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

Baker, RD and McHale, IG

http://dx.doi.org/10.1016/j.ejor.2016.08.043

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

Rose D. Baker and Ian G. McHale*

Centre for Sports Business, Salford Business School, University of Salford, UK.

August 12, 2016

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.

Keywords: Bradley-Terry, connectivity, paired comparisons, ranking, rating, sports, shrinkage.

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

*email address: i.mchale@salford.ac.uk

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 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 and 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 esti-

mating 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 (Berrut and Trefethen, 2004, Baker and Jackson 2014), 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 and 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.

Let there be n_p players in total in the dataset. Let the lth player have n_l tabulated strengths $\lambda_{l1} \cdots \lambda_{ln_l}$, with log-mean $y_l = \ln\{\sum_{k=1}^{n_l} \lambda_{lk}/n_l\}$.

Let there be n_m matches in total in the dataset, and let the winner of the jth match be player number i_{j1}, the loser number i_{j2}. The numbers of sets won are respectively $s_{i_{j1}}, s_{i_{j2}}$, where $s_{i_{j1}} < s_{i_{j2}}$. Let $P_j(\boldsymbol{\lambda}_{i_{j1}}, \boldsymbol{\lambda}_{i_{j2}})$ be the probability that player i_{j1} wins, where $\boldsymbol{\lambda}$ denotes the vector of tabulated strengths. Of course, P is simply a function of the two player strengths as interpolated from the $\boldsymbol{\lambda}$ values at the time of the jth match. The log-likelihood is then

$$\ell_0 = \sum_{j=1}^{n_m} \{s_{i_{j1}} \ln P_j + s_{i_{j2}} \ln(1 - P_j)\}.$$

The key idea that allows us to obtain time varying player strengths which are shrunken, is to assume that the log-means y_l are normally and independently distributed $Y_L \sim N[\mu, \phi^2]$. The posterior likelihood is then obtained by multiplying the likelihood by the product of the prior distributions for each player's mean log-strength (the normal probability densities). The logarithm of the posterior pdf is then

$$\ell = \ell_0 - \frac{1}{2} \sum_{l=1}^{n_p} \{\frac{(y_l - \mu)^2}{\phi^2}\} - \frac{n_p}{2} \ln \phi^2 - \frac{n_p}{2} \ln(2\pi). \tag{3}$$

Our procedure is empirical Bayes because we estimate the 'prior' parameters μ and ϕ from the data.

The maximum of (3) with respect to μ is trivial and is obtained when $\hat{\mu} = \sum_{l=1}^{n_p} y_l/n_p$. Note that one of the strengths needs to be fixed, because (3) is invariant under addition of a constant to the (logged) strengths (scaling of the unlogged strengths).

For estimating ϕ the situation is more complicated since $\ell \to \infty$ as $\phi \to 0$. Thus, to estimate ϕ we need to carry out the integration of the profile likelihood (having used $\hat{\mu}$ in (3)) to obtain a finite value for $\hat{\phi}$.

We use Laplace's method to approximate the integrations. First, we use a Taylor's series expansion of ℓ about the optimum λ values that maximise ℓ, which we denote by λ^*, so that

$$\ell_0 \simeq \ell_0(\lambda^*) + \frac{1}{2} \sum_{l=1}^{n_p} \partial^2 \ell / \partial y_l^2 (Y_l - y_l)^2, \tag{4}$$

where the derivatives are taken at the λ^* values; $\partial^2 \ell / \partial y_l^2 = \partial^2 \ell_0 / \partial y_l^2 - 1/\phi^2$ and $\partial^2 \ell / \partial y_l^2 < 0$.

Then integrating the posterior pdf over the Y_l we obtain the marginal likelihood

$$\ln \mathcal{L} = \ell(\lambda^*) - \frac{1}{2}\sum_{l=1}^{n_p} \ln\{1 + \phi^2 |\partial^2 \ell / \partial y_l^2|\}. \tag{5}$$

Here the Y_l have been assumed independent. To check whether this was a reasonable assumption, we calculated the average correlation resulting between bootstrapped estimates. We used the parametric bootstrap so that results of matches were simulated from the fitted model parameters, and a new set of bootstrap parameters were estimated. We found that the average correlation was only 0.0001 suggesting that the assumption of independence is a reasonable one.

Our procedure for estimating the time varying strengths is as follows:

1. estimate μ using $\hat{\mu} = \sum_{l=1}^{n_p} y_l / n_p$.
2. optimise (5) to estimate ϕ.
3. using the estimated value of ϕ, maximise (3) with respect to the tabulated strengths.

Steps 2 and 3 above are reasonable since for a given value of ϕ^2 the curvatures can be regarded as constant, so we simply maximise (3). For the women's data set described below, we find the estimate of $\phi \simeq 0.4 \pm 0.01$.

As mentioned above, we use a log-normal prior distribution. We experimented with other priors generated by imagining a series of say 2γ fictitious sets, in which a player wins γ and loses γ sets against the average player with strength ξ. Use of this prior shrinks players' strengths towards the mean ξ. Under the Bradley-Terry model, a strength y would give rise to a term $f(y) \propto (y\xi)^\gamma/(y+\xi)^{2\gamma}$. If this is to be a pdf, we require $\int_0^\infty f(y)\,\mathrm{d}y = 1$. Changing the variable of integration, the integral can be evaluated to obtain

$$f(y) = \frac{\xi^{-1}(y\xi)^\gamma}{B(\gamma+1, \gamma-1)(y+\xi)^{2\gamma}},$$

where B denotes the beta function. Hence $\frac{\gamma+1}{\gamma-1}(y/\xi)$ follows the F-distribution with $2\gamma+2$ and $2\gamma-2$ degrees of freedom. This prior distribution has the advantage that its parameter γ has a simple interpretation, but the great disadvantage of computational complexity. In our experiments, the results were very similar to when the log-normal was used suggesting the rankings are quite robust to choice of prior. However, since using the log-normal distribution prior is much less computationally demanding it is the prior we adopt here.

2.3 Optimising the model parameters, β and N

The values of β, the parameter of the continuum of paired comparisons model in (1), and N, the total number of nodes in the model, were estimated iteratively.

The likelihood-maximisation procedure described above was run for various values of β. The value that resulted in the highest value of the likelihood was $\beta \simeq 2.35$.

To estimate N we calculated the Akaike Information Criterion (AIC) and selected the value of N which resulted in the minimum AIC. As the number of nodes N varies, the node placement also changes for players with more than one node. Hence the AIC changes somewhat jerkily as N increases. To alleviate this unwanted noise in the procedure, we first calculated the AIC for ten chosen values of N. We then used a quadratic regression to identify the relationship between N and the AIC. From this relationship we identified the value of N which corresponded to the minimum AIC. This procedure was done iteratively with the process for finding β. The minimum AIC was found when $N \simeq 1600$.

2.4 Connectivity - do players fall into disjoint sets?

A hitherto undiscussed assumption of this type of time-varying paired comparisons model is that the competitors are 'connected'. To introduce the idea of connectivity consider an example of three players, A, B and C. The players may 'connect' directly or indirectly. For example, an indirect connection between A and C is that A plays B who plays C; A and C are connected via B. A direct connection between A and C is of course if A plays C. If not all players are connected in a dataset, then there are two or more disjoint sets of players. Thus, the strength of the players in one set could not be measured against that of players in the other sets, and so the strengths of all players could not then be compared and the results of our time-varying paired comparisons model would be meaningless.

As far as we know, this concept is new in being applied to paired comparisons models and there is no standard way to check or measure connectivity. Here we propose checking connectivity using a simple methodology borrowed from computer science. First, one starts with the first player in the list of matches, and her name is stored in a type of two-ended list called a *deque*. All players she played with are added to the end of the list, then the algorithm proceeds by adding those additional players with whom the second player in the list played. Eventually, all players who played together directly or indirectly are in the list. In this case, the list then contained all 1123 players, so all players are connected.

If the players fall into two or more disjoint sets, the data can give no clue as to how strong each set is relative to the other(s). Only the prior term ties the strengths together. In future work on paired comparisons models, this check of connectivity should be performed, otherwise the results may be spurious. This problem is a special case of the more general problem of identifiability, i.e. of whether a parameter could theoretically be estimated, given a very large amount of data (e.g. Casella and Berger 2002).

3 Rankings confidence intervals

In previous work on time-varying rating of players, for example, Baker and McHale (2014), results of model fitting are given in terms of an estimated maximum strength (α_i) and a standard error. To obtain the final "all-time rankings" the players are ordered in terms of the point estimate of the maximum strength achieved by the player. This is of course absolutely typical of any quantitative analysis in which parameters have been estimated from data. However, in the case of rankings, the standard error on the point estimate of strength is particularly difficult to interpret. Here, we instead evaluate a confidence interval for the ranking itself. In doing so, it is possible to see just how well the players are differentiated from one another.

Calculating these confidence intervals on ranking is not as straightforward as it may initially seem. The procedure we propose here is to compute the confidence intervals using the parametric bootstrap replicates. Each player's rank was found for each replicate, and the confidence limits read off from the sorted array of ranks for a player. This was easy to compute since the parametric bootstrap was already being computed to calculate standard errors on player strengths.

Several modelling choices have been made including node allocation and placement, and choice of prior distribution, and as these are varied, the resulting rankings will change. In particular, we found that the ranks of weaker players were more sensitive to these choices. Somewhat reassuringly, the ranks of the very top players (indeed those that are of primary interest here) were much less sensitive. Of course, the main reason for the sensitivity of the ranks of the weaker players to modelling choices is likely to be shortage of data. We therefore believe that the ranking confidence intervals will capture this

uncertainty on ranks.

4 Data

We obtained data on the results of women's tennis matches in the four Grand Slams: the Australian Open, The French Open, Wimbledon and the US Open, since the Open Era of tennis began in 1968 to the Australian Open in 2016. Unlike for the men's game, results had to be sourced from several locations including `www.tennis-data.co.uk` for the later years (since 2000), and `www.tennis24.com` for earlier years.

A problem with merging data from different sources was that one player may appear in the merged database under two different names. To deal with this, and other issues required a lot of effort. To detect potential errors, a naïve name conversion program was written which resulted in many error flags which needed to be investigated further. The errors were mainly of three types:

1. A player had married and changed her name, sometimes temporarily. For example, Chris Evert became Chris Evert-Lloyd and subsequently reverted to her maiden name. To address this problem, hyphenated names were flagged up and checked.

2. Names were sometimes spelt slightly differently across (and within) data sources. To address this, the Damerau-Levenshtein edit distance (Damerau, 1964) was computed, and names not more than 2 characters different were flagged up and checked.

3. Only a player's initial was given. For example E. Makarova could be Ekaterina or Elena. Flags were raised for players with long playing histories and subsequently investigated. Some of them turned out to be more than one player.

Any detected error was investigated and corrected. This type of 'data soaping' is an essential but little-reported aspect of data analysis. Finally, the data quality was checked against lists of aggregate results by player on wikipedia. For example, our match result database could be used to calculate the number of finals reached by a player and this could be compared to the equivalent figure on wikipedia.

In total, we have results of 21,921 matches including 1123 players playing 46,864 sets. Unlike Baker and McHale (2014) and other authors before them, we do not need to discard players who played in only a handful of matches. This is a consequence, and major advantage, of the empirical Bayes approach we adopt here.

Before presenting our model results, Table 1 shows the top players, ranked by number of Grand Slam titles won. Steffi Graf tops the list with 22 titles. Serena Williams is in second, but it is worth noting Margaret Court's record - a success rate of 50% in the 22 tournaments entered in the Open Era. However, the whole point of using a ratings model is to account for the strength of the opposition. It is possible that some players in this list benefited from 'weak' eras with little quality opposition, whilst others may not have won as many as their true ability should have awarded them with because they

competed in an era of great strength.

Table 1: Greatest women's tennis players according to the number of Grand Slam titles won in the Open Era (since 1968 to the Australian Open in 2016).

Rank	Name	Australian Open	French Open	Wimbledon	US Open	Total	Win Span	Open era entries	Titles per entry
1	Steffi Graf	4	6	7	5	22	1987-1999	56	0.39
2	Serena Williams	6	3	6	6	21	1999-2015	61	0.34
3	Chris Evert	2	7	3	6	18	1974-1986	56	0.32
3	Martina Navratilova	3	2	9	4	18	1978-1990	67	0.27
5	Margaret Court[a]	4	3	1	3	11	1968-1973	22	0.50
6	Monica Seles	4	3	0	2	9	1990-1996	40	0.23
7	Billie Jean King[b]	0	1	4	3	8	1968-1975	35	0.23
8	Evonne Goolagong Cawley[c]	4	1	2	0	7	1971-1980	34	0.21
8	Justine Henin	1	4	0	2	7	2003-2007	34	0.21
8	Venus Williams	0	0	5	2	7	2000-2008	67	0.10

[a] Margaret Court played 25 Grand Slam tournaments before the Open Era began, winning an additional 13 titles, giving her an overall win rate (titles per entry) of 0.51 (=24 titles in 47 entries).

[b] Billie Jean King played 15 Grand Slam tournaments before the Open Era began, winning an additional 4 titles, giving her an overall win rate of 0.24 (= 12 titles in 51 entries).

[c] Evonne Goolagang Cawley played 1 Grand Slam tournament before the Open Era began, with no titles, giving her an overall win rate of 0.20 (= 7 titles in 35 entries).

5 Results

Table 2 shows the all-time greatest women's tennis players as ranked by maximum one year strength. Just as she tops the Grand Slam titles rankings list, Steffi Graf tops our list. Serena Williams, on the other hand, drops from second place to fifth. This fall is likely to be a consequence of two factors: first, there is a greater volatility in her results (e.g. it has taken her longer to get to 21 Grand Slam titles than it took Graf to get to 22), and second, she may be playing against weaker opposition.

In addition to the estimated strengths, Table 2 also shows the confidence interval for the ranking. Any one of four players could actually be ranked number one (Graf, Navratilova, Seles and Serena Williams).

Table 3 shows the rankings based on maximum 3-year strengths and maximum lifetime strengths. The 3-year and lifetime strengths are calculated as the area under the strength curve over a three-year period, or the entire career of the player respectively. Martina Navratilova tops both of these lists pushing Steffi Graf into second place. It is interesting to note the confidence interval on Monica Seles's ranking. Seles was subject to an on-court attack in 1993 in which a man stabbed her in the back. At the time, she was the top ranked player in the world and was a fierce rival of Steffi Graf. Her career never fully recovered after the attack but our results suggest that if it had, she might well have gone on to win many more titles.

Figure 1 shows a plot of the evolution of the top two players' strengths according to our model: Steffi Graf and Martina Navratilova. Graf and Navratilova met nine times in Grand Slam tournaments with Navratilova having a 5-4 record, winning the last Grand Slam match between the two of them despite being 34 at the time. Between 1987 and 1989 they faced each other in three consecutive and enthralling Wimbledon Singles finals. However, despite Navratilova holding the better head-to-head record, our model suggests that at her very best, Graf would have a higher probability of winning a set (and hence a match) against her rival. In addition to the two players' strengths, also shown on the plot are the Grand

Table 2: Greatest women's tennis player according to the maximum strength achieved by a player during her career with the value of that natural logarithm of strength and the year it was achieved. Standard errors on estimated log-strengths are shown in parentheses.

Rank	Player	Strength	Year	95% CI Ranking
1	S. Graf	7.392 (0.18)	1989	(1, 3)
2	M. Navratilova	7.015 (0.05)	1985	(1, 4)
3	M. Seles	6.711 (0.08)	1994	(1, 9)
4	C. Evert	6.141 (0.07)	1977	(2, 8)
5	S. Williams	5.777 (0.12)	2004	(1, 10)
6	J. Henin	5.532 (0.09)	2008	(2, 14)
7	B. J. King	5.048 (0.09)	1974	(3, 16)
8	G. Sabatini	4.103 (0.11)	1992	(2, 16)
9	A. S. Vicario	4.076 (0.03)	1995	(4, 17)
10	M. Court	3.947 (0.06)	1974	(3, 18)
11	M. Hingis	3.882 (0.03)	1998	(4, 19)
12	E. Goolagong	3.802 (0.02)	1976	(4, 21)
13	K. Clijsters	3.71 (0.03)	2005	(7, 21)
14	L. Davenport	3.708 (0.05)	2000	(8, 24)
15	V. Williams	3.624 (0.01)	2002	(9, 22)
16	H. Mandlikova	3.434 (0.07)	1986	(8, 24)
17	J. Capriati	3.192 (0.16)	1991	(8, 26)
18	M. J. Fernandez	3.189 (0.05)	1992	(8, 30)
19	H. Sukova	3.115 (0.04)	1986	(8, 34)
20	M. Sharapova	2.928 (0.04)	2007	(11, 31)

Slam title victories. Both were serial winners but it is interesting to note how Navratilova's strength trajectory follows the typical, almost deterministic, relatively quick rise during the early part of her career, followed by a longer decline towards retirement. Graf's trajectory, on the other hand, has several marked peaks and troughs.

5.1 Goodness of Fit and model criticism

To assess goodness of fit we considered the observed and expected numbers of matches won for a period of 10 years around a player's peak strength. If the probability of winning a set is p, the probability of winning 2 or 3 sets out of 3 (and so winning the match) is $3p^2(1-p)+p^3 = p^2(3-2p)$. For the top 50 players, a chi-squared goodness of fit test on the probabilities of winning a match gave $X^2[50] = 30.2$, suggesting that the observed and predicted numbers of matches won were in good agreement.

Since we are using an empirical Bayes approach to shrink the estimated strengths towards the grand mean, it is interesting to investigate the amount of shrinkage taking place. To do so, we calculate the Higgins and Thompson (2002) I^2 statistic. First, one must calculate $Q = \sum_{l=1}^{p}(y_l - \mu)^2/\sigma_l^2$ which is given by $Q = 2\Delta\ell$, where $\Delta\ell = \ell_0(\phi \to \infty) - \ell_0(\phi = 0)$. From this, the I^2 statistic is given by $I^2 = 100(Q-(p-1))/Q$. For our data, this is 89%. In meta-analysis, a high I^2 means that a random effect is needed, i.e. the different studies appear to be measuring different values for the treatment effect. Here of course we expect player mean strengths to be different, so a high value of I^2 such as 89%, means that the data provides evidence that there are differences in the player strengths and that we are differentiating between them. A low I^2 would have meant that most of the apparent difference in player performances was noise.

Table 3: All time greatest tennis player according to a player's maximum three-year strength (columns 2 to 5) and total lifetime strength (final four columns). Standard errors on estimated log-strengths are shown in parentheses.

Ranking	Player	Strength	Years	95% CI Ranking	Player	Strength	Years	95% CI Ranking
1	M. Navratilova	6.681 (0.36)	1985 to 1987	(1, 4)	M. Navratilova	2.793 (1.74)	1973 to 2004	(1, 2)
2	S. Graf	6.409 (0.89)	1988 to 1990	(1, 5)	C. Evert	3.866 (1.21)	1971 to 1989	(1, 4)
3	M. Seles	6.312 (0.42)	1993 to 1995	(1, 8)	S. Williams	3.048 (1.16)	1998 to 2016	(2, 4)
4	C. Evert	5.717 (0.38)	1976 to 1978	(2, 8)	S. Graf	4.027 (2.04)	1983 to 1999	(2, 6)
5	J. Henin	5.007 (0.56)	2008 to 2010	(1, 11)	M. Maleeva	1.472 (0.35)	1982 to 2005	(4, 5)
6	S. Williams	4.911 (0.88)	2002 to 2004	(3, 12)	V. Williams	2.184 (0.88)	1997 to 2016	(5, 8)
7	B. J. King	4.594 (0.44)	1973 to 1975	(2, 14)	M. Seles	3.484 (1.83)	1989 to 2003	(1, 13)
8	A. S. Vicario	4.01 (0.06)	1994 to 1996	(3, 15)	L. Davenport	2.044 (0.81)	1991 to 2008	(7, 11)
9	E. Goolagong	3.713 (0.1)	1975 to 1977	(2, 17)	B. J. King	2.529 (1.14)	1968 to 1983	(4, 15)
10	M. Hingis	3.672 (0.19)	1997 to 1999	(4, 16)	A. S. Vicario	2.468 (1.18)	1987 to 2002	(6, 17)
11	V. Williams	3.468 (0.15)	2001 to 2003	(6, 18)	E. Goolagong	2.283 (1.04)	1968 to 1983	(7, 18)
12	K. Clijsters	3.47 (0.22)	2004 to 2006	(6, 19)	Conchita Martinez	1.707 (0.73)	1988 to 2005	(9, 16)
13	G. Sabatini	3.498 (0.63)	1991 to 1993	(6, 22)	T. Austin	1.666 (0.77)	1977 to 1994	(8, 18)
14	M. Court	3.412 (0.49)	1973 to 1975	(7, 22)	M. Pierce	1.758 (0.43)	1990 to 2006	(10, 18)
15	L. Davenport	3.325 (0.35)	1999 to 2001	(8, 22)	P. Shriver	1.449 (0.52)	1978 to 1996	(11, 18)
16	H. Mandlikova	2.976 (0.42)	1985 to 1987	(8, 23)	H. Sukova	1.568 (0.75)	1981 to 1998	(10, 21)
17	J. Capriati	2.927 (0.17)	2002 to 2004	(10, 25)	K. Clijsters	2.717 (0.77)	1999 to 2012	(1, 85)
18	H. Sukova	2.87 (0.26)	1985 to 1987	(8, 34)	V. Wade	1.45 (0.41)	1968 to 1985	(13, 22)
19	M. J. Fernandez	2.834 (0.33)	1991 to 1993	(10, 33)	Z. Garrison	1.46 (0.39)	1980 to 1996	(15, 23)
20	T. Austin	2.76 (0.13)	1978 to 1980	(10, 31)	W. Turnbull	1.042 (0.52)	1970 to 1989	(17, 24)

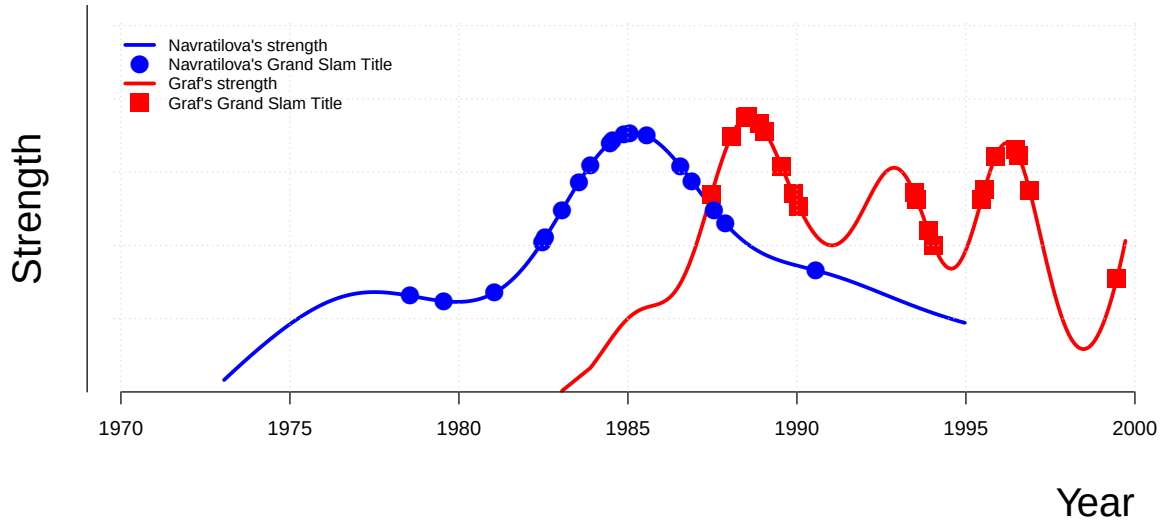

Figure 1: Estimated strength trajectories for two of the greatest players in women's tennis: Steffi Graf and Martina Navratilova.

6 Conclusions

In this paper we have presented an empirical Bayes methodology for fitting time-varying ratings models. The advantages of adopting the empirical Bayes approach are first, that account is taken of players who do not play in many matches or have a high volatility in results, and second, that the strengths of players winning or losing all of their matches can still be estimated.

In addition to the basic modelling procedure we present two additional concepts: rankings confidence intervals and connectivity. Calculating confidence intervals for the ranks themselves makes interpretation of the results much simpler than using standard errors on the estimated strengths alone. Here for example, we find that any one of four players could actually be the best player in the history of women's tennis: Steffi Graf, Martina Navratilova, Monica Seles or Serena Williams. Of course, if one had to give a single name, Steffi Graf in 1988 would be the 'point estimate' having attained the highest strength of any one player.

References

Baker. R. D. and Jackson, D., (2014), Statistical application of barycentric rational interpolants: an alternative to splines, *Computational Statistics*, **29** (5), 1065-1081.

Baker, R.D. and McHale, I.G. (2014). A dynamic paired comparisons model: Who is the greatest tennis player? *European Journal of Operational Research*, 236(2), 677-684.

Baker, R.D. and McHale, I.G. (2015). An empirical Bayes' procedure for ranking players in Ryder Cup golf. *Journal of Applied Statistics*, DOI:10.1080/02664763.2015.1043869.

Berrut, J. P. and Trefethen, L. N. (2004). Barycentric Lagrange interpolation, *SIAM Review*, 46, 501-517.

Casella, G. and Berger, R. L. (2002), Statistical Inference (2nd ed.), Duxbury, California.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors, *Communications of the ACM* **7** (3): 171-176.

Glickman, M. E. (1993). Paired comparison models with time-varying parameters, Ph.D. Dissertation, Department of Statistics, Harvard University, Cambridge, Massachusetts.

Glickman, M.E. (2001). Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, 28(6), 673-689.

Harvey, A. and Koopman, S. J. (1993), Forecasting Hourly Electricity Demand Using Time-Varying Splines, *Journal of the American Statistical Association* **88**, 1228-1236.

Held, L. and Vollnhals, R. (2005). Dynamic rating of European football teams. *IMA Journal of Management Mathematics* **16**, 121-130.

Higgins J. P. T. and Thompson S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**, 1539-1558.

Knorr-Held, L. (2000). Dynamic rating of sports teams. *Journal of the Royal Statistical Society, Series D* **49**, 261-276.

Hunter, D.R. (2004). MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics.* 32, 384-406.

Maritz, J. S. and Lwin, T. (1989). Empirical Bayes methods (2nd. ed.), Chapman and Hall, New York.

Stern, H. (1990). Models for Distributions on Permutations, *Journal of the American Statistical Association*, 85, 410, 558-564.