top

Utrecht University

Tractable inference for latent ability in competitive games:

approximating Bayesian rank-ordered models using a simpler proportion model

Research report
words count: 2446

Marta Sech
Supervised by Erik-Jan Van Kesteren

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences MSc

Contents

Chapter 1

Introduction

In many contexts we come across data that take the form of rankings. This type of data is present, for example, in researches concerning the preferences of individuals for different candidates in elections (H. Stern 1993 Koop and Poirier 1994 or for different consumer goods (Beggs, Cardell, and J. Hausman 1981). In these cases, the aim is usually to investigate individual's choice decisions (Yu, Gu, and Xu 2019). Another area where ranking data are common is in sport and game analytics, where the goal is rather to use game results to estimate the ability of competitors. This is done for example in Kesteren and Bergkamp 2023 with Formula One race results, in Ali 1998 with horses-races or in Graves, Reese, and Fitzgerald 2003 with NASCAR competitions.

If ranking data want to be analysed, models tailored for this typology of data are needed. There are several known parametric models in the literature. One of the most popular is the one proposed by Luce 1959 and Plackett 1975. Rank- Ordered Logit models (ROL) (J. A. Hausman and Ruud 1987) extended it by incorporating covariates.

Although the validity of the ROL models has been widely acknowledged, their execution can be complex in the practice for two reasons. The first one concerns the estimation, which can become computationally burdensome when the number of competitors to be ranked is large: Alvo and Philip 2014 report fifteen competitors as the threshold in this sense. Secondly, the interpretation of parameters is not straightforward, as it requires multiple steps to have them related back to the outcome (refer to Algorithm 1 to understand how the rank of a competitor can be recovered from his ability parameter).

These challenges lead researchers to use alternative models for rating competitors. For example, a linear regression was used in Eichenberger, Stadelmann, et al. 2009, a multilevel approach in Bell et al. 2016 and a Beta regression in S. Stern et al. 2008
The approach that will be investigated in this paper is based on the transformation of the rank order of competitors given in each game into a proportion of competitors beaten for each player. This outcome can then be modeled using a Beta regression with dummies variables representing competitors as predictors. The associated regression coefficients can

then be used to provide the desired rating of competitors. The downside of this approach is that it makes the incorrect assumption that the achieved ranking is (conditionally) independent of the achieved rank of the other competitors in the game. For example, if an averaged-skilled runner enters a local competition, he may ends up in the best positions and achieve a high proportion of competitors beaten. Nonetheless, if the same athlete enters a national level race where only the best contenders compete, his performance will be worse and so its proportion of competitors beaten.

In summary, the Rank-Ordered Logit model and Beta regression both have advantages and disadvantages in their use. Which one is preferable and when is currently unknown. In this project we will try to evaluate this and therefore answer the following research question: under which conditions (if any) is a proportion model a reasonable approximation for a complex rank-ordered model?

This report is structured as follows: first, in Chapter 2 additional background about models for ranking data is provided and Plackett-Luce model is introduced and explained. In the same chapter we also discuss the use of alternative models, with a particular focus on Beta regression. In Chapter 3, the simulation study used to answer the research question is delineated, alongside with the methodology used to compare the two models. In Chapter 4, we present simulations results and insights while in Chapter 5 we apply both models to a real-world example for illustration purposes. We end the paper with conclusions and suggestions for researchers who want to perform inference on ranking data.

Chapter 2

Theoretical background

2.1 Models for ranking data

Literature offers a wide range of parametric models to deal with ranking data, nicely summarised by Critchlow, Fligner, and Verducci 1991. These models are typically explained in terms of judges' preferences for different items. As the framework context of this article is competitive games, in this project games assume the role of judges and the competitors participating in each game represent the ranked items.

If competitors' abilities want to be inferred from ranking data,Bradley and Terry 1952 model can possibly be used. This model is suited to analyse paired-comparison data, for which the probability that a player i wins against another player j p_{ij} is modeled as:

$$p_{ij} = F(\lambda_i - \lambda_j),$$

where F is the logistic function and $\lambda_i = log(\alpha_i)$ for all i. α_i is a positive value representing the ability of the $i-th$ competitor and can therefore be used to rate competitors in head-to-head competitions. Nonetheless, this model can also be applied to a multi-competitor game context where n competitors compete against each other in each game. This is done by converting the ranking of competitors in each game into the $n(n-1)/2$ set of possible paired comparisons between them. Still, this may not represent the most optimal approach, as it would cause a loss of efficiency in the estimation of the ability parameters, as shown in Zhang 2021.

This is the reason why in multi-competitor contexts the most popular choice is to use the Plackett-Luce model, which belongs to the category of "order statistics models", according to the classification proposed in Critchlow, Fligner, and Verducci 1991. Thurstone 1927 can be considered the forerunner of this class of models, which are based on the idea that the position of each competitor in the rank is determined by their value on an unobserved and random variable, referred to as "utility" by Thurstone. The relative positions of competitors on this scale determine their rank: higher utilities are associated with a

better position. This property can be expressed through the following equation:

$$p(\pi) = p(y_{[1]} > y_{[2]} > \cdots > y_{[n]}),$$

where $p(\pi)$ represents an observed ranking of competitors and $y_{[1]}, ..., y_{[n]}$ is the ordered vector of the latent utilities. Thurstone hypothesised that each utility can be decomposed into $y_i = \theta_i + \epsilon_i$ where θ_i is a real constant representing the expected utility of $i - th$ competitor and the ϵ_i represents the error term. He further assumed that ϵ_i are random and independent vectors with a cumulative distribution function corresponding to a standard normal distribution. Later models assumed other types of distributions. While Henery 1983 used a Gamma distribution, Luce 1959 and Plackett 1975, opted for using a Gumbel (type 1 extreme value) distribution, for which utilities have the following cumulative function:

$$F(y_i \mid \boldsymbol{\theta}) = \exp\left(-\exp(-(y_i - \theta_i))\right)$$

This model has become particularly used in numerous extensions and applications since it leads to a closed form for the probability of each observed ranking of n competitors π:

$$p(\pi) = \prod_{i=1}^{n-1} \frac{\exp(\theta_i)}{\sum_{l=i}^{n} \exp(\theta_l)}$$

This form implies that the probability of each ranking can be expressed as the product of the top-choice probabilities, i.e. the product of the probability the first competitor is ranked first among all the competitors times the probability the second competitor is ranked first among all competitors except for the first one and so on. Indeed, the probability p_i that competitor i is ranked first in a set of n competitors, is given by:

$$p_i = \frac{\exp(\theta_i)}{\sum_{i=1}^{n} \exp(\theta_i)}$$

This formulation ensures that the model respects the Independence of Irrelevant Alternatives axiom (IIA)(Tversky 1972). This axiom postulates that the probability that a competitor i outperforms another competitor j in a game is irrelevant from the composition of the group of competitors entering the game.

Rank-Ordered Logit models (J. A. Hausman and Ruud 1987) represents a generalisation of the Plackett-Luce model as they also include covariates, either competitor-specific, game-specific or specific on their combination and interactions. These models can also handle cases where only a subset of players is ranked or when ties exist.

The approach that will be used in this article is inspired by the one used in Glickman and Hennessy 2015 who applied the ROL model to the results given by women's Alpine downnhill skiing competitions recorded over a decade and use them to rate skiers abilities over time. In Glickman article the latent utilities of Thurstone are interpreted in terms of

latent "performances" of players: every competitor in each game produces a performance which is mainly determined by their ability. Higher performances are associated with a better position in the ranking. Under this specification, θ_i can be regarded as the ability of competitors $i = 1, ..., n$.

2.1.1 Alternative choices

Rank-Ordered Logit models represents the gold standard approach for inferring the abilities of competitors from the ranks given in multiple competitions. However, easier to interpret and implement approaches can be used to achieve the same goal.

Sport fans all over the world may just decide to use the ranking of competitors from the last game and use it to compare them. Nonetheless, this simplistic approach would not allow the comparison between competitors entering different games and neither the possibility to investigate the impact of possible factors of interest on their abilities.

Another solution can be the one adopted by Eichenberger, Stadelmann, et al. 2009, who modeled the finishing positions of drivers in Formula One races using a simple linear regression with a dummy variable for each driver and a series of covariates as predictors. The comparison between the regression coefficients associated with each dummy variable can then give insights on the relative abilities of competitors. However, this approach assumes normally distributed residuals, which is often not seen in practice.

The use of a Beta regression model can overcome this problem, as it assumes a more flexible distribution for the dependent variable. S. Stern et al. 2008, for example, applied this model to the victory margins of teams in crickets and use it to evaluate team strengths. This kind of model can be used to deal with outcome constrained to a specific range, typically between 0 and 1, such as proportions.

The approach that will be tested in this project focuses on the transformation of the ranking of players resulting from each game into a proportion of competitors beaten for each competitor in each game. We will then make use of the parameterization of the Beta regression model proposed by Ferrari and Cribari-Neto 2004. This directly models the mean of the distribution, in such a way that the density function is defined as:

$$f(y|\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi) \cdot \Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

where $0 < \mu < 1$ is the mean parameter and $\phi > 0$ is the precision one, with larger values of ϕ associate to smaller variance, for fixed μ. Under this specification, if y_i represents the proportion of competitors beaten by competitor $i = 1, ..., n$, the variance of y_i is defined as:

$$Var(y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi}.$$

The variance is therefore a function of the mean, allowing the model to accommodate situations of heteroskedasticity (Cribari-Neto and Zeileis 2010).
The mean of y_i can instead be expressed as:

$$E(y_i) = \mu_i = g^{-1}(\boldsymbol{x}\boldsymbol{\beta}),$$

where $g()$ is a link function that needs to be strictly monotonic and twice differentiable and $\boldsymbol{x}$ represents the vector of dummy variables each of them associated to $n-1$ competitors and $\boldsymbol{\beta}$ is the vector of associated regression coefficients. A common choice for $g()$ is the logistic function as it makes the parameters interpretable in terms of odds-ratio. If the logit link is chosen, the mean of the distribution of y_i can be defined as:

$$\mu_i = \frac{e^{\boldsymbol{x}\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}\boldsymbol{\beta}}}.$$

The estimated $\boldsymbol{\beta}$ can then be used to rate the competitors with higher coefficients associated with more skilled players.
The underlying assumption of this specification of the Beta regression model is that the dispersion parameter is constant for all competitors. This adds to the generic assumption of independence of observations. Both these hypotheses may not be respected in the context of competitive games, and their impact on the estimation of ability parameters will be evaluated in the current project.

Chapter 3

Methods

To determine under which conditions a Beta regression model can well approximate the Rank-Ordered Logit model in the estimation of the ability parameters in a multi-competitor game context, a model-based simulation study was conducted. We simulated data according to Algorithm 1:

Algorithm 1: ROL-based simulation of rankings of competitors in multiple games

Input:

- number of games n_games;
- total number of competitors n_comp;
- number of entrants per game n_per_game;
- standard deviation of ability parameters σ.

Output: ranking of competitors in n_games

Step 1. A vector of abilities $\boldsymbol{\theta}$ is created, by randomly sampling n_comp values from a $Normal(0, \sigma)$ distribution;

for *each of* n_games **do**

  Step 2. Competitors indices $comp\_ids$ are created by randomly sampling n_per_game values without replacement from the interval (1,n_comp);

  Step 3. A vector Y of n_per_game performance values is generated from a Gumbel($\boldsymbol{\theta}[comp\_ids]$) distribution;

  Step 4. Y is sorted in descending order;

  Step 5. The indices of competitors associated to each performance in the ordered Y generated in Step 4 provides the ranking of competitors in the game;

end

Step 6. Rankings of competitors from each game are combined together;

Factors that were changed in the simulation are: total number of competitors, number of entrants per game, number of games and standard deviation of abilities. A summary of the different levels used for each of them is given in Table 3.1.

Factor	Levels
Total number of competitors	n_comp= 15, 50, 250, 500
Number of entrants per game	n_per_game = 2, 8, 20, 40
Number of games	n_game = 20, 50, 250, 1250
Standard deviation of abilities	σ = 0.1, 1, 10

Table 3.1: Different levels used for each factor in the simulations.

- Total number of competitors
  The first level (n_comp=15) represents the threshold mentioned by Alvo et Philip (2014) referring to the point after which the ROL model becomes too computationally intense, the second one (n_comp=50) represents the number of competitors typical of real world competitions such as ski races, and the remaining two levels, (n_comp=250 and n_comp=500) aim to simulate contexts where the number of potential competitors is considerably large, as is the case in online games.

- Number of entrants per game
  The tested levels for this factor replicate the number of competitors competing against each others in different sports. For example, n_per_game =2 simulate the outcome of a head-to-head competition such as a tennis match. The second one (n_per_game =8) reassemble an athletics competition; the third one (n_per_game= 20) approximate the number of entrants in a Formula One race and the last level (n_per_game= 40) the number of cars competing in NASCAR races.

- Number of games
  This factor constitutes the sample size as each game represents one observation. The uncertainty around the estimate of the ability parameters for both models is expected to decrease for increasing number of games.

- Standard deviation of abilities
  Different variances were chosen to imitate different scenarios ranging from situations in which players are characterised basically by the the same ability (σ=0.1) to context where they significantly differ (σ=10), as is often the case in sports where the same competitors are winning every time.

For each combination of these levels 1000 dataset were generated.
The Rank-Ordered Logit model and the Beta regression one were then fitted to each dataset. Parameters were estimated using a Bayesian approach through the use of Stan statistical modeling language (Team et al. 2018). Bayesian estimation methods were chosen

because they better adapt to model ranking data. Indeed, as stressed by Guiver and Snelson 2009, maximum likelihood estimation algorithms for Plackett-Luce model may not reach convergence and lead to over-fitting in presence of sparse data.
Performance of both models was evaluated by looking at the mean squared error between the known ability parameters and the estimated ones. The estimates were provided by the mean of the posterior distributions of the parameters and 95% credible intervals were also computed to give a measure of the uncertainty around them. An example of a graphical representation of how the comparison between the two models could look like is given in Figure 3.1.[1]

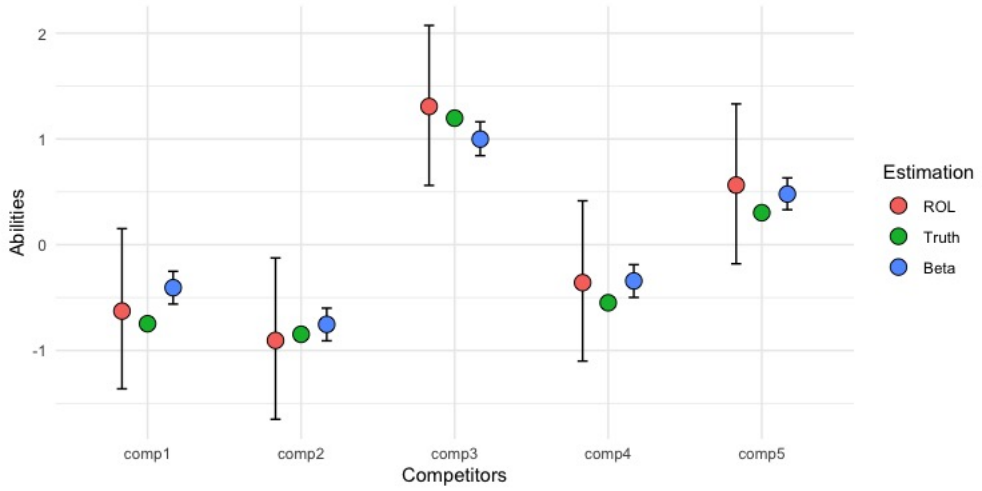

Figure 3.1: True abilities and ROL and Beta estimated ones with 95% credible intervals for five competitors

The analysis was conducted using package "CmdStanR" (Gabry and Češnovar 2021) in R studio version 4.2.1 (Posit team 2023).

[1]The example dataset used for this representation simulated 5 competitors competing against each other in 100 games with a standard deviation for the ability parameters equals to 1.

Bibliography

Ali, Mukhtar M (1998). "Probability models on horse-race outcomes". In: *Journal of Applied Statistics* 25.2, pp. 221–229.

Alvo, Mayer and LH Philip (2014). *Statistical methods for ranking data*. Vol. 1341. Springer.

Beggs, Steven, Scott Cardell, and Jerry Hausman (1981). "Assessing the potential demand for electric cars". In: *Journal of econometrics* 17.1, pp. 1–19.

Bell, Andrew et al. (2016). "Formula for success: multilevel modelling of Formula One driver and constructor performance, 1950–2014". In: *Journal of Quantitative Analysis in Sports* 12.2, pp. 99–112.

Bradley, Ralph Allan and Milton E Terry (1952). "Rank analysis of incomplete block designs: I. The method of paired comparisons". In: *Biometrika* 39.3/4, pp. 324–345.

Cribari-Neto, Francisco and Achim Zeileis (2010). "Beta regression in R". In: *Journal of statistical software* 34, pp. 1–24.

Critchlow, Douglas E, Michael A Fligner, and Joseph S Verducci (1991). "Probability models on rankings". In: *Journal of mathematical psychology* 35.3, pp. 294–318.

Eichenberger, Reiner, David Stadelmann, et al. (2009). "Who is the best Formula 1 driver? An economic approach to evaluating talent". In: *Economic Analysis and Policy* 39.3, p. 389.

Ferrari, Silvia and Francisco Cribari-Neto (2004). "Beta regression for modelling rates and proportions". In: *Journal of applied statistics* 31.7, pp. 799–815.

Gabry, Jonah and Rok Češnovar (2021). "cmdstanr: R Interface to'CmdStan'". In: *URL: https://mc-stan. org/cmdstanr, https://discourse. mc-stan. org*.

Glickman, Mark E and Jonathan Hennessy (2015). "A stochastic rank ordered logit model for rating multi-competitor games and sports". In: *Journal of Quantitative Analysis in Sports* 11.3, pp. 131–144.

Graves, Todd, C Shane Reese, and Mark Fitzgerald (2003). "Hierarchical models for permutations: Analysis of auto racing results". In: *Journal of the American Statistical Association* 98.462, pp. 282–291.

Guiver, John and Edward Snelson (2009). "Bayesian inference for Plackett-Luce ranking models". In: *proceedings of the 26th annual international conference on machine learning*, pp. 377–384.

Hausman, Jerry A and Paul A Ruud (1987). "Specifying and testing econometric models for rank-ordered data". In: *Journal of econometrics* 34.1-2, pp. 83–104.

Henery, Robert J (1983). "Permutation probabilities for gamma random variables". In: *Journal of applied probability* 20.4, pp. 822–834.

Kesteren, Erik-Jan van and Tom Bergkamp (2023). "Bayesian analysis of Formula One race results: disentangling driver skill and constructor advantage". In: *Journal of quantitative analysis in sports* 19.4, pp. 273–293.

Koop, Gary and Dale J Poirier (1994). "Rank-ordered logit models: An empirical analysis of Ontario voter preferences". In: *Journal of Applied Econometrics* 9.4, pp. 369–388.

Luce, R Duncan (1959). "On the possible psychophysical laws." In: *Psychological review* 66.2, p. 81.

Plackett, Robin L (1975). "The analysis of permutations". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 24.2, pp. 193–202.

Posit team (2023). *RStudio: Integrated Development Environment for R*. Posit Software, PBC. Boston, MA. URL: `http://www.posit.co/`.

Stern, Hal (1993). "Probability models on rankings and the electoral process". In: *Probability models and statistical analyses for ranking data*. Springer, pp. 173–195.

Stern, Steven et al. (2008). "Ranking international limited-overs cricket teams using a weighted, heteroskedastic logistic regression with beta distributed outcomes". In.

Team, Stan Development et al. (2018). *Stan modeling language users guide and reference manual, version 2.18. 0*.

Thurstone, Louis L. (1927). "A Law of Comparative Judgment". In: *Psychological Review* 34, pp. 273–286.

Tversky, Amos (1972). "Elimination by aspects: A theory of choice." In: *Psychological review* 79.4, p. 281.

Yu, Philip LH, Jiaqi Gu, and Hang Xu (2019). "Analysis of ranking data". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 11.6, e1483.

Zhang, Sanqian (2021). "Building upon Bradley-Terry and Plackett-Luce: some methods for modeling paired comparison and rank order data". PhD thesis. Harvard University.