Sports Analysis: Hierarchical Modeling of Basketball Game Scores

Jeffrey Bond Wang[1], Jordan Hoffmann[1]

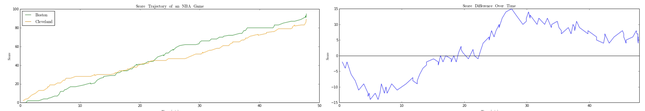
[1]Dept. of Applied Mathematics, Harvard University, Cambridge, MA 02138

Objective

To model the distribution of predicted outcomes of an NBA game based on the remaining time and the current scoring rate of the two teams based on play-by-play analysis.

Motivation

Every year, millions of people around the globe tune into various sporting events. Popular sports use a wide variety of scoring strategies and have significantly different tempos. For example, a soccer game with only 1 or 2 total goals stands in sharp contrast to an NBA game where when combined, the two teams commonly score in excess of 200 points! We seek to model how different changes in score (and at different times) affect the eventual outcome of various common sports. Given that the NBA playoffs are currently ongoing, for this poster we focus on the NBA.



Model Theory and Parameters

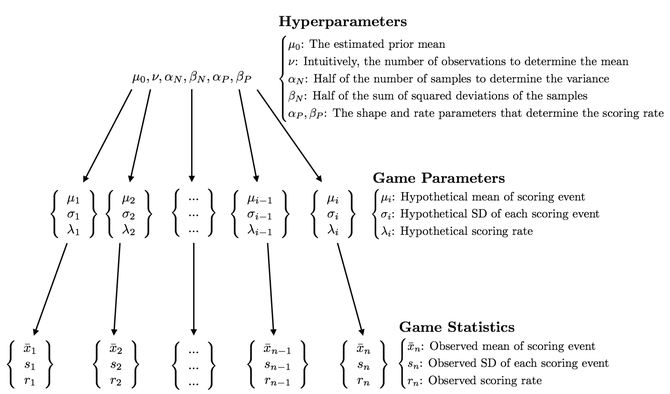
Define the change in score differential with t minutes remaining to be $S(t)$. We define positive score differential to be favoring the home team. We go ahead and model the differential to be given by:

$$P(S(t)=s|\mu,\sigma,\lambda)=\sum_{j=1}^{\infty}\mathcal{N}\left(\frac{s}{j};\mu,\frac{\sigma}{\sqrt{j}}\right)\text{Poisson}(j;\lambda t)+\begin{cases}e^{-\lambda t} & \text{if } s=0\\ 0 & \text{else}\end{cases}$$

We can also define our priors based on the conjugate priors of our distribution as

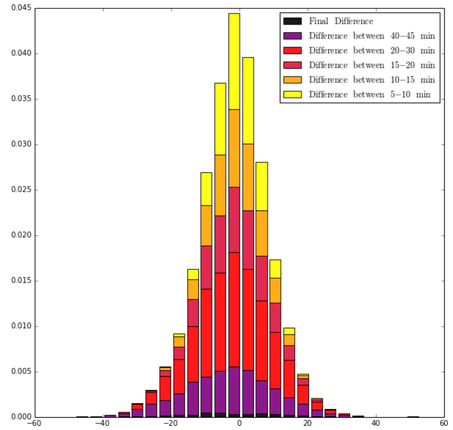
$$(\mu,\sigma)\sim\text{NormalInverseGamma}(\mu_0,\nu,\alpha_N,\beta_N),$$
$$\lambda\sim\text{Gamma}(\alpha_P,\beta_P).$$

Using this, we can design a hierarchical model of the form:

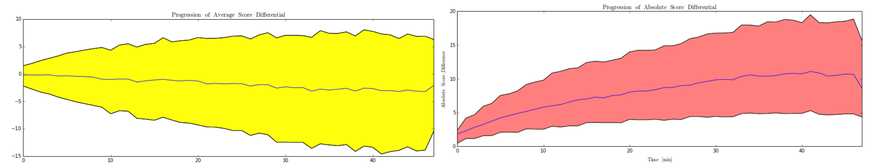


Data Analysis

- We use a dataset from the 2009-2010 NBA season where the Boston Celtics won the Eastern Conference.
- The dataset provides play-by-play analysis giving the score and time for every scoring event in each game.
- We begin by attempting to gain an intuition for the dataset by looking at various parameters:
- **Right:** we show the distribution of score difference after a set amount of time



- **Below:** On the **left** we show the mean score differential over time. As expected, the mean is around 0 at all times. The yellow region shows the signed RMSD from the mean as a function of time. As expected, the value grows with time. On the **right**, the plot is the same except the absolute value of score difference is considered.

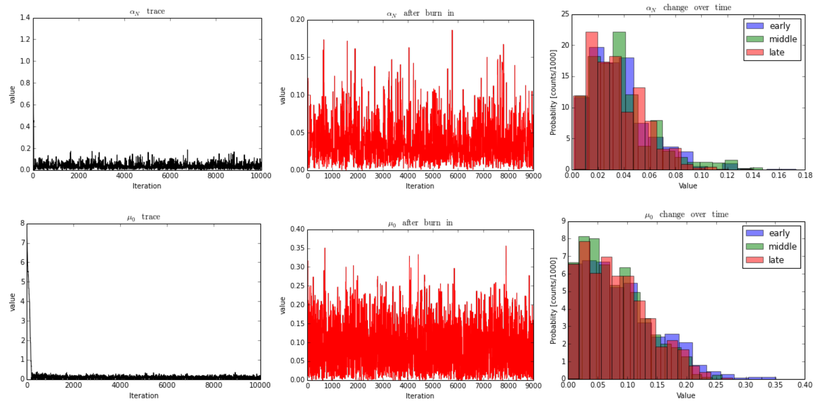


Part 1: Training Hyperparameters

Using a Gibbs sampler, we can derive the mean of our hyper parameters for each team. For the Boston Celtics, we find that:

$$\langle\alpha_N\rangle=0.033,\langle\beta_N\rangle=0.091,\langle\alpha_P\rangle=0.42,\langle\beta_P\rangle=8.4,\langle\mu_0\rangle=0.080,\langle\nu\rangle=13.4$$

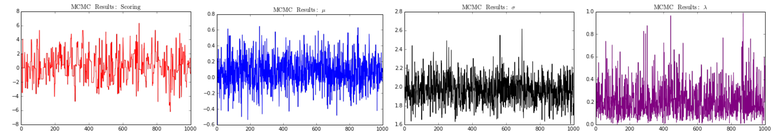
Below: we plot a selection of the output showing the convergence of three of our hyperparameters. In the **left** column we plot the entire trace. In the **middle** column we plot the period after burn in. In the **right** column we plot histograms showing the sampled distribution at various parts of the trace.



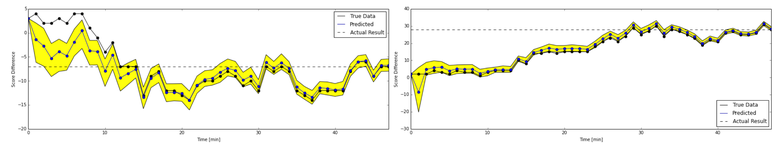
Part 2: Predicting Game Results

Using the mean of the hyperparameters trained in the first part, we utilized a Gibbs Sampler and Metropolis Hasting algorithm to allow one to predict game results.

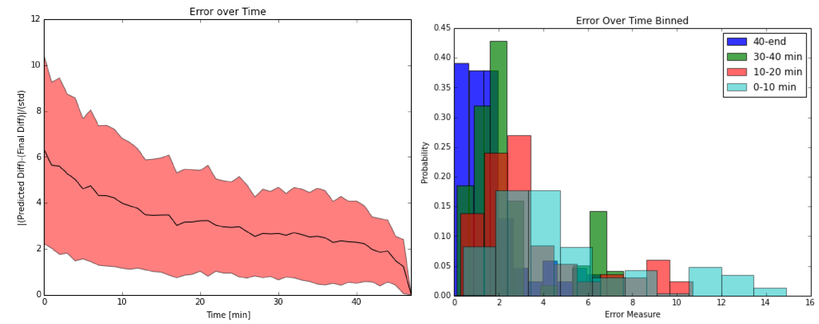
- We find that our distributions drawn using a MCMC converge nicely:



- We can then compare our predicted outcome of a basketball game to the true outcome as a function of time. **Below** we show two games (on the **left** CHA-BOS and on the **right** BOS-CHI). In both cases, we see some volatility early on, and then convergence towards the true final outcome.



- The Boston Celtics played 81 games in the 2009-2010 season. We can run our predictive model using the play-by-play data from each game and compute μ and σ of the predicted outcomes at every minute during each game. We can then compute: $|\mu_{\text{predictions}}-\Delta|/\sigma_{\text{predictions}}$ where Δ represents the final score difference.



Future Work

The model that we have made is general enough to be applied to any sport that has a scoring structure where over time points are accumulated. This would allow our model to be used to study many of the world's most popular sports including football (soccer), hockey, and American football (to name a few). We also plan on doing an analysis where we look at how the model changes when trained on different NBA teams.

References

Parsed data: `eecs.harvard.edu/~elaine/sousvide/stories/basketball.html`

Email: jeffreywang@college.harvard.edu, jhoffmann@fas.harvard.edu