**Comparing Two or More Ranking Procedures**

By

Ross Larsen

Wayne Larsen

Tanner Johnson

Abstract

In the United States newspapers sport sections often have listings of the latest rankings of college or high school sports teams. During the month of December these weekly rankings may include college football, college basketball (both men and women), college volleyball, and local high school teams. Typically, college rankings list the top 25 teams in each sport. Normally such rankings are determined by a poll of coaches or sports writers; a totally subjective procedure.

Some authors have suggested mathematical methods for ranking teams eliminating the subjective nature of the polls.

The purpose of this paper is review two such mathematical methods **and to introduce a method for comparing two such methods.**

Pairwise comparisons are useful in: (a) ranking sports teams, (b) taste comparisons, and (c) employee ratings . It may have applications in other fields as well such as education such as in admitting students to graduate school via faculty vote. The LARC system was developed by Larsen and Allen (date) and has desirable properties for a pairwise comparison systems. Namely it: (a) has no subjectivity, purely based on mathematical principles, (b) has applications to any pairwise comparison framework, such as sports teams, chess, and the other applications listed above. (c) handles schedules that are unbalanced, (d) is ‘directionally invariant;’ that is an object of comparison can never lose ranking for winning or gain ranking when it loses, (e) is symmetric in wins and loses, (f) provides a ranking for all objects (e.g., not just the top 25), and is (h) practical, it makes intuitive sense.

This paper will use the sports teams terminology even though the pairwise comparisons can be used in many other contexts. We do include an example of employee rankings.

The LARC system uses a Bayesian approach. Each team is assumed to have a ‘strength.’ The larger the team’s strength, the more likely the team will win an individual game. Traditionally, LARC has used the Bradley-Terry model (Bradley & Terry, 1952) for the probability that team i will beat team j as follows.

**Probability ( team i beats team j) = Si / ( Si + Sj)**

where Si and Sj are the strengths of team i and team j respectively.

An alternative model is the Thurstone-Mosteller (Thurstone, 1927; Mosteller, 1951),

**Probability ( team i beats team j) = F( Si - Sj)**

where **F** is the cumulative standard normal distribution.

Note for the Bradley-Terry situation the strengths are non-negative numbers; for the Thurstone-Mosteller model the strengths can be either positive or negative. Of course, these strengths are unknown; this is where the Bayesian approach is used to estimates the strengths based on the won lost records and schedules of each team.

The purpose of this article is to see which of the two modeling schemes (Bradley-Terry vs. Thurstone-Mosteller) in the LARC system perform the best with (a) simulated data, and (b) real-world sports data. Specifically the hypotheses will be:

1. Given a small schedule (20 games) simulated assuming the Bradley-Terry model how much does the choice of model used in the LARC system affect the results?
2. Given a small schedule (20 games) simulated assuming the Thurstone-Mosteller model how much does the choice of model used in the LARC system affect the results?
3. Given a full season of unbalanced sports team simulated data (i.e., modeled after basketball) which of the models perform better?
4. Given a full season of real sports team data (i.e., NBA basketball) which of the models perform better?
5. For the sports analogy we will include a home-court advantage term in the posterior to see if it improves performance with real data.

Performance will be defined two ways: (a) as the ability to correctly predict a comparison given that it is removed from the dataset given the rest of the data for the real and simulated data and (b) the bias and variance of the estimated strengths of the objects in the simulated data.

**Literature Review**

Bradley and Terry’s (1952) method of pairwise comparisons has been in use for over 50 years in such methodologies such as the Larsen-Allen Ranking by Computer (LARC) scheme (Allen, 1979). It involves the gamma distribution and has the likelihood function as mentioned previously.

The Thurstone-Mosteller model (Thurstone, 1927; Mosteller, 1951) on the other hand assumes a gaussian distribution of the strengths as mentioned previously. The Thurstone-Mosteller model has been used in fields such as psychology and statistics and in preferential learning (Brochu, de Freitas, & Ghosh 2007) and in the ranking of chess players (Elo, 1978).

Handley (2001) compared to the two methods in regard to image quality assessments and concludes that the Bradley-Terry Method has better mathematical properties allowing for more statistical procedures (such a comparing groups through hypothesis testing) to be easily implemented. Cattelan (2012) compares both methods and points out strengths and weaknesses of both pointing out that the Thurstone-Mosteller model performs well with dependent data and discusses the packages that are available for the implementation of these methods.

Others abandon the standard parametric approaches entirely and attempt to solve the problem through more flexible models such as B-splines (Winsberg & Ramsay, 1981).

Nevertheless, to our knowledge, there has never been an attempt to compare the two methodologies in regard to: (a) bias, and (b) predictive ability in sports-like scenarios. It will inform the literature to test these two methodologies on real world data as well as simulated data sets to make a case which one performs the best.

References

Allen, A. S. (1979). *The ranking of teams competing in pairwise comparisons*. Brigham Young University.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired

comparisons. *Biometrika*, *39*(3/4), 324-345.

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data.

*Statistical Science*, *27*(3), 412-433.

Elo, A. E. (1978). *The rating of chess players, past and present*. Arco Pub.

Eric, B., de Freitas, N., & Ghosh, A. (2008). Active preference learning with discrete choice data. In

*Advances in neural information processing systems*(pp. 409-416).

Handley, J. C. (2001, April). Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired

comparison models for image quality assessment. In Proc. IS&T *PICS* Conf. (Vol. 1, pp. 108-112).

Mosteller, F. (1951). Remarks on the method of paired comparisons: I. The least squares solution assuming

equal standard deviations and equal correlations. *Psychometrika*, *16*(1), 3-9.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, *34*(4), 273.

Winsberg, S., & Ramsay, J. O. (1981). Analysis of pairwise preference data using integrated B-splines.

*Psychometrika*, *46*(2), 171-186.