European Journal of Operational Research 000 (2016) 1–6

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

An empirical Bayes model for time-varying paired comparisons ratings: Who is the greatest women's tennis player?

Rose D. Baker, Ian G. McHale*

Centre for Sports Business, Salford Business School, University of Salford, Salford M5 4WT, UK

ARTICLE INFO

Article history:
Received 23 March 2016
Accepted 17 August 2016
Available online xxx

Keywords:
Bradley–Terry
Paired comparisons
Ranking
Sports
Shrinkage

ABSTRACT

We present a methodology for fitting a time-varying paired comparisons model using an empirical Bayes approach. The model simultaneously avoids two problems that typically arise with paired comparisons data: first, that extreme values of estimated strengths can occur for competitors appearing in and winning a small number of games, producing absurd rankings, and second, that the time-varying strengths 'balloon' over time. The empirical Bayes approach automatically shrinks the strength estimates towards the mean, thus avoiding both issues. We present our model and demonstrate its use in the setting of tennis in search of an answer to the question: who is the greatest women's player of all time. Our results suggest that Steffi Graf is a strong candidate, but, using confidence intervals on the rankings themselves, others cannot be ruled out.

1. Introduction

Baker and McHale (2014) presented a methodology to estimate time-varying ratings for paired comparisons and used the model to answer the question: "who is the greatest men's tennis player in the Open Era?". Here we address the 'sister' question: "who is the greatest women's tennis player in the Open Era?". Although answering this question would be of great interest to sports fans, the main intellectual novelty in the current paper lies in the improvement of the underlying model used to estimate time varying strengths.

One might think that the task of ranking women tennis players would be very similar to that of ranking men players. However, the characteristics of women's tennis, and the resulting differences in the data set of results, mean that a more robust model is needed. Specifically, in women's tennis, matches are best out of three sets instead of best out of five, reducing the amount of data considerably. With less data, estimating time-varying strengths becomes more challenging, and hence one cannot simply take software written for studying men's tennis, and use it for studying women's tennis.

The improvement in the modelling approach is needed to deal with several issues arising from the characteristics of the women's game, and these issues are not unique to fitting ratings models to sports data. The first problem arises as a result of competitors playing different numbers of matches. This means that there is more uncertainty in the estimate of strength for a competitor playing in relatively few matches, compared to a competitor playing in many matches. As a result, we may have the perverse situation where a player with just a few victories is rated as being better (but with a larger standard error) than a player with a marginally lower win rate, which was however achieved over many more matches. The second related issue is caused by players winning (or, more likely, losing) all of their matches so that the estimate of strength tends towards infinity (or zero). Indeed, Hunter (2004) decided to drop any player from the data set winning or losing all of their competitions from the estimation. This is clearly an unsatisfactory strategy. These two issues are general problems for paired comparisons models and have long been a thorn in the side of analysts fitting ratings models. A third issue, specific to fitting time varying comparisons models, is that of 'ballooning'; because the strengths of competitors are relative, fitted strengths can move up and down arbitrarily over time, and this must somehow be prevented. If left unaddressed, these issues can result in the analyst obtaining a rankings table with unlikely, unexpected and possibly spurious ratings.

Our solution to these problems is to assume that all player strengths come from some 'prior' distribution of competitor strengths. Taking this approach deals with all these problems. Further, the intuition behind the approach makes sense in the setting of sport: assuming that competitors are drawn from a population in which ability itself is a random variable having some distribution among players is realistic: some players will have a higher

* Corresponding author.
E-mail address: i.mchale@salford.ac.uk (I.G. McHale).

http://dx.doi.org/10.1016/j.ejor.2016.08.043
0377-2217/© 2016 Elsevier B.V. All rights reserved.

strength than the mean, whilst others will be weaker than the mean strength; but most will have near average strength. As we observe the players competing and winning and losing matches, we will be able to ‘update’ the estimate of their strength as we get a better understanding of what their underlying ability is.

This approach is a type of ‘shrinkage’ (Maritz & Lwin, 1989) in that we take an estimate of a player’s strength (the prior) and update it in the light of information (match results). In comparison to the estimate of strength that would be obtained without using a prior, the empirical Bayes estimate is ‘shrunken’ towards the mean of the prior distribution. Of course, the amount of shrinkage decreases as more evidence is gathered that the player is different from the average.

Assuming a prior distribution for parameters is nothing new and is the essence of Bayesian statistics itself. We should point out however that our approach is frequentist since in the empirical Bayes method, part of a hierarchical prior distribution is estimated from the data. Indeed, empirical Bayes is an accepted part of the frequentist toolset. Baker and McHale (2015) present an empirical Bayes’ methodology for use in the case of a static paired comparisons model. However, the situation here is made much more complicated than would normally be the case because we are looking to estimate *time-varying* strengths for each competitor. As such, there is no single strength for each competitor, rather there is a ‘line’ of strengths.

Unlike here, much of the previous work on the analysis of sporting results has used stochastic models of strengths in order to rank the competitors. Glickman (1993) presented a dynamic Bradley–Terry model for chess and Glickman (2001) presented a state-space model which allowed for the mean and the variance of the evolution process to be stochastic and demonstrated the model by rating National Football League teams and chess players. Knorr-Held (2000) used the Kalman filter to estimate dynamic ratings for sports teams. These types of stochastic models are representative of what happens in team sports, where individual players come and go and the resulting change in performance could be modelled as a random process. However, for sports like tennis, individuals compete, and there is a strong deterministic component to the evolution of their strength, which typically peaks and then falls off slowly towards retirement. Although models which allow for a stochastic evolution of strengths can of course be used to model individual sports, the use of a ratings model that allows for a deterministic evolution of player strengths seems more natural, and is the methodology we adopt here.

The paper is organised as follows. In the next section, we describe the basic time-varying paired comparisons model, the empirical Bayes modifications to the basic model, and the procedure for estimating the parameters of the model, including the idea of connectivity. Section 3 presents a simple idea of calculating confidence intervals on rankings. Our data set for the women’s Grand Slam tennis is described in Section 4, before the results of our model, and model diagnostics are presented in Section 5. Some conclusions are given in Section 6.

2. Time-varying model

As in Baker and McHale (2014), the basic building block of our model is the continuum of paired comparisons models, first presented in Stern (1990), but first expressed in terms of the distribution function of the beta distribution by Baker and McHale (2014). The probability that player i beats player j is given by

$$p_{ij} = B(\beta, \beta)^{-1} \int_0^{\alpha_i/(\alpha_i+\alpha_j)} y^{\beta-1}(1-y)^{\beta-1}\,\mathrm{d}y, \tag{1}$$

where player i’s strength is α_i, β is a parameter to be estimated and B denotes the beta function. For $\beta = 1$, the model reduces to the familiar Bradley–Terry model, whilst as $\beta \to \infty$ the Thurstone–Mosteller model is obtained. The over-lying unit of victory in tennis is the match. However, there are smaller units of competition: a point, a game and a set. As Baker and McHale (2014) did for men’s tennis, we use the unit of victory as the set. This means information is retained in the data regarding the margin of victory (2–0 in sets suggests a stronger performance than 2–1). However, we do not use the game as the unit of victory because this can result in counter-intuitive results. For example, a player may win a match 7–6, 1–6, 7–6. If the game were used as the unit of victory, then the winner would be deemed to have a lower estimated strength of the two competing players given that the loser, in fact, won more games than the winner. Using the set score (2–1) does not have this weakness.

The time-varying strength is modelled using the barycentric rational interpolant (Baker & Jackson, 2014; Berrut & Trefethen, 2004), so that the strength of player i at time t is given by

$$\alpha_i(t) = \frac{\sum_{k=1}^{n_i} w_{ik}\lambda_{ik}/(t-t_{ik})}{\sum_{k=1}^{n_i} w_{ik}/(t-t_{ik})} \tag{2}$$

where λ_{ik} is the kth fitted strength of player i, i.e. the strength at time t_{ik}. To differentiate between $\alpha_i(t)$ and λ_{ik}, we call the latter the *tabulated strength*. Of course, at time t_{ik}, the two are the same. There are n_i such nodes for player i, and we use weights of order zero such that $w_{ik} = (-1)^k$.

One might wonder whether strengths could be forecast using (2). There is a small amount of work on forecasting using splines (e.g. Harvey & Koopman, 1993), so it is possible that an analogous method could be developed using the barycentric method. However, our focus here is on the use of the method for interpolating and smoothing noisy data.

2.1. Node allocation

Our first improvement on the Baker and McHale (2014) methodology comes in the allocation of nodes to each player. A large number of nodes results in over-parameterisation, whilst if there are too few nodes, the model cannot respond to the changing strengths of players appropriately.

Rather than use the complicated and somewhat ad-hoc formula in Baker and McHale (2014), we propose a simpler algorithm, which in our tests provides better results: specify N, the required total number of nodes in the model (the total for all players in the model). Then if s sets were played in total, there should be a node for every s/N sets played. Of course, s/N must be at least unity, which means that the actual number of nodes allocated will exceed N. This system means that many players who played rarely only have one node and are assumed to have a constant strength, whereas players who played a lot have more nodes. For players with more than one node, nodes were regularly spaced in time to include the first and last match dates for that player. We discuss how we found the optimum value of N, the total number of nodes, in Section 2.3 below.

2.2. Empirical Bayes model extension: shrinkage

The second and major contribution to the literature here is to adopt an empirical Bayes methodology whereby we assume player strengths are random variables drawn from some underlying distribution. After experimentation with different prior distributions, it was decided that the prior mean strength of each player should be a random variable from the log-normal distribution, but of course, the methodology presented here can be used with other prior distributions, as discussed at the end of this section. We now set up the mathematical terminology of our empirical Bayes methodology in terms of tennis.