Bayesian Skill Ranking

Leland Chen, Joseph Huang, Ryan Thompson

March 16, 2011

1 Introduction

Template

2 Related Work

Let's talk about [1] and [2] and then just follow the citations that those guys make to introduce the general challenges and state-of-the art.

Talk about the history of ELO [3] and Glicko [4], and the ongoing debate between logistic and gaussian (logit vs. probit).

Basketball brings two key challenges to the field of Bayesian Skill estimation.

- Results are not binary Win/Lose outcomes.
- Results are influenced by the skill differential of more than two players.

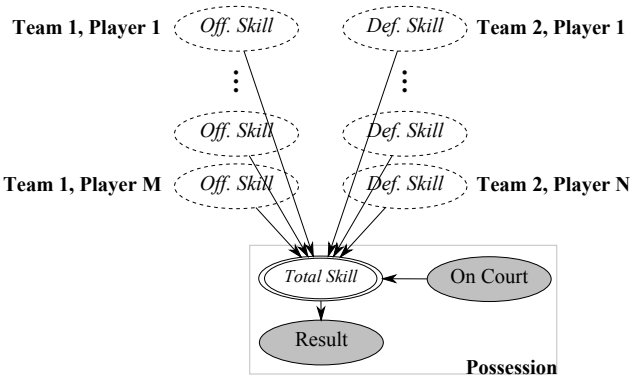
3 Method

3.1 Traditional Networks

We model the result of each possession as an independent and identically distributed random variable, *Result*. An NBA roster has 12 players, but only five of them are on the court for a given team during any single posession.

To reduce noise, we ignore "garbage time" posessions. To simplify the model we will only consider possessions in the first half of games.

We begin with a very basic model that follows traditional "skill ranking" in the sense that each player has an unobserved (or parameterized) skill that has some distribution, and using inference (or parameter estimation) we can determine the value of each player's skill. In this network, we decided each player would have a hidden offsensive and defensive skill value, shared across possessions:



These skills would contribute deterministically in some way to an "effective" total skill differential between the two teams, and then the *Result* variable would be one of four outcomes:

- $R = 0$ Offensive Team Scores nothing, change of possession (e.g. turnover, defensive rebound, etc.)
- $R = 1$ Offensive Team Scores 1 point
- $R = 2$ Offensive Team Scores 2 points
- $R = 3$ Offensive Team Scores 3 points

Note: We are ignoring the rare 4 point plays.

The *On Court* random variable "multiplexes" between those players that are on the court and those that are not.

3.2 Issues

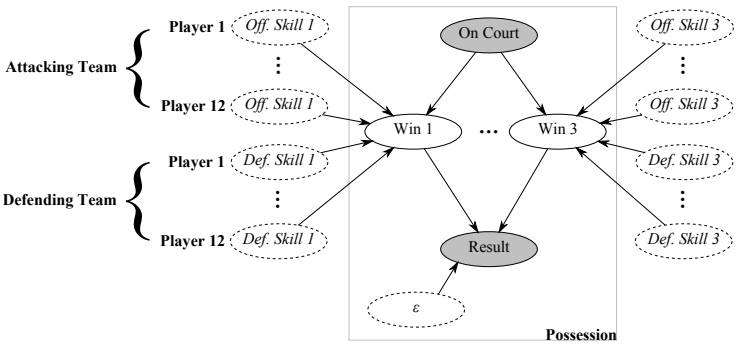
These traditional models [1] have difficulty capturing the proper causalities when Results can have multinomial-valued outcomes. For example, regardless of the parameterization of the *Results* CPD, both $\Pr\{R = 2\}$ and $\Pr\{R = 3\}$ would depend on the same skills of the same players. The relative distribution of outcomes $\Pr\{R = 2\}$ vs. $\Pr\{R = 3\}$ would be shared across all "units" (i.e. all combinations of *On Court* assignments).

In reality, a team that scores $R = 3$ half the time and $R = 1$ half the time is just as good as a team that scores $R = 2$ all the time. However, any traditional Win/Lose model will unfairly penalize the likelihood of one of these teams over the other and leads to under-fitting. This might suggest that we have a random variable that represents $\mathrm{E}[\text{points scored}]$, but this doesn't pass the clarity test.

Secondly, there is a lot of value in being able to compare with the state-of-the-art in the Win/Lose based Bayesian Skill Ranking literature. Specifically, there is common debate over logistic vs. Gaussian skill/performance distributions and we wish to be sensitive to that conversation in this project. Having a multinomial outcome makes it difficult to directly compare logistic vs. Gaussian in the standard skill/performance framework because there is no consensus in the literature about how to extend these models just to support ties [5], let alone general multinomial outcomes.

3.3 Proposed Network

To address some of the shortcomings discussed in **Section 3.2**, we propose the following:



In this network, each player is represented by three skill parameters:

1. Skill 1: Ability to score (or defend against) one-point opportunities
2. Skill 2: Ability to score (or defend against) two-point opportunities
3. Skill 3: Ability to score (or defend against) three-point opportunities

For notational convenience, let

$$A\left(\text{Skills}\right) \triangleq \exp\left(\sum_i \text{Off. Team, Player i Off. Skill } k\right)$$

$$B\left(\text{Skills}\right) \triangleq \exp\left(\sum_i \text{Def. Team, Player i Def. Skill } k\right)$$

where

$$\begin{aligned}\vec{\theta}_k = &\{\text{Off. Team, Player i Off. Skill } k\}_k \\ &\cup \{\text{Def. Team, Player i Def. Skill } k\}_k\end{aligned}$$

On each possession, there are three binary hidden random variables defined by a probability function with parameters that we will learn. In particular, each of the "Win" random variables depends on the corresponding skills of players on court as input (which are learned):

1. *Win 1*: True if there was a **guaranteed opportunity** to score one point at some point during the possession.
2. *Win 2*: True if there was a **guaranteed opportunity** to score two points at some point during the possession.
3. *Win 3*: True if there was a **guaranteed opportunity** to score three points at some point during the possession.

Note: The term "opportunity" is not intended in the typical basketball sense, where it often means "the opportunity to take a shot".

Each of the *Win k* events has a CPD parameterized by $\vec{\theta}_k$ (whose values we will learn during training), which depends on the skills of the players. In the basic model of player skills, the skill of a five-man team is the sum of the skills of its players.

For example, the Bradley-Terry model corresponds to:

$$\Pr\left\{\text{Win}_k = \textbf{True} \mid \text{Skills}\right\} = \frac{A\left(\text{Skills}\right)}{A\left(\text{Skills}\right) + B\left(\text{Skills}\right)}$$

Alternatively, the Thurstone Case V model corresponds to:

$$\Pr\left\{\text{Win}_k = \textbf{True} \mid \text{Skills}\right\} = \Phi\left(A - B\right)$$

The logit (Bradley-Terry) and probit(Thurstone Case V) are the main Binary Response Models used throughout the history of the skill ranking literature [3, 4, 1, 2].

However, the power of this Bayesian Network construction is that any Binary Response Model can be plugged in modularly. One can easily compare the performance of Cauchy [6], Log-Log and Complementary Log-Log [7], Scobit [8], etc. without changing the network or the core inference algorithm.

For notational convenience, we will write $W_k = w_k^1$ for $\text{Win}_k = \textbf{True}$, and $W_k = w_k^0$ when **False**.

To understand the relationship between $\{Win\}$ and *Result*, let's go through an example. Imagine that for a particular five-man unit, the θ are known and the resulting probabilities are:

$$\begin{aligned}\Pr\left\{W_1 = w_1^1\right\} = 5\% \quad &, \quad \Pr\left\{W_1 = w_1^0\right\} = 95\% \\ \Pr\left\{W_2 = w_2^1\right\} = 35\% \quad &, \quad \Pr\left\{W_2 = w_2^0\right\} = 65\% \\ \Pr\left\{W_3 = w_3^1\right\} = 10\% \quad &, \quad \Pr\left\{W_3 = w_3^0\right\} = 90\%\end{aligned}$$

This means, on each possession

- the offensive team has a 10% probability of scoring 3 points (i.e. $R = r^3$ with 90% probability)
- the offensive team has a $(100\% - 10\%) \times 35\%$ probability of scoring 2 points (i.e. $\Pr\left\{R = r^2\right\} = 31.50\%$)
- the offensive team has a $90\% \times 65\% \times 5\%$ probability of scoring 1 point (i.e. $\Pr\left\{R = r^1\right\} = 2.925\%$)
- the offensive team has a 55.575% chance of scoring 0 points

Essentially, *Result* simply selects the highest-scoring decision available during the possession. If *Win 3* is true, *Result* is 3; If *Win 3* is false but *Win 2* is true, *Result* is 2; etc.

Result can be specified with either a tree-CPD or table-CPD, but it has the following probabilities.

$R \mid W_1, W_2, W_3$	$R = r^0$	$R = r^1$	$R = r^2$	$R = r^3$
$w_3^1 w_2^1 w_1^1$	ε	ε	ε	$(1 - 3\varepsilon)$
$w_3^1 w_2^1 w_1^0$	ε	ε	ε	$(1 - 3\varepsilon)$
$w_3^1 w_2^0 w_1^1$	ε	ε	ε	$(1 - 3\varepsilon)$
$w_3^1 w_2^0 w_1^0$	ε	ε	ε	$(1 - 3\varepsilon)$
$w_3^0 w_2^1 w_1^1$	ε	ε	$(1 - 3\varepsilon)$	ε
$w_3^0 w_2^1 w_1^0$	ε	ε	$(1 - 3\varepsilon)$	ε
$w_3^0 w_2^0 w_1^1$	ε	$(1 - 3\varepsilon)$	ε	ε
$w_3^0 w_2^0 w_1^0$	$(1 - 3\varepsilon)$	ε	ε	ε

The single parameter ε is inspired by the noisy-max and essentially indicates unmodelled errors (e.g. offensive or defensive mistakes). The better our model fits the actual flow of the game, the smaller ε should be.

4 Implemention

Maximum Likelihood starts with:

$$\begin{aligned}L &= \prod_{\mathcal{D}} \Pr\left\{\mathcal{D} \mid \vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \varepsilon\right\} \\ &= \prod_{\mathcal{D}} \sum_{\substack{w_1 \in \text{Val}(W_1) \\ w_2 \in \text{Val}(W_2) \\ w_3 \in \text{Val}(W_3)}} \\ &\qquad \Pr\left\{R = \mathcal{D}_r, W_1 = w_1, W_2 = w_2, W_3 = w_3 \mid \vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \varepsilon\right\}\end{aligned}$$

$$= \prod_{\mathcal{D}} \sum_{\substack{w_1 \in \mathrm{Val}(W_1) \\ w_2 \in \mathrm{Val}(W_2) \\ w_3 \in \mathrm{Val}(W_3)}} \Pr\left\{R = \mathcal{D}_r \mid W_1 = w_1, W_2 = w_2, W_3 = w_3, \cancel{\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3}, \varepsilon\right\} \Pr\left\{W_1 = w_1 \mid \vec{\theta}_1, \cancel{\vec{\theta}_2, \vec{\theta}_3, \varepsilon}\right\} \Pr\left\{W_2 = w_2 \mid \vec{\theta}_2, \cancel{\vec{\theta}_1, \vec{\theta}_3, \varepsilon}\right\} \Pr\left\{W_3 = w_3 \mid \vec{\theta}_3, \cancel{\vec{\theta}_1, \vec{\theta}_2, \varepsilon}\right\}$$

In the M-step, we assume $\mathcal{D}$ has observed $W_1 = \mathcal{D}_1$, $W_2 = \mathcal{D}_2$, and $W_3 = \mathcal{D}_3$ (after soft assignments during the E-step) which provides global decomposition allowing us to log-maximize each section of the network separately:

$$L = \left(\prod_{\mathcal{D}} \Pr\left\{R = \mathcal{D}_r \mid W_1 = \mathcal{D}_1, W_2 = \mathcal{D}_2, W_3 = \mathcal{D}_3, \varepsilon\right\}\right) \left(\prod_{\mathcal{D}} \Pr\left\{W_1 = \mathcal{D}_1 \mid \vec{\theta}_1\right\}\right) \left(\prod_{\mathcal{D}} \Pr\left\{W_2 = \mathcal{D}_2 \mid \vec{\theta}_2\right\}\right) \left(\prod_{\mathcal{D}} \Pr\left\{W_3 = \mathcal{D}_3 \mid \vec{\theta}_3\right\}\right)$$

4.1 Maximum Likelihood: ε

$$\operatorname*{argmax}_{\varepsilon} \left\{\prod_{\mathcal{D}} \Pr\left\{R = \mathcal{D}_r \mid W_1 = \mathcal{D}_1, W_2 = \mathcal{D}_2, W_3 = \mathcal{D}_3, \varepsilon\right\}\right\} = \operatorname*{argmax}_{\varepsilon} \left\{(1 - 3\varepsilon)^{M_{modelled}} (\varepsilon)^{M_{\mathrm{noise}}}\right\}$$

with solution

$$\varepsilon = \frac{1}{3} \frac{M_{\mathrm{noise}}}{M_{\mathrm{noise}} + M_{\mathrm{modelled}}}$$

where

$$M_{\mathrm{modelled}} \triangleq \mathrm{M}\left[r^3, w_3^1\right] + \mathrm{M}\left[r^2, w_3^0, w_2^1\right] + \mathrm{M}\left[r^1, w_3^0, w_2^0, w_1^1\right] + \mathrm{M}\left[r^0, w_3^0, w_2^0, w_1^0\right]$$

and M_{noise} is a count of the remaining observations such that $M_{\mathrm{modelled}} + M_{\mathrm{noise}} = M$, the total number of observations.

4.2 Maximum Likelihood: θ_i

Let

$$\Pr\left\{W_i = w_i^0 \mid \vec{\theta}_i\right\} \triangleq \frac{1}{1 + \exp\left(\Delta_i\right)}$$

so that

$$\Pr\left\{W_i = w_i^1 \mid \vec{\theta}_i\right\} \triangleq \frac{1}{1 + \exp\left(-\Delta_i\right)}$$

where

$$\Delta_i \triangleq (\theta_{i,\mathrm{Off.P1}} + \ldots + \theta_{i,\mathrm{Off.P12}}) + (\theta_{i,\mathrm{Def.P1}} + \ldots + \theta_{i,\mathrm{Def.P12}})$$

Without loss of generality, we can assume that the defensive θs are negative numbers, that reduce the overall $\Pr\{Win\}$ when added. (This is simpler than subtracting positive θ_{Def} and then having to keep track of positive and negative signs the whole time.)

Now, we wish to compute:

$$\operatorname*{argmax}_{\vec{\theta}_i} \left\{\prod_{\mathcal{D}} \Pr\left\{W_i = \mathcal{D}_i \mid \vec{\theta}_i\right\}\right\}$$

when W_i is fully observed with the solution expressed with respect to sufficient statistics of $\vec{\theta}_i$.

This reduces to weighted logistic regression. We use a weighted variant of the basic Newton-Raphson logistic regression techniques [9].

probit: [10]

4.3 Expectation-Maximization: E-step

During the E-step all parameters are fixed, so we simply evaluate probabilities directly. For each datapoint $\langle \mathcal{D}_r, \mathcal{D}_c \rangle$ we create soft-datapoints:

- $\langle \mathcal{D}_r, \mathcal{D}_c, w_1^1, w_2^1, w_3^1 \rangle$ with weight proportional to $\Pr\left\{\mathcal{D}_r, \mathcal{D}_c, w_1^1, w_2^1, w_3^1 \mid \vec{\theta}, \varepsilon\right\}$
- $\langle \mathcal{D}_r, \mathcal{D}_c, w_1^1, w_2^1, w_3^0 \rangle$ with weight proportional to $\Pr\left\{\mathcal{D}_r, \mathcal{D}_c, w_1^1, w_2^1, w_3^0 \mid \vec{\theta}, \varepsilon\right\}$
- $\vdots$
- $\langle \mathcal{D}_r, \mathcal{D}_c, w_1^0, w_2^0, w_3^0 \rangle$ with weight proportional to $\Pr\left\{\mathcal{D}_r, \mathcal{D}_c, w_1^0, w_2^0, w_3^0 \mid \vec{\theta}, \varepsilon\right\}$

4.4 Expectation-Maximization: Initialization

Since soft assignments to W_3 represent roughly the probability of scoring three points during a possession, we can initialize $\Pr\left\{W_3 = w_3^1\right\}$ to the fraction

$$w_3^{\mathrm{init}} := \frac{\mathrm{M}\left[r^3\right]}{M}$$

Similarly, soft assignments to W_3 and W_2 suggest that

$$\frac{\mathrm{M}\left[r^2\right]}{M} \to \Pr\left\{r^2\right\} \approx \left(1 - \Pr\left\{W_3 = w_3^1\right\}\right) \times \Pr\left\{W_2 = w_2^1\right\}$$

is the probability of scoring two points during a possession. So let's initialize

$$w_2^{\mathrm{init}} := \left(\mathrm{M}\left[r^2\right] \div M\right) \div \left(1 - w_3^{\mathrm{init}}\right)$$

And for the same reason,

$$w_1^{\mathrm{init}} := \left(\mathrm{M}\left[r^1\right] \div M\right) \div \left(1 - w_2^{\mathrm{init}}\right)$$

With these soft assignments we can begin EM on the M-step.

4.5 Implementation Considerations

E-Step MLE of θ are not closed-form but require iteration. We choose closed-form Hessian-based weighted regression. Overshooting and step size, stopping criterion, underflow and pruning points, are all real considerations.

5 Results

We started by running a single dataset.

Then more

Then full division

6 Analysis

Metrics: epsilon test/training vs. dataset size $E[Prdatapoint]$ M_count

"Battier is known as..., but was -18 total in the 5 games he played against the southwest in 2008-2009, which is reflected in our results as..."

7 Conclusions

Bayesian prior/smoothing Coach skill More binary response models

8 References

References

[1] R. Herbrich, T. Minka, and T. Graepel, "TrueSkillTM: A Bayesian skill rating system," *Advances in Neural Information Processing Systems*, vol. 20, pp. 569–576, 2007.

[2] R. Coulom, "Whole-history rating: A bayesian rating system for players of time-varying strength," *Computers and Games*, pp. 113–124, 2008.

[3] A. Elo, *The rating of chessplayers, past and present*. Batsford, 1978.

[4] M. Glickman, "Parameter estimation in large dynamic paired comparison experiments," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 48, no. 3, pp. 377–394, 1999.

[5] D. Hunter, "MM algorithms for generalized Bradley-Terry models," *The Annals of Statistics*, vol. 32, no. 1, pp. 384–406, 2004.

[6] "Maximum likelihood estimation," Charles H. Franklin, June 2005, http://users.polisci.wisc.edu/franklin/Content/MLE/Lecs/MLELec07p4up.pdf.

[7] J. Long, *Regression models for categorical and limited dependent variables*. Sage Publications, Inc, 1997.

[8] J. Nagler, "Scobit: an alternative estimator to logit and probit," *American Journal of Political Science*, vol. 38, no. 1, pp. 230–255, 1994.

[9] "Logistic regression," Jia Li, September 2008, http://www.stat.psu.edu/ jiali/course/stat597e/notes2/logit.pdf.

[10] E. Demidenko, "Computational aspects of probit model," *Mathematical Communications*, vol. 6, no. 2, 2001.