1 Summary

I have implemented and tested a rating system analogous to the Elo system used in chess.

The United States Chess Federation (USCF) is enthusiastic about their adoption of the Elo system for several reasons: they hold invitational tournaments for various national championships, they use it to help select national teams for Olympiads, they award national titles (Master and Senior Master) based on ratings, and finally, they believe that the rating system encourages players at all levels to compete more often in tournaments. Similarly, international titles such as International Master and International Grandmaster depend on a version of the ELO system used by the International Chess Federation (FIDE).

2 Introduction

There are many sources that rank teams in competitive sports, for example coaches polls in college football. Ranks give an ordering, who's first, who's second, etc. The presumption is that if player or team A is ranked higher than B, then A is more likely to win against B than to lose. How much more likely? Ranks don't answer that question. Rating systems do.

Ratings systems exist for chess, go, table tennis, scrabble, soccer and other sports. Master points in duplicate bridge are not a rating system - once earned, master points never decrease. Rating systems are especially important in competitions with unbalanced schedules. In a double round-robin tournament like The English Premier League season, ratings will typically agree with the league standings.

Rating systems are an example of what is known as a latent-variable model: we don't directly observe the strength of the players, we observe the results of matches between participants. A rating system uses the results of matches to estimate the player's strength. If A has rating R_A, and B has rating R_B, then some function of those ratings predicts the probability that A beats B. Rating systems imply ranks: if $R_A > R_B$, then player A is ranked above player B.

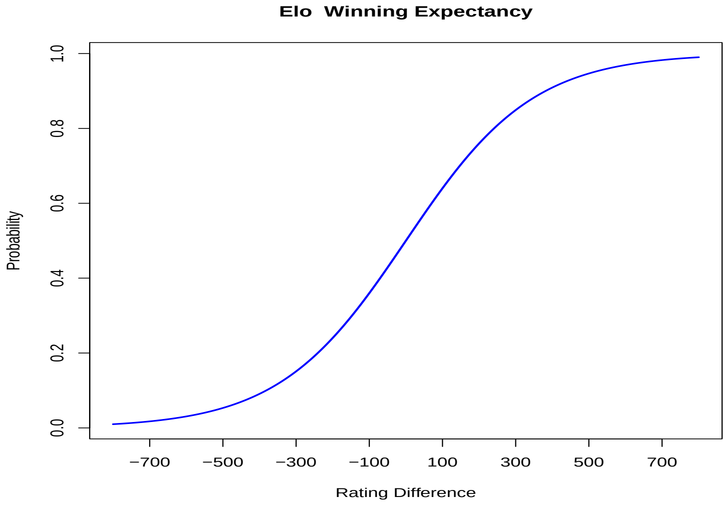
Arpad Elo, a physicist and chess player, created a statistically based rating system for chess in the 1950's. Since computers were not generally available, he made numerous assumptions and approximations to simplify the computations. The Elo system updates ratings using the simple formula

$$R_{new} = R_{old} + K(W - W_E)$$

where W is the observed number of wins and W_E is the expected number, computed using the logistic model

$$P_{A,B} = \frac{10^{(R_A - R_B)/400}}{1 + 10^{(R_A - R_B)/400}}.$$

Here is a graph of the logistic curve underlying the Elo system used in chess:



The curve shows the probability that a player will beat an opponent based on the rating difference. If the difference is near 0, the probability is near 1/2. If the player is rated 200 points higher than their opponent, the probability is about 3/4. Conversely, if the player is rated 200 points lower than their opponent, the probability of winning is about 1/4.

The Elo system was simple to implement with a hand calculator, which was important in the pre-computer era. The United States Chess Federation adopted the Elo system in 1960, and FIDE, the international chess federation, adopted it in 1970. The international soccer federation (FIFA) now usues it to rate national teams.

The Elo system does have significant disadvantages. First, it makes no provision for the differing levels of uncertainty in different players' ratings.

Second, the update does not depend on the whole network of results, but only on the results of a single participant's matches. Suppose that A beats opponent B who is much higher rated, but B has, in the same tournament, also lost to several other weaker players. A gets credit in the Elo system for beating a much stronger player, even though the results of the other matches suggest that B is not playing at that level. To put it in statistical terms, the Elo system ignores the covariances between parameter estimates.

With the advent of modern computational power, we are no longer constrained to use a system designed for hand calculators.

3 Pétanque

The FPUSA has adopted a rating system based on the logistic model:

$$\mathbb{P}(A \text{ beats } B) = \frac{e^{\beta(R_A - R_B)}}{1 + e^{\beta(R_A - R_B)}}$$

or equivalently,

$$\log\left(\frac{P_{A,B}}{1 - P_{A,B}}\right) = \beta(R_A - R_B).$$

where β is a scale factor, which is typically chosen to make the ratings easier to interpret, and R_A and R_B are the ratings of two participants. I chose $\beta = \log(2)/100$, so that a 100 point rating difference corresponds to $2:1$ odds in favor of the higher rated player. In other words, we expect a player rated 100 points higher than an opponent to win 2/3 of the time. Because odds are multiplicative, a 200 point difference corresponds to $4:1$ odds, a 300 point difference corresponds to $8:1$ odds, etc.

One may estimate the ratings R_i by various methods. For example, maximum likelihood and Bayes methods are both easy to implement with modern computers. What does that mean? Imagine that there are only two players, A and B. They play 100 games. If each wins 50 games, then our best guess is that $\mathbb{P}(A \text{ beats } B) = P_{A,B} = 1/2$, or $R_A = R_B$. If A wins 67 and loses 33, then we guess that $\mathbb{P}(A \text{ beats } B) = P_{A,B} \approx 2/3$. Using the scale parameter $\beta = \log(2)/100$, then our best guess for the rating difference is $R_A - R_B = 100$. With many players, and many results, we do essentially the same thing: pick the ratings that best fit the observed results.

Maximum Likelihood estimation has the disadvantage that anyone with all wins or all losses can't be rated. While that may not be a fatal flaw, a more principled system is to start everyone with a seed rating called a prior distribution, and update according to Bayes Theorem. That is the system I have implemented and tested.

Finally, note that in this model there is no absolute scale: only differences between ratings matter. Thus we can pick arbitrary values for the average rating and scale without changing the model.

In light of the fact that pétanque tournaments may be singles, doubles, or triples, the logistic model needs to have three alternative forms:

Singles Two players: A and B

$$\log\left(\frac{P_{A,B}}{1-P_{A,B}}\right) = \beta(R_A - R_B).$$

Doubles Four players, $A1$ and $A2$ on Team A, $B1$ and $B2$ on Team B:

$$\log\left(\frac{P_{A,B}}{1-P_{A,B}}\right) = \beta((R_{A1} + R_{A2})/2 - (R_{B1} + R_{B2})/2).$$

Triples $A1$,$A2$ and $A3$ on Team A, $B1$, $B2$ and $B3$ on Team B:

$$\log\left(\frac{P_{A,B}}{1-P_{A,B}}\right) = \beta((R_{A1} + R_{A2} + R_{A3})/3 - (R_{B1} + R_{B2} + R_{B3})/3).$$

While those formulas base the probabilities on the average ratings of teams, it is possible that using the sum of the team ratings instead of the average may fit better. That is an empirical question that remains to be answered. Using the sum instead of the average would tend to compress the rating scale. Mark Glickman, a statistician friend at Harvard who is working on similar problems, agrees with my decision to use the average.