Bayesian Skill Ranking

Leland Chen, Joseph Huang, Ryan Thompson

March 16, 2011

1 Introduction

In this paper, we examine professional basketball games in the context of Bayesian Skill Ranking. By modeling every possession as a data point parameterized by a number of points scored and a binary-valued vector of players on the court, we hope to infer player skill levels in six abilities - one-point, two-point, and three-point offense and defense. There are many existing rating systems, but most are designed for single-player games with simple binary outcomes. In order to successfully model basketball outcomes, we built a network with a likelihood function that can be maximized via Expectation-Maximization. We model our main variables with both a logit and probit function, which we use to estimate each player's skill rankings.

2 Related Work

For head-to-head competitions (Player A vs. Player B) with binary outcomes (Player A Wins, Player A Loses), there exist many well-understood techniques for compressing player performances into a single skill parameter. ELO [1] is the most widespread such ranking system. ELO was initially designed around a probit model that determined probability of Player A defeating Player B as a function of the difference of their skill ranking. However, the model was soon changed to the better-performing logit Bradley-Terry model, as suggested originally by [2].

Other ranking systems, such as Glicko [3], model the probability of winning as both a logit and probit function, which are very similar in distribution but have different convergence properties and perform differently at the tails of the distributions.

More recent developments in ranking systems deal with team behavior and performance, and Bayesian networks where skill influences performance, and performance variables are combined to determine winning probabilities. TrueSkill [4] is used for Halo rankings in both individual and cooperative modes, but it only uses Gaussian priors for skill and performance. Whole-History Rating [5] focuses on varying skill parameters over time, to account for players improving with practice and declining with age.

The Elo rating system is used for calculating relative skill levels of players in two-player games, but has been adapted to team sports like soccer, football, basketball, and baseball. The performance of players/teams is inferred from wins, losses, and draws against other players/teams, while depending on the ratings of the opponent and the results scored against them. (Elo also has some mathetmatical issues of that have been addressed through different means). Elo only considers the final score, which is not always indicative of how competitive the game was, and when used for teams, does not consider the players individually. Therefore, our intent is to build a more sophisticated model that determines 6 different skills (3 on offensive and 3 on defense) for each player using Bayesian Networks. By doing so, we can make better inferences on how well these sets of players contribute as a team on each possession and how likely they are to score 0, 1, 2, or 3 points per possession against their opponent.

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

- Results are not binary Win/Lose outcomes - or at least it is nearly impossible to distinguish between players abilities when only examining the final result of a game.

- Results are influenced by the skill differential of more than two players.

- A player's skill level is at least two parameters - offense and defense.

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 possession.

To reduce noise, we ignore "garbage time" possessions. To simplify the model we will only consider possessions in the first half of games, or in the case of games that go to overtime, the entire game. This ensures that we take a lot of the gamesmanship and coaching strategy out of the data - we won't have to deal with starters resting or bench players playing a lot of minutes in a blowout.

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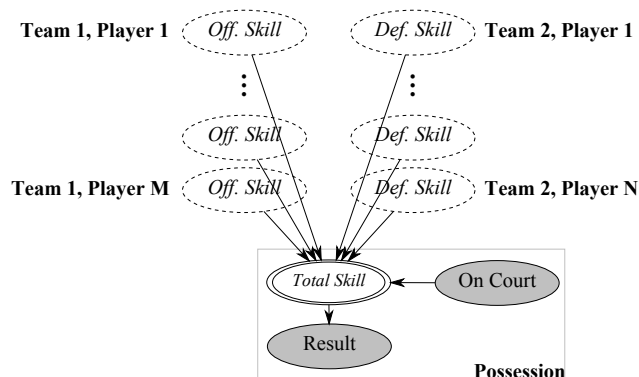
These skills would contribute deterministically in some way to an "effective" total skill differential between the two teams, and then the *Result* variable would be one of four outcomes:

- $R = 0$ 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 [4] 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 [6], let alone general multinomial outcomes. This is part of our rationale for designing a network with a modular framework for different density functions.

3.3 Proposed Network

To address some of the shortcomings discussed in **Section 3.2**, we propose the following:
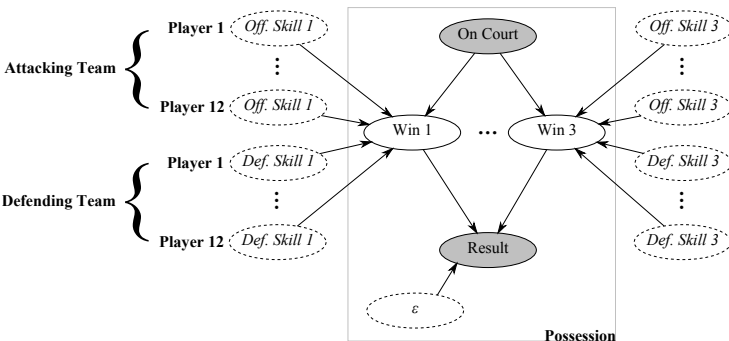
In this network, each player is represented by three skill parameters:

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

$$A\,(\text{Skills}) \triangleq \sum_i (\text{Off. Team, Player i Off. Skill } k) * C_i$$

$$B\,(\text{Skills}) \triangleq \sum_i (\text{Def. Team, Player i Def. Skill } k) * C_i$$

where C_i is the ith element of the vector $\vec{C}$. $\vec{C}$ is a vector of indicator functions. For any particular possession, exactly ten elements of $\vec{C}$ are 1 (there are ten basketball players on the court at once). All the other elements are zero.

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:

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

Alternatively, the Thurstone Case V model corresponds to:

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

The logit (Bradley-Terry) and probit (Thurstone Case V) are the main Binary Response Models used throughout the history of the skill ranking literature [1, 3, 4, 5].

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 [7],

Log-Log and Complementary Log-Log [8], Scobit [9], etc. without changing the network or the core inference algorithm.

For notational convenience, we will write $W_k = w_k^1$ for Win$_k$ = **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:

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

This means, on each possession

- the offensive team has a 10% probability of scoring 3 points (i.e. $R = r^3$ with 10% 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.

$R\|W_1,W_2,W_3$	$R = r^0$	$R = r^1$	$R = r^2$	$R = r^3$
$w_3^1 w_2^1 w_1^1$	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

We will perform Parameter Estimation with Missing Data. Each player's skill is treated as a fixed parameter of the Win CPDs. ε is also a parameter. Each possession from our dataset is treated as an i.i.d. observation from the joint distribution of the entire network. **Result** and **OnCourt** are always observed, $Win~1$, $Win~2$, and $Win~3$ are always missing/hidden.

The general algorithm, then, consists of iteratively maximizing:

$$L = \prod_{\mathcal{D}} \Pr\left\{\mathcal{D} \mid \vec{\theta_1}, \vec{\theta_2}, \vec{\theta_3}, \varepsilon\right\}$$

$$= \prod_{\mathcal{D}} \sum_{\substack{w_1 \in \mathrm{Val}(W_1) \\ w_2 \in \mathrm{Val}(W_2) \\ w_3 \in \mathrm{Val}(W_3)}}$$

$$\Pr\left\{R = \mathcal{D}_r, W_1 = w_1, W_2 = w_2, W_3 = w_3, \mathcal{D}_C \mid \vec{\theta_1}, \vec{\theta_2}, \vec{\theta_3}, \varepsilon\right\}$$

$$= \prod_{\mathcal{D}} \sum_{\substack{w_1 \in \mathrm{Val}(W_1) \\ w_2 \in \mathrm{Val}(W_2) \\ w_3 \in \mathrm{Val}(W_3)}}$$

$$\Pr\left\{R = \mathcal{D}_r \mid W_1 = w_1, W_2 = w_2, W_3 = w_3, \cancel{\mathcal{D}_C}, \cancel{\vec{\theta_1}, \vec{\theta_2}, \vec{\theta_3}}, \varepsilon\right\}$$
$$\Pr\left\{W_1 = w_1 \mid C = \mathcal{D}_c, \vec{\theta_1}, \cancel{\vec{\theta_2}, \vec{\theta_3}, \varepsilon}\right\}$$
$$\Pr\left\{W_2 = w_2 \mid C = \mathcal{D}_c, \vec{\theta_2}, \cancel{\vec{\theta_1}, \vec{\theta_3}, \varepsilon}\right\}$$
$$\Pr\left\{W_3 = w_3 \mid C = \mathcal{D}_c, \vec{\theta_3}, \cancel{\vec{\theta_1}, \vec{\theta_2}, \varepsilon}\right\}$$

In the E-step, we perform inference on the eight possible combinations of $Win~1$, $Win~2$, and $Win~3$ and make soft-assignments to each of the eight versions each datapoint.

In the M-step, we perform maximum likelihood estimation of parameters $\vec{\theta_1}$, $\vec{\theta_2}$, $\vec{\theta_3}$, and ε using the same algorithms as if $\mathcal{D}$ had completely observed $W_1 = \mathcal{D}_1$, $W_2 = \mathcal{D}_2$, and $W_3 = \mathcal{D}_3$ (weighted by the soft assignments of the E-step).

During the M-step, we can take advantage of global decomposition allowing us to log-maximize each section of the network separately:

$$\ell = \left(\sum_{\mathcal{D}} \ell\left\{R = \mathcal{D}_r \mid W_1 = \mathcal{D}_1, W_2 = \mathcal{D}_2, W_3 = \mathcal{D}_3, \varepsilon\right\}\right)$$
$$\left(\sum_{\mathcal{D}} \ell\left\{W_1 = \mathcal{D}_1 \mid C = \mathcal{D}_c, \vec{\theta_1}\right\}\right)$$
$$\left(\sum_{\mathcal{D}} \ell\left\{W_2 = \mathcal{D}_2 \mid C = \mathcal{D}_c, \vec{\theta_2}\right\}\right)$$
$$\left(\sum_{\mathcal{D}} \ell\left\{W_3 = \mathcal{D}_3 \mid C = \mathcal{D}_c, \vec{\theta_3}\right\}\right)$$

$C = \mathcal{D}_c$ is the $OnCourt$ variable. It is always observed and assumed to have a uniform prior so it cancels out during any arg max operation.

4.1   M-Step: Maximum Likelihood ε

Using the table CPD for from **Section 3.2** for $\Pr\{R = \mathcal{D}_r \mid W_1 = \mathcal{D}_1, W_2 = \mathcal{D}_2, W_3 = \mathcal{D}_3, \varepsilon\}$ and collecting like terms, we find:

$$\underset{\varepsilon}{\mathrm{argmax}}\left\{\prod_{\mathcal{D}} \Pr\left\{R = \mathcal{D}_r \mid W_1 = \mathcal{D}_1, W_2 = \mathcal{D}_2, W_3 = \mathcal{D}_3, \varepsilon\right\}\right\}$$

$$= \operatorname*{argmax}_{\varepsilon} \left\{ (1 - 3\varepsilon)^{M_{\text{modelled}}} (\varepsilon)^{M_{\text{noise}}} \right\}$$

with solution

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

where

$$
\begin{aligned}
M_{\text{modelled}} \triangleq \ & \text{M}\left[r^3, w_3^1\right] \\
& + \text{M}\left[r^2, w_3^0, w_2^1\right] \\
& + \text{M}\left[r^1, w_3^0, w_2^0, w_1^1\right] \\
& + \text{M}\left[r^0, w_3^0, w_2^0, w_1^0\right]
\end{aligned}
$$

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

4.2   M-Step: Maximum Likelihood θ_i

In this project, we implement the two major pairwise comparison models: Bradley-Terry, and Thurstone Case V.

In Bradley-Terry, each *Win i* random variable follows a logistic distribution. Let

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

so that

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

where

$$\Delta_i \triangleq \left[\theta_{i,\text{Off.P1}}, \ldots, \theta_{i,\text{Def.P12}}\right] \vec{C}$$

and the *OnCourt* variable $\vec{C}$ is a vector of indicator functions. For any particular possession, exactly ten elements of $\vec{C}$ are 1 (there are ten basketball players on the court at once). All the other elements are zero.

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 logistic regression in general, and weighted logistic regression due to the soft-assignments of the E-step. We use a weighted variant of the basic Newton-Raphson logistic regression technique [10].

In Thurstone Case V, each *Win i* random variable follows a Gaussian distribution.

$$\Pr\left\{ W_i = w_i^0 \mid \vec{\theta}_i \right\} \triangleq \Phi\left(\Delta_i / \sigma\right)$$

replaces the logit function with the probit function. Again, we use a weighted variant of [11], the basic Newton-Raphson procedure for probit regression. Here, $\sigma^2 = 10$ as player performance is assumed to be unit variance, and outcome is the sum of 10 player performances.

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 $\Pr\left\{ \mathcal{D}_r, \mathcal{D}_c, w_1^0, w_2^0, w_3^0 \mid \vec{\theta}, \varepsilon \right\}$

Since each datapoint has only eight possibilities, we easily compute the exact probability $\Pr\left\{ \mathcal{D}_r, \mathcal{D}_c, w_1^1, w_2^1, w_3^1 \mid \vec{\theta}, \varepsilon \right\}$ using the chain rule for Bayesian networks.

4.4   Expectation-Maximization: Initialization

We will experiment with a few initialization strategies. One possibility is to initialize all skills are the same. Here,

$$\theta = \vec{0}$$

suggesting $\Pr\{W_i\} = 0.5$ for all i, and therefore

$$\varepsilon = \frac{1}{3} \frac{\text{M}\left[r^3\right]/2 + \text{M}\left[r^2\right]/4 + \text{M}\left[r^1\right]/8 + \text{M}\left[r^0\right]/8}{\text{M}}$$

Another approach is to realize that soft assignments to, e.g. W_3 represent roughly the probability of scoring three points during a possession. Then we would initialize $\Pr\left\{W_3 = w_3^1\right\}$ to the fraction

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

Similarly, soft assignments to W_3 and W_2 suggest that

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

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

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

And for the same reason,

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

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

4.5 Implementation Considerations

Computing the E-Step maximum likelihoods of θ are not closed-form but require iteration and use of the numerical approximation method Newton-Raphson, which uses the gradient and inverse of the Hessian to reach the maximum log-likelihood. The majority of the running-time is spent in these functions, computing the Moore-Penrose pseudoinverse of the Hessian.

Of additional concern is the problem of overshooting the maximal point and actually decreasing our log-likelihood by updating θ by too large a distance. In order to prevent this, whenever we detect a log-likelihood that is lower than our previous, we halve the step size and repeat until we either increase the log-likelihood or are satisfied that we have arrived at the maximum.

With the logit and probit functions, numerical underflow is also an issue - the tail ends of distributions tend to infinitesimal probabilities. We prune out these points when we detect them in order to avoid underflow and NaN results.

5 Results

Our initial run was one game from 2007, when the Detroit Pistons played at Cleveland. For the single game, our epsilon was on the order of 10e-7. After that, we ran on all Western Conference Southwest Division games from the 2008-2009 NBA season. With the logit function, we achieved an epsilon of .0266 and a log-likelihood of -2955.9, while the probit function achieved an epsilon of .0154 and a log-likelihood of -2969.2. The Southwest Division games encompassed 71 players, so we fit 71*6 = 426 parameters in total, from 3152 possessions.

5.1 Evaluation

We tried separating 25% of the possessions for testing and 75% of the possessions for training. When doing this, we notice that the empirical expectation

$$\mathrm{E}\left[R\right] \approx \frac{1}{m}\sum_{R} r\,\mathrm{M}\left[r\right] = 1.0486$$

is 1.0486 points per possession (over 2364 possessions) in the training data, yet it predicts

$$\mathrm{E}\left[\mathrm{E}_{\mathrm{training}}\left[R|C\right]\right] = 1.0183 \text{ or } 1.0217$$

points per possession (over 788 possession) for the possessions in the test set.

The actual number of points scored in the test set was 1.0178, which indicates that the model is accounting for player skills.

Prediction against test

Estimated Points Per Possession
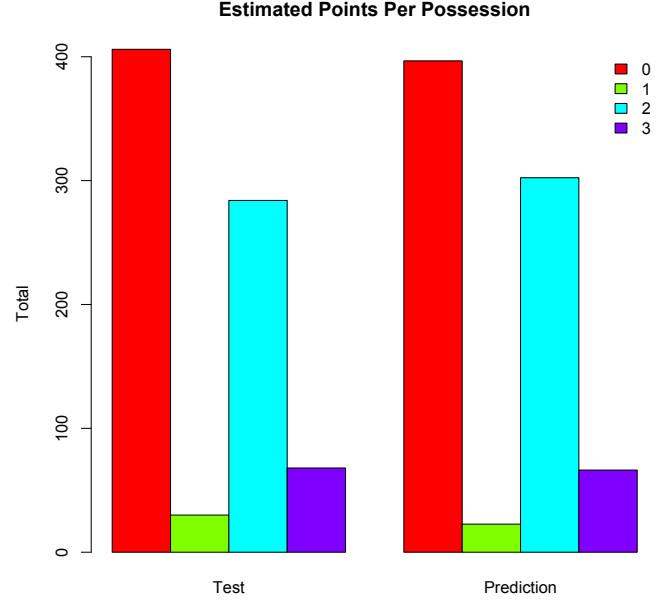
6 Analysis

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

Looking at our single game data on February 4th, 2007 between Detroit and Cleveland, we can see that Billups, Maxiell, James, Wallace, and Varejao were the top 5 2 point offensive players. There were not a lot of 3 point possessions in the game (only 13), so the more efficient scorers were rewarded with high 2 point offensive skill numbers. Billups shot only 4/9 from the floor but by shooting 9/10 from the FT line, this equated to 4 additional 2 point plays. Maxiell only played 5 minutes but in those few possessions, the team converted on both 2 point and 1 point possessions. The bottom 5 2 point defense that day were Snow, Hunter, Marshall, Gibson, and Webber. When these players were on the court, they allowed their opponents to score more than the expected number of 2 point plays. What is interesting to note here is that Marshall and Gibson have better than average 2 point offensive skills but lower than average 2 point defensive skills.

As we trained with more data, including all games in the head to head between Detroit and Cleveland that year, the offensive and defensive skills started to converge closer to the mean. This is due to reducing variance that can occur with one game and few minutes. For example, Maxiell, who only played 5 minutes in the previous game no longer remained in the top 5 of the 2 point offensive skills. The bottom 5 2 point defensive skills were dominated by Cleveland players. We can see that Detroit got the best of Cleveland this year, winning 3 out of 4 games and only losing a close one in OT.

"[NBA West Southwest Intradivision games] In the 2008-2009 NBA season, Shane Battier was on the Second team NBA All-Defensive Team. However, he had a non-adjusted +/- of -18 in the games he played against the Southwest division, which is reflected in the results of only -0.255, and

a below average 1.460 in his defensive skill ranking against 2 pt and 3 pt, respectively.

Surprisingly, Ryan Bowen had 1 pt defensive skills of -932.78, 2 pt defensive skills of -2.16, and 3 pt defensive skills of -18.774. Bowen only participated in one of seven of his team's intradivision games, playing 16 minutes and ending with a non-adjusted $+/-$ of -3. However, when he was on the floor, the opponent had 32 possessions but only $R(3) = 2$ (6.25% of possessions) which is lower than the typical 8.48

(Jekyll and Hyde like numbers). On offense, Matt Carroll posted a 1 pt offensive skill of -13.171, 2 pt offensive skill of 2.414, and 3 pt offensive skill of -19.257. $+2.414$ puts him second out of all players in intradivision games in the Southwest division. His defensive skill for 1 pt was –170.3, 2 pt 2.665, and 3 pt -18.11. So why the disparity in the numbers? It turns out that Carroll only played in one game for about 5 minutes. While in the game, his opponent never scored 1 point or 3 points on any of their possessions, and hence the "great" defensive skill. On the offensive end, Carroll's team also did not score 1 point or 3 points on any of their possessions. They did convert on 4/9 (44.4%) possessions for 2 points, when the league average is closer to 38

7 Conclusions

Bayesian prior/smoothing Coach skill - can add coach as another variable, and see if they are using the most optimal lineups against their opponent. More binary response models

8 References

References

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

[2] R. Bradley and M. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[3] 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.

[4] 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.

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

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

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

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

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

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

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