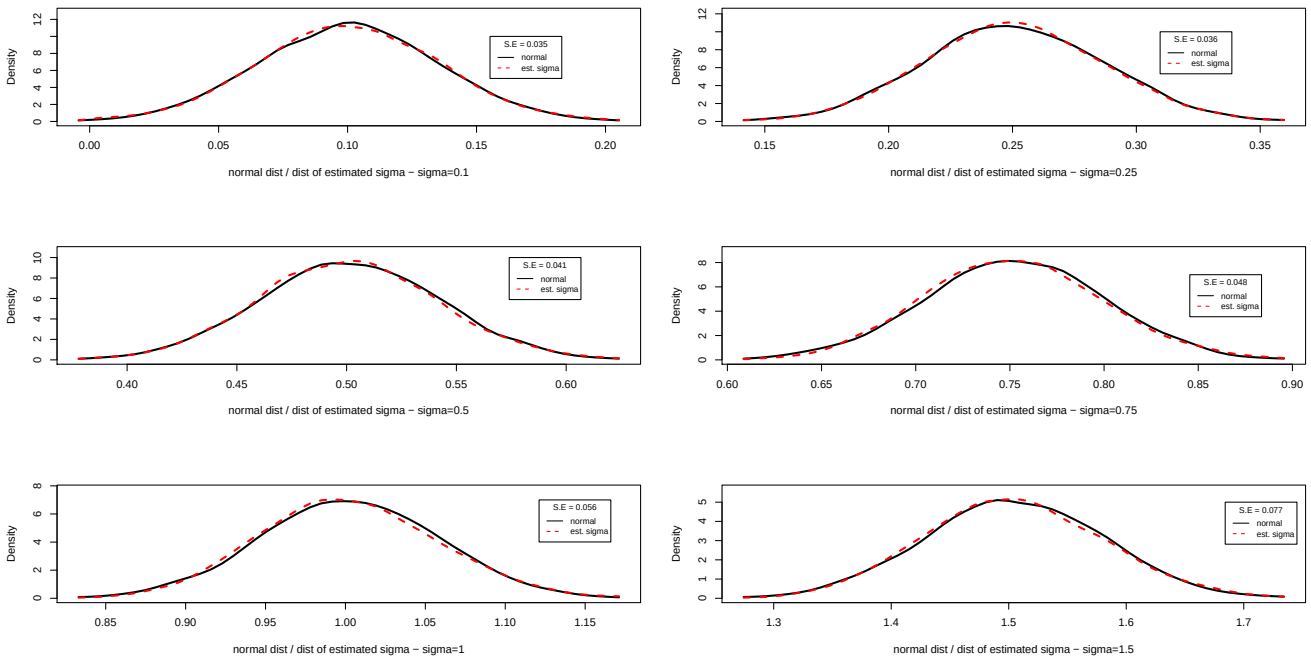

NOTE: Plots are given for modal values of the number of clubbed cells in each case. Simulation is performed for strengths from lognormal distribution, with different shape parameters sigma. The plots are compared to a chi squared distribution

shows that the tail of distribution (essential for p-value calculation) is almost overlapping.

In addition to verifying the applicability of p-values, the simulation study of different parameters of lognormal is used to briefly explore the distribution of the estimated parameter and estimated strengths. It is observed that the estimate of the strength from the lognormal based percentile strength model follows a normal distribution. It is also noted that the standard error of such estimates increase when the true parameter decreases. A visual comparison of the distribution of the estimated parameter with a corresponding normal distribution is reported in Figure 3. With such a variation in the estimated parameter, its impact on the 90% confidence interval of various strengths is shown in Figure 4. Like the parameter estimate, the estimated strength is also found to follow a normal distribution. The distributional properties of such estimates are not explored in this article.

*Alternative model selection criteria:* A Model having higher number of parameters may not be preferable if it provides only a marginally better fit. Typically to address that, model selection

Figure 3: Distribution of the estimated shape parameter vis-a-vis Normal Distribution



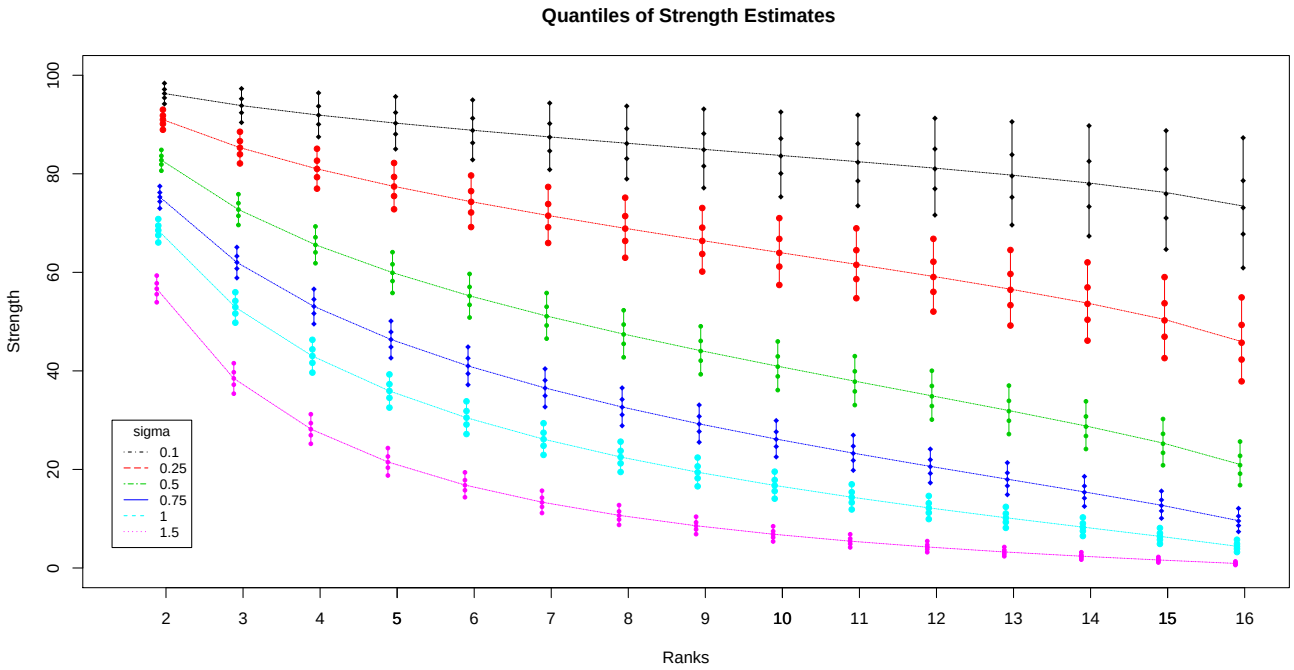

NOTE: This is from simulated rank-percentile BT models based on lognormal distribution with different shape parameters sigma.

criteria $AIC = 2p - 2\ln(L)$ is used where, $p$ is the number of parameters estimated to get the strength-vector and $L$ is the maximized likelihood value for a given model. Alternatively, a model selection criteria which incorporates the sample size ($n$) is given by $AICc = AIC + \frac{2p(p+1)}{n-p-1}$. Lower values of these selection criteria imply a better model.

### 3.6 Change Point Testing

A typical change point test can be done within the nested model framework to explore the possibility of temporal change in structure in any of the model. However, the scope of the change point test is broader. For example, the combined data for men and women NCAA basketball tournament can be analyzed effectively postulating that the same model (with identical parameters) fit both the groups. Alternatively, repetitions for women and men can be analyzed separately to yield two different estimated strength-vectors. The test adopted in this nested framework can be explicitly used to test for a single strength-vector (null) against two different strength vector (al-

Figure 4: Comparison of estimated strength




NOTE:For each rank, 5 percentiles, 95%, 75%, 50%, 25%, 5%, of its strength estimates are indicated in the plot. The mean of the estimates, almost coinciding with the median, is connected by a line plot. Since the estimated strengths are almost unbiased, this also provides a visual comparison of the strengths across ranks for varying shape parameter of lognormal distribution. The strength of the first rank is fixed as 100 and hence not plotted.

ternate). This approach effectively verifies if the same relative strength-vector explains the ranking order in men and women tournament or the vectors are significantly different. The test statistic for the change point test is $2\left(\text{loglikelihood}_{\text{alternate}} - \text{loglikelihood}_{\text{null}}\right)$. The test-statistic follows a chi-square distribution with the degrees of freedom given by the difference in the number of parameters estimated by the two models. The results of this test for various models that provide good fit are reported in Section 4. The test can also be modified to verify if the round based evolving strength estimate is significantly different from the constant strength estimate.

## 4 ANALYSIS

In this section the results of all the models and change point analysis applied to NCAA basketball tournament for men and women are discussed.

A comparative study of three alternative path for computing MLE, as given in Section 3.1, is summarized in Table 3. The table provides details of the fit and the estimates. The strength estimates are found to be approximately equal. Similar strength estimates lead to similar RML as well. Consequently, without loss of generality, all subsequent analysis for maximum likelihood strength models is done using MM algorithm.

Table 3: Comparison of MLE and RML estimates: NCAA Basketball - Men and Women combined

| | ML (optim) | | ML (exp. strength) | | ML (MM alg.) | |
|---|---|---|---|---|---|---|
| | MLE | RML | MLE | RML | MLE | RML |
| Test Statistic | 18.54 | 19.18 | 18.53 | 19.17 | 18.54 | 19.24 |
| p-value | 10.04% | 8.42% | 10.06% | 8.45% | 10.02% | 8.28% |
| neg LL | 1457.76 | 1458.31 | 1457.76 | 1458.31 | 1457.77 | 1458.35 |
| AIC | 2945.52 | 2946.62 | 2945.52 | 2946.62 | 2945.53 | 2946.70 |
| AICc | 2948.19 | 2949.29 | 2948.19 | 2949.29 | 2948.19 | 2949.36 |

NOTE: The degrees of freedom is 12 and the number of parameters is 15, in each case.

*Women:* Results from all the models fitted to the women's data is reported in Table 4. The table reports critical information pertaining to model validation and selection criteria viz. AIC, AICc and p-values along with the minimum negative log-likelihood estimated from the model.

Table 4: Summary of Results: NCAA Women's Basketball

| Model | d.f. | TS | p-value | neg LL | parameters | AIC | AICc |
|---|---|---|---|---|---|---|---|
| Straight-Rank | 20 | 62.77 | 2.6E-06 | 551.89 | 0 | 1103.77 | 1103.77 |
| Reverse-Rank | 23 | 139.19 | 1.3E-18 | 597.39 | 0 | 1194.78 | 1194.78 |
| Pre-Decided Strength | 21 | 125.12 | 8.2E-17 | 589.08 | 0 | 1178.15 | 1178.15 |
| *MLE using MM* | 1 | 7.88 | 0.50% | 509.44 | 15 | *1048.88* | 1056.38 |
| RML | 3 | 24.74 | 1.8E-05 | 522.35 | 15 | 1074.69 | 1082.19 |
| Rank-percentile based on | | | | | | | |
| Normal dist. | 21 | 140.33 | 1.2E-19 | 601.99 | 1 | 1205.98 | 1206.03 |
| Triangular dist. | 22 | 222.74 | 3.8E-35 | 656.01 | 0 | 1312.01 | 1312.01 |
| Symmetric Beta dist. | 18 | 87.40 | 4.2E-11 | 567.44 | 1 | 1136.89 | 1136.94 |
| Asymmetric Beta dist. | 17 | 25.78 | 7.86% | 527.93 | 2 | 1059.87 | 1060.02 |
| Exponential dist. | 21 | 53.57 | 0.01% | 542.29 | 1 | 1086.58 | 1086.63 |
| **Lognormal dist.** | 17 | 14.48 | **63.27%** | 523.67 | 1 | 1049.34 | **1049.39** |
| Gamma dist. | 18 | 25.77 | 10.52% | 527.93 | 1 | 1057.85 | 1057.91 |
| Weibull dist. | 17 | 21.60 | 20.06% | 526.16 | 1 | 1054.33 | 1054.38 |
| Pareto dist. | 16 | 46.85 | 7.2E-05 | 534.33 | 1 | 1070.65 | 1070.70 |

NOTE: Rank-percentile BT models based on lognormal provides a superlative fit to the women's data.

Rank-percentile BT model based on lognormal distribution is found to be the best fit with a largest p-value of 63.27% and minimum AICc value of 1049.39. The AIC criteria for the MLE model is found to be the least. This is because the AIC criteria does not sufficiently penalize the number of parameters. Rank-percentile BT model based on weibull distribution also provides a good estimate of the strength with a p-value of 20.06%. Besides, rank-percentile models based on gamma distribution also has a large p-value of 10.52%. Such high p-values indicate that the ranking procedure for women has been consistent over the years. Though, straight-rank model

does not fit the data well, it fares better than pre-decided strength or rank-percentile models based on triangular distribution.

Table 5: Summary of Results: NCAA men's Basketball

| Model | d.f. | $\chi^2$ TS | p-value | Neg.LogLkhd | # par. | AIC | AICc |
|---|---|---|---|---|---|---|---|
| Straight-Rank | 26 | 56.32 | 0.05% | 942.95 | 0 | 1885.90 | 1885.90 |
| Reverse-Rank | 28 | 71.76 | 1.1E-05 | 952.89 | 0 | 1905.79 | 1905.79 |
| Pre-Decided strength | 28 | 85.28 | 1.1E-07 | 958.05 | 0 | 1916.10 | 1916.10 |
| MLE using MM | 9 | 16.75 | 5.27% | 918.62 | 15 | 1867.25 | 1872.05 |
| RML | 9 | 16.67 | 5.41% | 918.63 | 15 | 1867.25 | 1872.05 |
| Rank-percentile from | | | | | | | |
| Normal dist. | 25 | 64.88 | 2.2E-05 | 947.70 | 1 | 1897.40 | 1897.44 |
| Triangular dist. | 29 | 145.56 | 1.8E-17 | 1002.44 | 0 | 2004.88 | 2004.88 |
| Symmetric Beta dist. | 26 | 66.08 | 2.4E-05 | 947.74 | 1 | 1897.49 | 1897.52 |
| Asymmetric Beta dist. | 24 | 39.80 | 2.25% | 932.17 | 2 | 1868.35 | 1868.45 |
| Exponential dist. | 26 | 49.48 | 0.36% | 935.38 | 1 | 1872.75 | 1872.79 |
| **Lognormal dist.** | 25 | 36.64 | **6.25%** | 931.80 | 1 | **1865.61** | **1865.64** |
| Gamma dist. | 25 | 39.80 | 3.06% | 932.17 | 1 | 1866.35 | 1866.38 |
| Weibull dist | 25 | 40.38 | 2.67% | 932.37 | 1 | 1866.75 | 1866.78 |
| Pareto dist | 25 | 52.48 | 0.10% | 942.35 | 1 | 1886.69 | 1886.73 |

NOTE: Rank-percentile BT models based on lognormal distribution is the best fit as per each of the three criterion. The corresponding strengths are plotted in Figure 5.

*Men:* The comparative study of the model fit for men's data is reported in Table 5. The rank-percentile BT model based on lognormal distribution is again found to be the best fit to the data. It has the largest p-value of 6.25% and minimum AIC and AICc value of 1865.61 and 1865.64 respectively, and may be seen as outperforming MLE and RML because of model prudence. MLE, RML, rank-percentile BT models based on gamma and weibull distribution also provide decent fit with the p-values given by 5.27%, 5.41%, 3.06% and 2.67% respectively. The fit validates the ranking for men. As discussed in Section 3.4, there is an apprehension that indeed the strength

of the various rank order does not stay the same in each round of the tournament. The analysis for the round based model has been carried out for rank-percentile BT model based on lognormal distributions and reported in Table 6. Despite the increase in model complexity and consequently a decrease in the degrees of freedom, the p-value improves to 7.25%. The improvement is seen for the model where the strengths evolve in the first three rounds and in the fourth round it is the same as that of 3rd round. AIC criteria (1864.33) however chooses a model which keeps the strength of the first three rounds as constant and allows the last round to have a different strength.

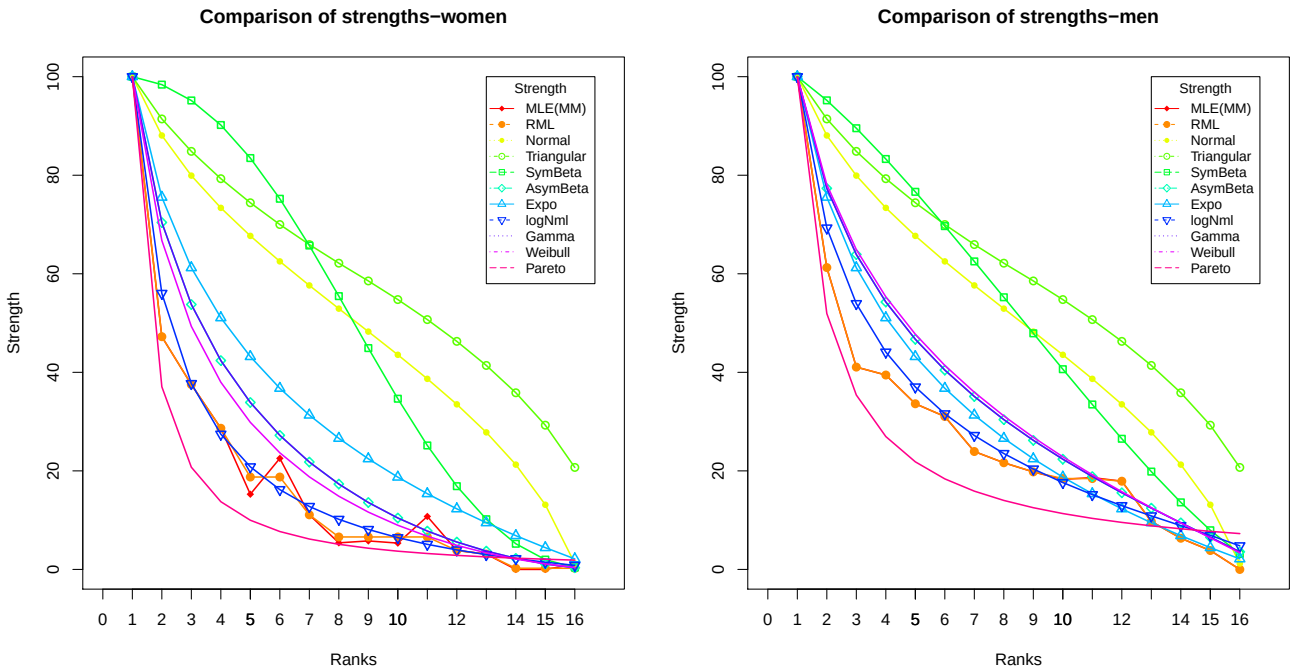
Table 6: Summary for round based BT models: NCAA men's Basketball

| Strength in Round | | | | df | TS | p-value | neg LL | # par | AIC | AICc |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | | | | | | | |
| Est1 | Est2 | Est3 | Est4 | 22 | 33.88 | 5.04% | 928.41 | 4 | 1864.83 | 1865.19 |
| Est1 | Est2 | Est3a | Est3a | 23 | 33.51 | **7.25%** | 931.05 | 3 | 1868.11 | 1868.32 |
| Est1a | Est1a | Est3a | Est3a | 24 | 36.39 | 5.02% | 931.74 | 2 | 1867.50 | 1867.60 |
| Est2a | Est2a | Est2a | Est4 | 24 | 37.49 | 3.90% | 930.16 | 2 | **1864.33** | 1864.44 |
| Est2a | Est2a | Est2a | H | 27 | 43.65 | 2.24% | 932.81 | 1 | 1867.63 | 1867.67 |
| Est1a | Est1a | H | H | 29 | 105.05 | 0.00% | 964.97 | 1 | 1931.94 | 1931.97 |
| Est1 | Est2 | H | H | 28 | 99.62 | 0.00% | 964.27 | 2 | 1932.55 | 1932.66 |
| Est1 | Est2 | Est3 | H | 24 | 40.53 | 1.87% | 931.06 | 3 | 1868.13 | 1868.34 |

NOTE: 'Est$j$' denotes strength estimated for round $j$ . 'Est1a' is the combined estimate of round 1 and 2. 'Est3a' is the combined estimate of round 3 and 4. 'Est2a' is the combined estimate of round 1,2 and 3. 'H' denotes that the strength of all teams are taken to be equal.

*Strength Estimates:* A comparison of the strength estimates for various models for men and women data is shown in Figure 5. The result shows that only the models with non increasing strengths estimate fared better in the comparative analysis done earlier in this section. It is also evident from the plot that the RML estimates flatten the curve where the ML estimates are not found to be consistent with the rank order. The extraordinary fit of the rank-percentile BT model based on lognormal distribution is evident from its proximity with the MLE strength. Other rank-

Figure 5: Strength estimates for various models for NCAA: Men and Women




percentiles BT models based on from gamma, weibull and asymmetric beta are also found to have a closer plot to the MLE curve.

*Combined data of men and women:* A comparative study of all the models for the combined data of men and women is reported in Table 7. Here the MLE model is found to be the best with a p-value of 10%. RML also fares well with a p-value of 8%. Though a good fit of RML also validates the ranking system of NCAA tournament as opposed to men and women differently, the poor fit of the distribution based rank-percentile BT model is not a surprise as it puts further constraints on the strength estimates. The strength estimates of all the models are displayed in Figure 6. Round based models were tried for the rank-percentile BT model based on lognormal distribution. None of these models provide a substantial improvement in the model fit. Hence as discussed in Section 3.6, a change point analysis is done to check if the strength estimates of men and women are significantly different. The (restricted) maximum likelihood estimates and estimates from the rank-percentile BT models based on asymmetric beta, lognormal, gamma and weibull distribution are considered for this change point test and the results are summarized in

Table 8. Invariably in all of the models considered, a low p-value indicates that the null-model (men and women have same strength estimates) is rejected. Hence it is concluded that a separate analysis is preferable for men vis-a-vis women.

Table 7: Summary of Results: NCAA Basketball Men's and Women's combined

| Model | d.f. | TS | p-value | neg LL | parameters | AIC | AICc |
|---|---|---|---|---|---|---|---|
| Straight-rank | 29 | 78.98 | 1.6E-06 | 1494.83 | 0 | 2989.66 | 2989.66 |
| Reverse-rank | 33 | 177.43 | 1.1E-21 | 1550.28 | 0 | 3100.56 | 3100.56 |
| Pre-decided-strength | 31 | 174.39 | 6.6E-22 | 1547.12 | 0 | 3094.24 | 3094.24 |
| **MLE using MM** | 12 | 18.54 | **10.02%** | 1457.76 | 15 | **2945.53** | **2948.19** |
| RML | 12 | 20.10 | 6.51% | 1458.65 | 15 | 2947.31 | 2949.98 |
| Rank-percentile based on | | | | | | | |
| Normal dist. | 32 | 177.05 | 5.3E-22 | 1549.68 | 1 | 3101.37 | 3101.39 |
| Triangular dist. | 33 | 339.54 | 1.4E-52 | 1658.44 | 0 | 3316.89 | 3316.89 |
| Symmetric Beta dist. | 28 | 136.15 | 3.3E-16 | 1524.86 | 1 | 3051.73 | 3051.76 |
| Asymmetric Beta dist. | 28 | 58.80 | 0.06% | 1477.15 | 2 | 2958.31 | 2958.37 |
| Exponential dist. | 30 | 58.25 | 0.15% | 1477.66 | 1 | 2957.33 | 2957.35 |
| Lognormal dist. | 27 | 49.73 | 0.49% | 1472.59 | 1 | 2947.19 | 2947.21 |
| Gamma dist. | 29 | 58.78 | 0.09% | 1477.14 | 1 | 2956.28 | 2956.30 |
| Weibull dist. | 28 | 58.30 | 0.07% | 1476.82 | 1 | 2955.65 | 2955.67 |
| Pareto dist. | 28 | 79.71 | 7.4E-07 | 1491.23 | 1 | 2984.47 | 2984.49 |
| Chi-sq dist. | 29 | 58.78 | 0.09% | 1477.14 | 1 | 2956.28 | 2956.30 |

NOTE: MLE model specification requires 15 parameters to be estimated from the data and hence gives a better fit. The poor fit for the other models is because of the change in structural form of strength in men vis-a-vis women, as seen in change point analysis via Table 8.

## 5 Summary and Conclusion

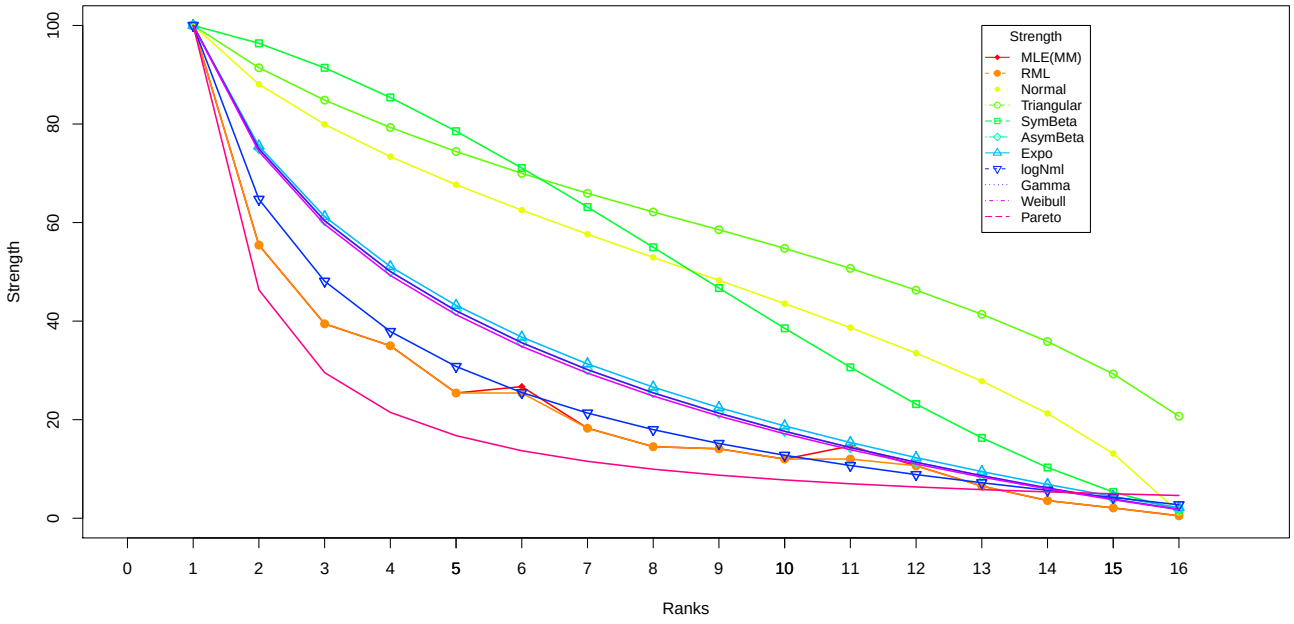
*Summary:* MLE and RML models incorporating the rank order in the BT estimate lack model prudence. RML also fails to provide a strictly monotonic strength. The strict monotonicity

Table 8: Summary of Results — Change point problem: Men vs Women NCAA

| | MLE | Asymmetric Beta | Log-Normal | Gamma | Weibull |
|---|---|---|---|---|---|
| Test Statistic | 59.40 | 34.10 | 34.24 | 34.09 | 36.58 |
| degrees of freedom | 15 | 2 | 1 | 1 | 1 |
| p value | 3.20E-07 | 3.93E-08 | 4.87E-09 | 5.27E-09 | 1.46E-09 |

NOTE: Rank-percentile BT models based on lognormal distribution is one parameter model. Hence the difference in parameters of null and alternate model is one (degrees of freedom for the chi squared distribution of the TS).

Figure 6: Strength estimates for various models for Men and Women combined




restriction and model prudence is addressed by considering the strengths to be percentiles of a suitably chosen parametric distribution. The proposed models with rank-percentile based strength from various parametric distributions provide interpretable estimates for tournaments where a prior ranking is necessary. Besides, the round based modification gives an outline to modify such models to make it adaptive within the framework of specific tournaments with multiple rounds of games. The work develops model selection and validation criteria.

It is seen that the strength estimates from the proposed models satisfactorily explain the win-loss records from NCAA basketball tournament. The observation validates the ranking procedure adopted by the NCAA ranking panel. It is also seen that though the rank order remains the same across various rounds of the tournament, the relative strength of various rank orders evolves through the rounds. A change point analysis done on the data reveals that there is indeed a significant difference between the relative strength of various rank orders of men and women.

The scope of the models is not limited to the NCAA context or to that of a standard knockout tournament. In the context of sports itself, while model validation may not be essential for a valid ranking, the approach can be explored as a sufficient condition for the validate of a ranking. On the other hand, if the teams are not ranked (e.g. in the case of a round robin tournament) one can apply the proposed models using a ranking based on the data. In such a case, a good fit would indicate a simple explanation of the win-loss records. The latter can be seen through short illustration of the IPL data.

*An illustration through the Indian Premiere League (IPL):* The BT model with the various modifications discussed in this work, including the distribution based rank-percentile models can be applied to other tournament formats, including round-robin tournaments. Unlike knockout, the challenge of modeling and statistical inference is less in the round-robin format, as each pair of teams face off an equal number of times. As an illustration, the results from the Indian Premiere League (IPL) — a franchise-based cricket tournament that has been very successful and popular in its six years of existence is considered. In each edition, teams play each other twice in the round robin format, culminating in the top four teams playing in the knockout format to decide the winner. Additional complication is encountered because IPL started with eight teams, with three new teams joining the league in later years while three teams, including two that joined later on, left. Given below in Table 9 is the win matrix, with $(i, j)$ -th element representing number of wins of team in $i$ -th row over team in the $j$ -th column.

Table 9: Win Matrix from Indian Premiere League (IPL)

| | CSK | DC | DD | KKR | KT | KXIP | MI | PWI | RCB | RR | SRH |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CSK | 0 | 6 | 8 | 8 | 1 | 8 | 6 | 4 | 8 | 8 | 2 |
| DC | 4 | 0 | 4 | 2 | 1 | 3 | 4 | 3 | 6 | 2 | 0 |
| DD | 4 | 7 | 0 | 5 | 1 | 5 | 6 | 3 | 5 | 6 | 0 |
| KKR | 4 | 7 | 6 | 0 | 0 | 6 | 2 | 4 | 6 | 6 | 1 |
| KT | 1 | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| KXIP | 4 | 7 | 7 | 5 | 1 | 0 | 7 | 3 | 7 | 3 | 0 |
| MI | 9 | 6 | 6 | 10 | 0 | 5 | 0 | 5 | 7 | 7 | 1 |
| PWI | 2 | 1 | 2 | 1 | 1 | 3 | 1 | 0 | 0 | 1 | 0 |
| RCB | 6 | 4 | 5 | 6 | 2 | 5 | 6 | 5 | 0 | 6 | 1 |
| RR | 5 | 7 | 6 | 5 | 1 | 8 | 5 | 4 | 5 | 0 | 2 |
| SRH | 0 | 0 | 2 | 1 | 0 | 2 | 1 | 2 | 0 | 1 | 0 |

As noted earlier, there is no ranking of teams in IPL. Hence, to implement the rank-based BT models, a ranking is introduced on the basis of % of wins of each team. The rank percentile model based on Lognormal distribution provided an excellent fit to the six-years of IPL result, considering all teams, with a p-value of 0.99 (Chi-square test statistic 11.45 and d.f.=26-1=25). This is highly impressive even while taking into account small number of observations. The estimated strengths of all the teams are reported in the Table 10.

Table 10: Strength of IPL teams: Lognormal Percentile Strength Model

| | IPL TEAMS | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | CSK | MI | SRH | RR | RCB | KXIP | KKR | DD | KT | DC | PWI |
| Strength | 100 | 85.05 | 75.88 | 69 | 63.32 | 58.34 | 53.75 | 49.32 | 44.86 | 40.02 | 34.03 |

NOTE: The estimated strengths are from rank-percentile BT models based on Lognormal distribution.

*Alternative Theoretical Treatise:* The test statistic, $T$ as developed in Section 3.5 can be alternatively framed by replacing $E_{ij}$ with $\tilde{E_{ij}}$ where $\tilde{E_{ij}} = n \times \pi^0_{ij} \times \text{Probability}[i \text{ plays } j]$. The new expected win would typically be smaller because of the knockout structure of the tournament. In practice, it would result into larger clubbing and hence greater information loss, i.e. asymptote would reach a lot later. The advantage of this approach is in having a deterministic number of

repetition and hence deterministic degrees of freedom. The development of such a test statistic and consequent inferences based on it is planned to be taken up in a follow-up study.

*Paired comparison:* Moving beyond sports tournaments and related work to general paired comparison framework, it is known for example that the ranking of consumer items or grading of edibles is done by marketing practitioners and food testers respectively. The problem of authenticating the ranking given by such evaluators or judges can be handled using the procedures outlined in this article. In particular, data from various paired comparisons of the objects ranked by experts can be collected from consumers in a suitable framework and analyzed using the models described in this article. Some similar case studies and related theoretical development of the survey designs, akin to the tournament structures, are planned to be taken up in a follow up work.

## Appendix A Clubbing Algorithm

**Algorithm 1:** Clubbing Algorithm

**input** : A matrix M of size $t \times t$ with many entries less than 5 and a zero-matrix A of same size

**output**: A matrix A of size $t \times t$ with almost all non-zero entries greater than 5

```
1  for i ← 1 to t − 1 do
2      columnsum ← 0 ;
3      rowsum ← 0 ;
4      for j ← t to i + 1 do
5          columnsum ← columnsum +M[i, j] ;
6          rowsum ← rowsum +M[j, i] ;
7          if ( ( rowsum ≥ 5 and columnsum ≥ 5) then
8              A[i, j] ← columnsum;
9              A[j, i] ← rowsum;
10             columnsum ← 0 ;
11             rowsum ← 0 ;
12         end
13     end
14 end
15 repeat all of the above lines with transpose(A) as the new M
```

## References


Agresti, A. (2002), *Categorical data analysis*, Vol. 359 John Wiley & Sons.

Annis, D. H., and Craig, B. A. (2005), "Hybrid paired comparison analysis, with applications to the ranking of college football teams," *Journal of Quantitative Analysis in Sports*, 1(1).

Bradley, R. A., and Terry, M. E. (1952), "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, 39(3/4), 324–345.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995), "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, 16(5), 1190–1208.

Cattelan, M., Varin, C., and Firth, D. (2013), "Dynamic Bradley–Terry modelling of sports tournaments," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1), 135–150.

Davidson, R. R., and Beaver, R. J. (1977), "On extending the Bradley-Terry model to incorporate within-pair order effects," *Biometrics*, pp. 693–702.

Glickman, M. E., and Jensen, S. T. (2005), "Adaptive paired comparison design," *Journal of statistical planning and inference*, 127(1), 279–293.

Graßhoff, U., and Schwabe, R. (2008), "Optimal design for the Bradley–Terry paired comparison model," *Statistical Methods and Applications*, 17(3), 275–289.

Graves, T., Reese, C. S., and Fitzgerald, M. (2003), "Hierarchical models for permutations: Analysis of auto racing results," *Journal of the American Statistical Association*, 98(462), 282–291.

Harville, D. A. (2003), "The selection or seeding of college basketball or football teams for postseason competition," *Journal of the American Statistical Association*, 98(461), 17–27.

Huang, T.-K., Weng, R. C., and Lin, C.-J. (2006), "Generalized Bradley-Terry models and multiclass probability estimates," *The Journal of Machine Learning Research*, 7, 85–115.

Hunter, D. R. (2004), "MM algorithms for generalized Bradley-Terry models," *Annals of Statistics*, pp. 384–406.

Mair, P., Hornik, K., and de Leeuw, J. (2009), "Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods," *Journal of statistical software*, 32(5), 1–24.

Rao, P., and Kupper, L. L. (1967), "Ties in paired-comparison experiments: A generalization of the Bradley-Terry model," *Journal of the American Statistical Association*, 62(317), 194–204.

Schwenk, A. J. (2000), "What is the correct way to seed a knockout tournament?," *American Mathematical Monthly*, pp. 140–150.

Strauss, D., and Arnold, B. C. (1987), "The rating of players in racquetball tournaments," *Applied statistics*, pp. 163–173.

Turner, H., and Firth, D. (2010), "Bradley-Terry Models in R: The BradleyTerry2 Package," *Biometrika*, 714(730), 498.