A point-based Bayesian hierarchical model to predict the outcome of tennis matches

Martin Ingram, Silverpond

September 21, 2017
Introduction

Predicting tennis matches is of interest for a number of applications:

- Coaching: Prediction models can provide useful feedback about who players should be able to beat and how they are improving over time
- Fan engagement: Who is the favourite? By how much? Who is currently the best player?
Approaches to tennis prediction

Broadly speaking, published tennis prediction models fall into three classes:

1. Regression models
2. Paired comparison models
3. Point based models
Regression models

- Regression models phrase match prediction as a regression task, using a suitable link function (logit/probit) to predict match outcomes.
- For example, Gilsdorf et al. [1] predict using a probit model including ranking, prize earnings and demographics.
Paired comparison models

- Paired comparison models model match outcomes by assuming that each player has a hidden latent ability.
- The probability of a player winning a match is modelled as a function of the difference of the two latent abilities, $\theta_1$ and $\theta_2$.
- For example, Elo typically uses the following likelihood:

$$p(p_1 \text{ wins} | \theta_1, \theta_2) = \frac{1}{1 + 10^{(\theta_2 - \theta_1)/400}}$$

(1)

- A version of Elo (with an optimised k-factor) devised by FiveThirtyEight [2] has been particularly popular in tennis.
- Other interesting paired comparison models exist but are not as popular for tennis (e.g. TrueSkill [3], Glicko [4]).
Point based models (I)

- Point based models use a model of a tennis match developed, among others, by Newton & Keller [5].
- It assumes that points on serve are independent and identically distributed (i.i.d.).
- This means that the probabilities $p_1$ and $p_2$ of winning a point on serve for players 1 and 2 are assumed constant throughout the match.
- Using recursive equations, it is possible to calculate the probability of holding serve, winning a set, winning a tiebreak, and winning the match as functions of only $p_1$ and $p_2$. 
For $p_1 = 0.63$ (ATP average), probability of holding serve is 79.4%

For $p_1 = 0.65$ and $p_2 = 0.60$, probability of player 1 winning a best-of-three match is 73.7%
Point based models (II)

- Point based prediction models predict $p_1$ and $p_2$ and then predict the match winner using the i.i.d. model.
- For example, Barnett and Clarke [6] propose to calculate the probability as:

$$f_{ij} = f_t + (f_i - f_{av}) - (g_j - g_{av})$$

- $f_t$: average serve-winning probability at the tournament
- $f_i$: the player’s average serve-winning probability
- $f_{av}$: the tour average serve-winning probability
- $g_j$: the opponent’s average return-winning probability
- $g_{av}$: the tour average return-winning probability
In a 2015 paper [7], Stephanie Kovalchik compares 11 published prediction models, including all three model classes, by predicting the ATP’s 2014 season.

The best representatives of each model type are:

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>Accuracy</th>
<th>Log loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression-based</td>
<td>Gilsdorf et al.</td>
<td>68%</td>
<td>0.61</td>
</tr>
<tr>
<td>Paired comparison</td>
<td>FiveThirtyEight Elo</td>
<td>70%</td>
<td>0.59</td>
</tr>
<tr>
<td>Point-based</td>
<td>Barnett &amp; Clarke</td>
<td>67%</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Point-based models have the highest log loss and lowest accuracy.
Pros and cons of the point-based approach

- In addition to the worse evaluation, the i.i.d. model is also proven to be wrong, albeit a “good approximation” (Klaassen & Magnus [8]). Players do not play i.i.d., although deviations are quite small.
- Why use it at all? The main attraction is the ability to make a great wealth of predictions beside match outcome:
  - Number of sets (e.g.: two or three sets?)
  - Set scores (e.g.: how likely is a tiebreak?)
  - Many more: number of games, number of points...
- In addition, in-play win probabilities can be calculated based on the score
Example: Set scores

At low $p_1$ and $p_2$ (average / below average servers), scores like 6-3 or 6-2 are likely.

At high $p_1$ and $p_2$ (very strong servers), a tiebreak becomes most likely.
Key question of this talk:

Can we build a better point-based model?
I wanted the model to account for the following factors:

- **Surface preferences**: Tennis is played on a number of surfaces (clay, grass, hard, indoor). Players often do better on one surface than another (e.g. Nadal: 10 French Open titles, 2 Wimbledon titles).

- **Tournament effects**: It is easier to win points on serve at some tournaments than at others, raising players’ averages (and making e.g. tiebreaks more likely).

- **Time dependence**: Player skills change over time
Likelihood

Split each tennis match into two “serve-matches” and use a binomial likelihood:

\[ y_i \sim Binomial(n_i, \theta_i) \]  

where:

- \( y_i \): Points won on serve in serve-match \( i \)
- \( n_i \): Points played on serve in serve-match \( i \)
- \( \theta_i \): Serve-winning probability in serve-match \( i \)
Modelling $\theta_i$

The model for $\theta_i$, the serve-winning probability in serve-match $i$, is:

$$
\text{logit}(\theta_i) = (\alpha_{s(i)}p(i) - \beta_{r(i)}p(i)) + (\gamma_{s(i)}m(i) - \gamma_{r(i)}m(i)) + \delta_{t(i)} + \theta_0 \quad (4)
$$

- $\alpha_{s(i)}p(i)$: server $s(i)$'s serving skill in period $p(i)$
- $\beta_{r(i)}p(i)$: returner $r(i)$'s returning skill in period $p(i)$
- $\gamma_{s(i)}m(i)$: server's additional skill on surface $m(i)$
- $\gamma_{r(i)}m(i)$: returner's additional skill on surface $m(i)$
- $\delta_{t(i)}$: adjustment to the intercept at tournament $t(i)$
- $\theta_0$: intercept
Modelling time dependence

Took some inspiration from Glicko [4]. Serve skills $\alpha$ and return skills $\beta$ follow a Gaussian random walk over time:

$$\alpha_{p+1} \sim N(\alpha_p, \sigma_\alpha^2) \quad (5)$$
$$\beta_{p+1} \sim N(\beta_p, \sigma_\beta^2) \quad (6)$$

$$\sigma_\alpha, \sigma_\beta \sim N(0, 1) \quad (7)$$

In other words, skills in the next period are a small normal jump away from the previous skills. Note priors on $\sigma$ are constrained to be positive when fit (unit half-normals).
Hierarchical priors

Initial skills, tournament intercepts and surface skills all have hierarchical priors:

\[
\delta \sim N(0, \sigma_\delta^2) \tag{8}
\]
\[
\gamma \sim N(0, \sigma_\gamma^2) \tag{9}
\]
\[
\alpha_1 \sim N(0, \sigma_{\alpha_0}^2) \tag{10}
\]
\[
\beta_1 \sim N(0, \sigma_{\beta_0}^2) \tag{11}
\]

All priors for the group \( \sigma \)'s are unit half-normals.
Model Checks & Validation
A point-based Bayesian hierarchical model to predict the outcome of tennis matches

External validation

- Use data from 2011 onwards to predict 2014
- With periods of 3 months, fit model four times, once for each quarter of 2014
- Use posterior estimates and i.i.d. model to predict win probabilities

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>Accuracy</th>
<th>Log loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression-based</td>
<td>Gilsdorf et al.</td>
<td>68%</td>
<td>0.61</td>
</tr>
<tr>
<td>Paired comparison</td>
<td>FiveThirtyEight Elo</td>
<td>70%</td>
<td>0.59</td>
</tr>
<tr>
<td>Point-based</td>
<td>Barnett &amp; Clarke</td>
<td>67%</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>Point-based</strong></td>
<td><strong>Proposed</strong></td>
<td><strong>68%</strong></td>
<td><strong>0.60</strong></td>
</tr>
</tbody>
</table>

→ Considerable improvement compared to previous best point-based model!
A point-based Bayesian hierarchical model to predict the outcome of tennis matches

Point-level validation

For Barnett & Clarke and the model, can also compute metrics on how well the serve-winning probabilities are estimated.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barnett &amp; Clarke</td>
<td>0.081</td>
<td>22.3%</td>
</tr>
<tr>
<td><strong>Proposed</strong></td>
<td>0.077</td>
<td>28.7%</td>
</tr>
</tbody>
</table>

→ Improved here too (as you would expect).
A point-based Bayesian hierarchical model to predict the outcome of tennis matches

Evaluation: Posterior predictive checks

Fit model from 2014 up to 2017 (pre US Open) with three-month periods. Replicate $y_i$ using 4,000 simulations of $\theta_i$, and compare test quantities computed on data to those on replications.

Mean: replications match data exactly ($p=0.50$).

Standard deviation: all replications have greater standard deviation than the data!
Evidence of underdispersion?

Compare one replication in more detail.

→ Good agreement in general, but at low expected $y$, the model underpredicts the points won on serve; at high expected $y$, it overpredicts. Evidence of underdispersion?
Results & Examples
A point-based Bayesian hierarchical model to predict the outcome of tennis matches

Results: Serve and return skills Q3 2017

One advantage of point-based model: can look at serve and return skills. Broadly agree with intuition; some interesting: Nadal very strong on serve, Schwartzman very strong on return.
Results: Overall skills

- Results mostly intuitive: e.g. Nadal and Federer shared all 4 Grand Slams this year
- Surprises: Kyrgios ranked highly (only number 18 in ATP Rankings); Wawrinka ranked low (number 4 in ATP rankings)
A point-based Bayesian hierarchical model to predict the outcome of tennis matches

Serve compared to return skill
Skill evolution – Renaissance of Federer (?) and Nadal

- Nadal and Federer share 4 Grand Slams after droughts of 5 years (Federer) and 3 years (Nadal). Big improvements suggested.
- But: Federer was better in 2015!
- Djokovic has declined a great deal.
- Nadal has improved greatly (particularly on serve). Moya to credit (coach since Dec ’16)?
I fit the model up to the start of the US Open. How would it have predicted the final: Nadal vs. Anderson? Nadal won 6-3 6-3 6-4.
Surface effects: Nadal vs. Anderson on grass

How would things change in a hypothetical match at Wimbledon?
Surface effects: Nadal vs. Anderson on clay

How would things change in a hypothetical match at the French Open?
Conclusions and future work

- Introduced a new point-based model with higher prediction accuracy than the previous best
- Takes into account surface effects and time-varying ability, as well as tournament effects
- Future work:
  - Accelerating model fit: Currently fit using Stan, takes about 80 minutes. Limited to period lengths of 1 month or over, and data of about 5 years or less. An approximate solution e.g. using variational Bayes or another approach would be of interest.
  - Different skill time evolution: e.g. like Glicko 2, where the jumps are drawn from another distribution [9], or maybe a Gaussian process.
  - Investigate alternatives to the Binomial link, such as the COM-Poisson model, which could handle underdispersion, and / or investigate causes of underdispersion further
Thank you for paying attention!
A point-based Bayesian hierarchical model to predict the outcome of tennis matches

Appendix: Comparison to ATP Rankings

Broad agreement, but exceptions highlight differences: model rates Kyrgios much higher (injuries), Thiem lower (plays a lot), Murray lower (declining this year). Wawrinka is a very variable player.
Kyrgios, Zverev lead among young players
Zverev climbing at fastest rate
Khachanov improving (slowly), Coric stagnating
Appendix: Clay skills

- Biggest clay boost: Rafael Nadal (unsurprisingly)
- Players from Latin America and Spain do very well
A point-based Bayesian hierarchical model to predict the outcome of tennis matches

Appendix: The tour is more competitive now than 2015

![Mean posterior serve vs. mean posterior return skills, Q3 2015](chart.png)
Appendix: Serve compared to return skill in 2017
References I


References III
