STATS 551 Project Proposal Presentation

Using a Bayesian Hierarchical Model to Predict NBA Scores

Andrew Soncrant
asoncran@umich.edu

Taylor Spooner
spoonert@umich.edu

University of Michigan

Agenda

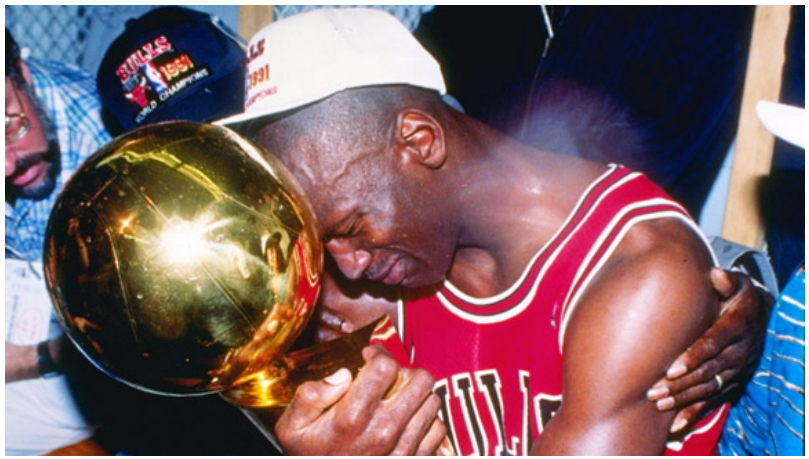
1. Motivation and Purpose
   - What is the problem of interest?
   - Why we chose this project?
2. The Data
   - What is the data and how it was collected.
   - Simple facts: size, structure, etc.
3. Plan for data analysis
   - The data analysis model
   - Preliminary analysis
   - Time line
   - Potential difficulties

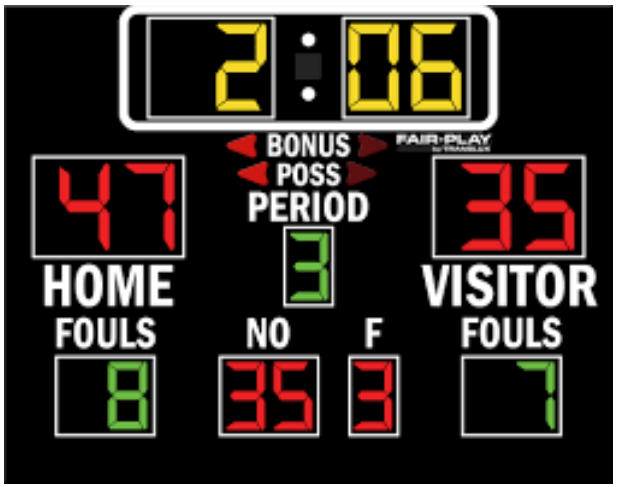
Agenda

1. Motivation and Purpose
   - What is the problem of interest?
   - Why we chose this project?
2. The Data
   - What is the data and how it was collected.
   - Simple facts: size, structure, etc.
3. Plan for data analysis
   - The data analysis model
   - Preliminary analysis
   - Time line
   - Potential difficulties

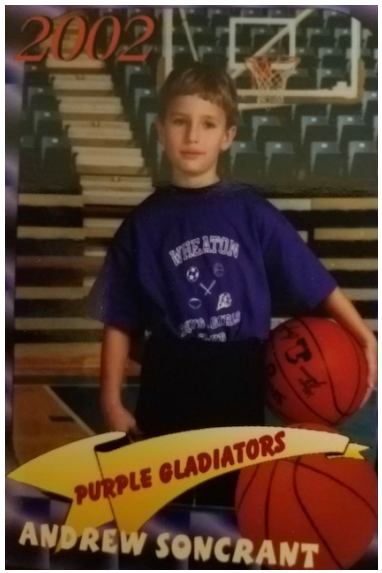
What is the problem of interest?

- Bayesian hierarchical methods to predict NBA Scores
- Use past data to create informative priors and clusters for present data
- Gain inference on overall home court advantage in NBA and offensive/defensive abilities for each team
- Predict scores and outcomes for NBA games





Why did we choose this problem?



- Statistical modeling for sports is a growing topic
- Both authors are fans of sports
- Extend results from Baio and Blangiardo [1]
  - Different sport where are points valued less
  - Additional techniques to improve prediction
- Yet to find NBA data modeled in this way

Agenda

1. Motivation and Purpose
   - What is the problem of interest?
   - Why we chose this project?
2. The Data
   - What is the data and how it was collected.
   - Simple facts: size, structure, etc.
3. Plan for data analysis
   - The data analysis model
   - Preliminary analysis
   - Time line
   - Potential difficulties

Data Source

- Source: `www.hoopsstats.com`
  - Data gather using web scraping
- Game box score stats for every regular season game in the 2015-2016 and 2016-2017 season
  - 1,230 games
  - 30 total NBA teams
  - Each team plays every other team at least once at home and on the road

Basic Structure

- Each row of the dataset represents one game (unlike the below picture)
- The two teams playing in the game, identifying which team was at home and which was away
- Response variables: Final scores for the home and away team
- 18 other box score statistics (rebounds, assists, etc.) for both the home and away team to be used as possible explanatory variables

Date	Opponent			Min	Pts	Reb	Ast	Stl	Blk	To	Pf	Dreb	Oreb	Fgm-a
Apr 28	vs Washington	L	99-115	48	99	35	26	9	4	22	22	24	11	37-78
				48	115	33	21	16	7	16	22	25	8	42-78
Apr 24	vs Washington	W	111-101	48	111	49	24	7	1	12	20	37	12	40-90
				48	101	46	19	7	5	14	25	37	9	36-85

The Use and Problem with Two Seasons

Slight look ahead

A major component of this project will be to create informative priors using existing data. In order to accomplish this we will use the 2015-2016 season to create informative priors and clustering for team performance.

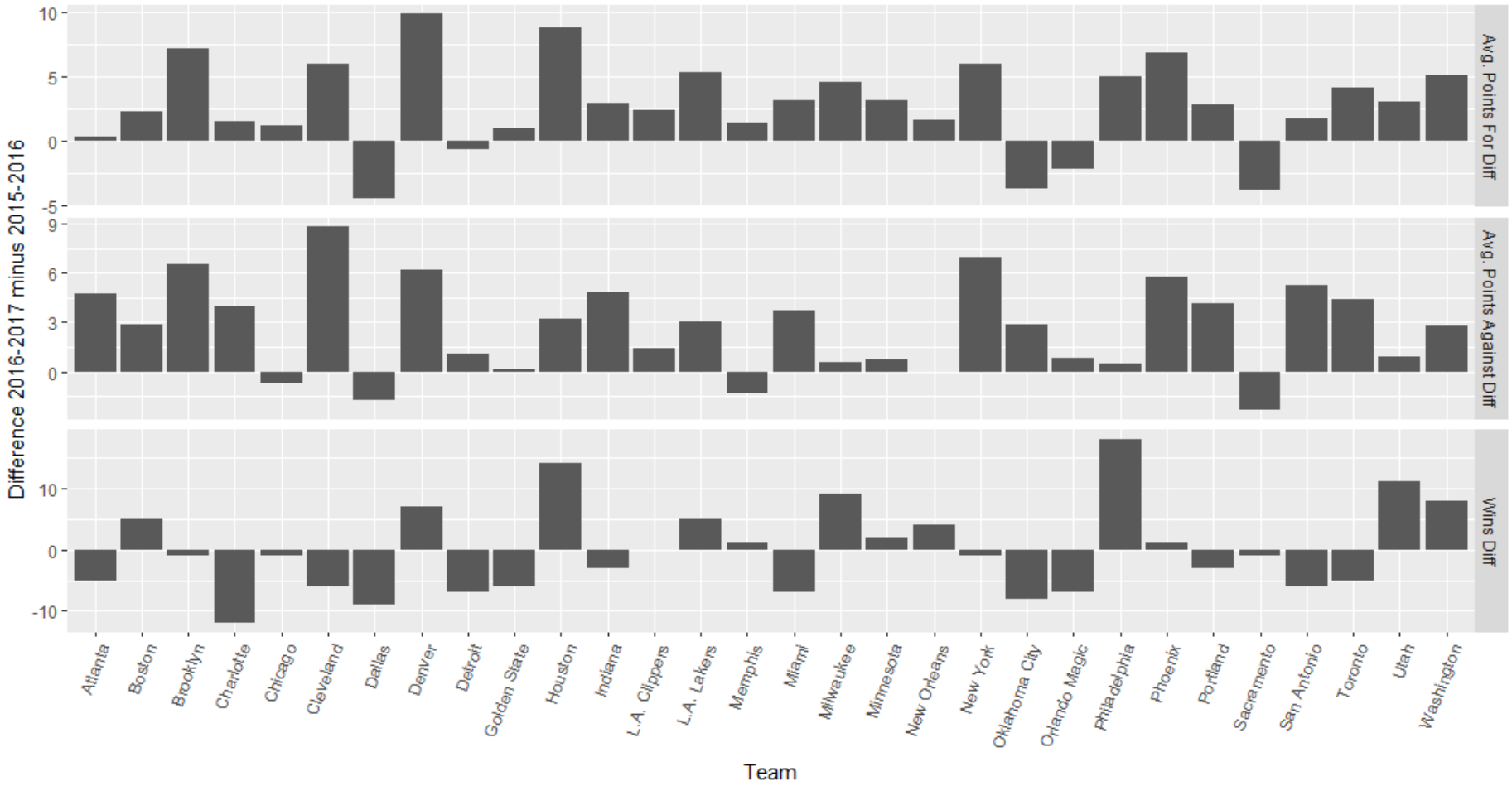
The Use and Problem with Two Seasons

Slight look ahead

A major component of this project will be to create informative priors using existing data. In order to accomplish this we will use the 2015-2016 season to create informative priors and clustering for team performance.

- Dependency or lack that of between seasons
- Considered splitting one season in half, using the first half of the season to inform the second half
  - Problem: structure of home/away series for each team would then not hold

The Use and Problem with Two Seasons



Median Changes (Points): Avg. Points Diff. = 2.8, Avg. Points Against Diff. = 2.8, Wins Diff = -1

Agenda

1. Motivation and Purpose
   - What is the problem of interest?
   - Why we chose this project?
2. The Data
   - What is the data and how it was collected.
   - Simple facts: size, structure, etc.
3. Plan for data analysis
   - The data analysis model
   - Preliminary analysis
   - Time line
   - Potential difficulties

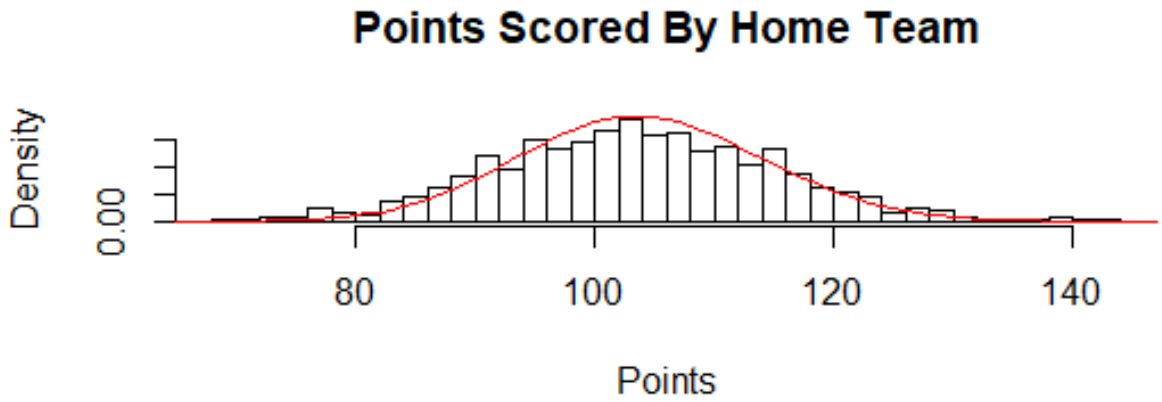
Motivation

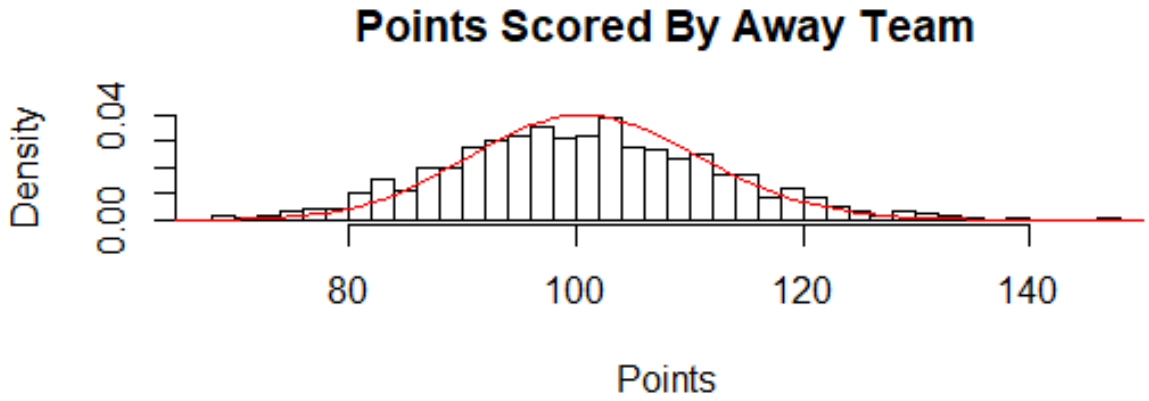
Model structure built off of the results from [1]

- Each game contains a pair or observed values, $\mathbf{y} = (y_{g1}, y_{g2})$ where y_{g1} and y_{g2} represent the points scored by the home and away teams respectively for each game $g \in 1, \ldots, G = 1230$.
- As games are independent of one-another, a Poisson model seems appropriate for modeling $\mathbf{y}$, that is, $y_{gi} \sim \text{Poisson}(\theta_{gi})$, $i \in 1, 2$. $(\theta_{g1}, \theta_{g2})$ represent the scoring rates for the home and away team in game $g - th$, respectively.

Initial Data Analysis: Is a Poisson model reasonable?

Here we plot home and away scores, and overlay the corresponding Poisson densities (using MLE estimates). Notice slight overdispersion, which we account for using a Poisson-logNormal model as per Millar [5].





The Model

The Model: Poisson-logNormal

$y_{gi} \sim \text{Poisson}(\theta_{gi})$

- $\log(\theta_{g1}) = \text{home} + off_{h(g)} + def_{a(g)}$
- $\log(\theta_{g2}) = off_{a(g)} + def_{h(g)}$

Where $h(g)$, $a(g)$ uniquely identify the home/away teams in the gth game.

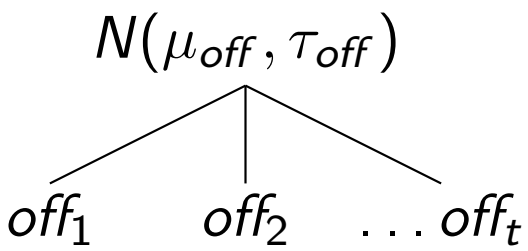
In words, we're modeling home team scoring as the home team's offensive ability plus the away team's defensive ability (which ideally is negative), plus some "home court advantage" parameter. We similarly model the away team's scoring, without home-court advantage.

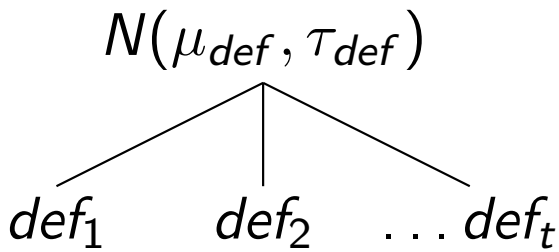
Priors– Initial Model

Parameters

We invoke a mixed-effects model. The home parameter takes a Gaussian prior with a large variance. The team specific off_t, def_t parameters are Gaussian from a common distribution.

- $home \sim N(0, 0.0001)$
- $off_t \sim N(\mu_{off}, \tau_{off})$
- $def_t \sim N(\mu_{def}, \tau_{def})$





Priors– Initial Model

Parameters

We invoke a mixed-effects model. The home parameter takes a Gaussian prior with a large variance. The team specific off_t, def_t parameters are Gaussian from a common distribution.

- $home \sim N(0, 0.0001)$
- $off_t \sim N(\mu_{off}, \tau_{off})$
- $def_t \sim N(\mu_{def}, \tau_{def})$

Identifiability constraints on team-specific parameters[1, 2]:

$$\sum_{t=1}^{T} off_t = 0, \sum_{t=1}^{T} def_t = 0$$

Priors– Initial Model

Parameters

We invoke a mixed-effects model. The home parameter takes a Gaussian prior with a large variance. The team specific off_t, def_t parameters are Gaussian from a common distribution.

- $home \sim N(0, 0.0001)$
- $off_t \sim N(\mu_{off}, \tau_{off})$
- $def_t \sim N(\mu_{def}, \tau_{def})$

Hyperparameters

$\mu_{off} \sim \text{Normal}(0, 0.0001)$
$\tau_{off} \sim \text{Gamma}(0.1, 0.1)$

$\mu_{def} \sim \text{Normal}(0, 0.0001)$
$\tau_{def} \sim \text{Gamma}(0.1, 0.1)$

Priors– Informative Model

- Using the posterior mean and variances for each team's offensive and defensive parameters, we hope to create a more informative model for the 2016-2017 season
- The model will stay the same, however the hyperparameters will now be a vector of values indexed by the team number and not have distributions themselves

$$home \sim N(\mu_{home}, \tau_{home}) \quad off_t \sim N(\mu_{off(t)}, \tau_{off(t)})$$
$$def_t \sim N(\mu_{def(t)}, \tau_{def(t)})$$

Where for a given $t = 1, \cdots, 30$, $\mu_{off(t)}$, $\tau_{off(t)}$, $\mu_{def(t)}$ and $\tau_{def(t)}$ are just numbers (the posterior means of their respective distributions from our 2015-2016 model) and μ_{home}, τ_{home} are the posterior mean and precision of the "home-court" advantage.

Overshrinkage in Hierarchical Models

- A problem of with hierarchical models is the consequence of over-shrinkage. With sports analysis it is seen in the two following ways: [1, 2]
  - Penalizing the good teams with a low mean prior distribution compared to ability
  - Awarding the bad teams with high mean prior distribution compared to ability

Mixture Model

Solution to overshrinkage: Mixture Model

- Define three groups of teams (good, medium, bad) for both offense and defense, obtain probabilities that each team is in that group
- For each team t, define two latent variables, $grp^{off}(t)$ and $grp^{def}(t)$ which can take values 1, 2, 3 for the three specified groups
- Find probabilities $\boldsymbol{\pi}^{\boldsymbol{off}} = (\pi_{1t}^{off}, \pi_{2t}^{off}, \pi_{3t}^{off})$ and $\boldsymbol{\pi}^{\boldsymbol{def}} = (\pi_{1t}^{def}, \pi_{2t}^{def}, \pi_{3t}^{def})$ that each team is in each group

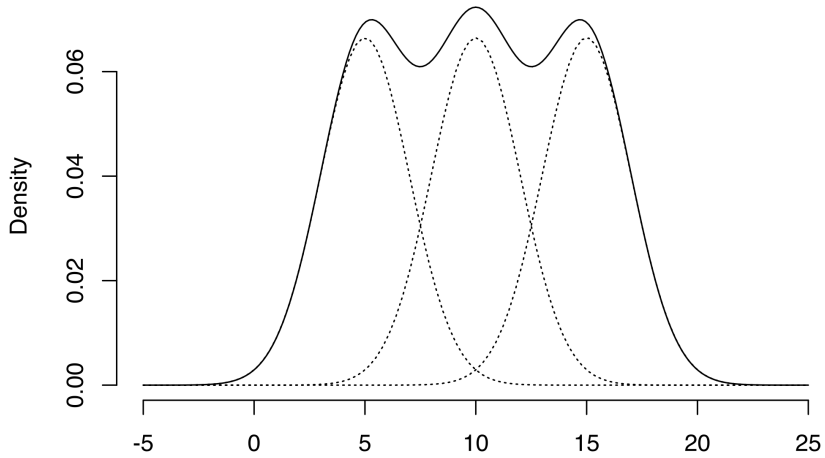
Mixture Model

Solution to overshrinkage: Mixture Model

- Define the offensive and defensive effects as such:

$$off_t = \sum_{k=1}^{3} \pi_{kt}^{att} \times nct(\mu_k^{att}, \tau_k^{att}, \nu = 4),$$
$$def_t = \sum_{k=1}^{3} \pi_{kt}^{def} \times nct(\mu_k^{def}, \tau_k^{def}, \nu = 4)$$

Where $nct()$ is a non-central t distribution as proposed by [1] and each hyperparameter is given its own distribution



Clustering: Posterior $\Rightarrow$ Prior

For each team, we use their 2015-2016 cluster distribution posterior means $(\hat{\pi}_t^{off}, \hat{\pi}_t^{def})$ to set the prior for 2016-2017. In this way, one season informs the next:

Offense

$$\pi_1^{off} \sim dir(\hat{\pi}_1^{off})$$
$$\vdots$$
$$\pi_t^{off} \sim dir(\hat{\pi}_t^{off})$$

Defense

$$\pi_1^{def} \sim dir(\hat{\pi}_1^{def})$$
$$\vdots$$
$$\pi_t^{def} \sim dir(\hat{\pi}_t^{def})$$

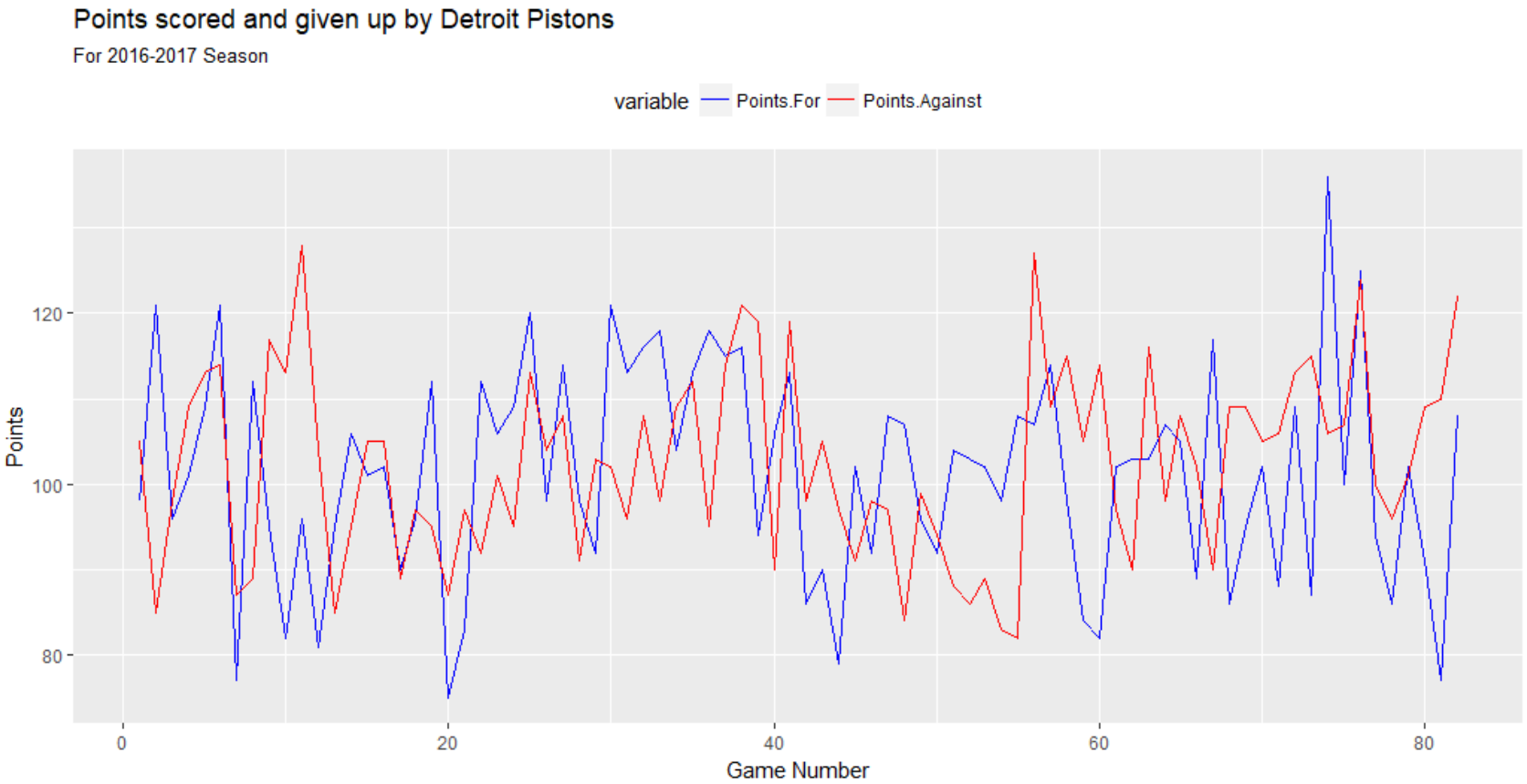
Comparison of Models

At the end of our analysis we will have created four models to gain inference and compare for the 2016-2017 NBA season:

1. Non-informative priors
2. Mixture Model, non-informative priors and group probabilities
3. Informative priors using 2015-2016 data
4. Mixture Model, informative priors and group probabilities

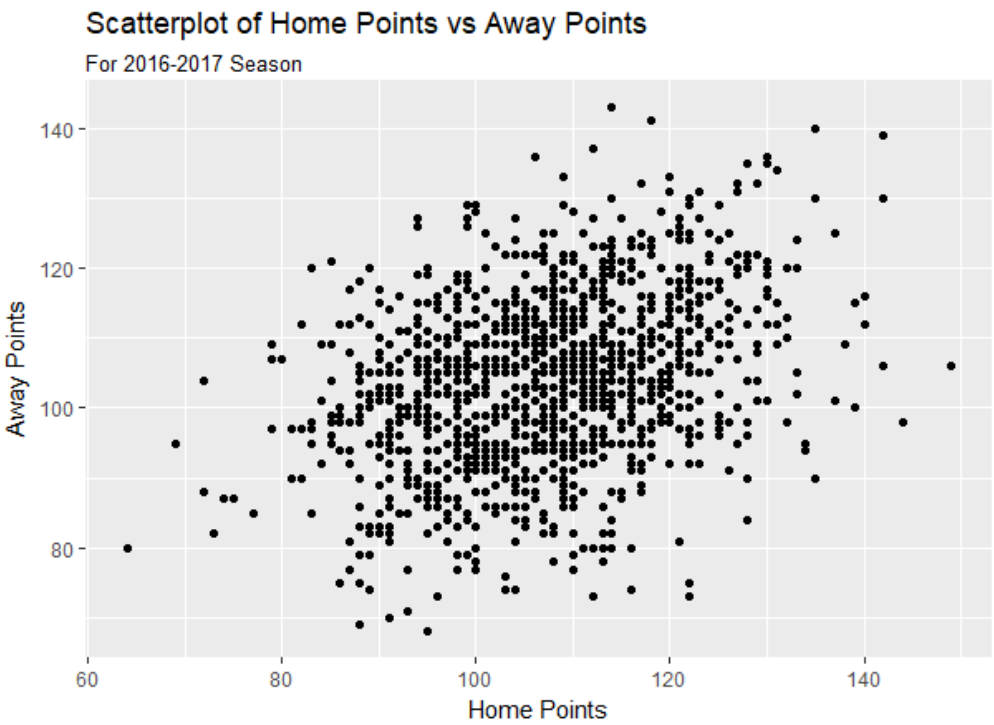
Compare each model's predictive performance to evaluate performance.

Difficulty– Variance in basketball scores



$$sd(Points.For) = 12.5 \quad sd(Points.Against) = 11.04$$

Potential Difficulty/Future Work



$r = 0.355$

- Correlation between home and away scores due to pace of play
- Bivariate Poisson likelihood accounts for scoring correlation [2]
- Use box score statistics (rebounds, turnovers, assists, PACE [3], etc.) to better predict offense and defense effects.

References I

[1] Baio, G. & Blangiardo, M. (2010), "Bayesian hierarchical model for the prediction of football results", *Journal of Applied Statistics* **37**, 253 - 264.

[2] Dimitris Karlis and Ioannis Ntzoufras. (2003), "Analysis of Sports Data by Using Bivariate Poisson Models", *Journal of the Royal Statistical Society. Series D (The Statistician)* vol. 52, no. 3, 2003, pp. 381393.

[3] "Hollinger's NBA Team Statistics." *ESPN*, www.espn.com/nba/hollinger/teamstats.

[4] *Google Cloud & NCAA ML Competition 2018-Men's*, Kaggle, www.kaggle.com/c/mens-machine-learning-competition-2018

References II

[5] Millar, Russell B. (2009), "Comparison of Hierarchical Bayesian Models for Overdispersed Count Data Using DIC and Bayes' Factors." *Biometrics*, vol. 65, no. 3, 2009, pp. 962-969.

[6] "NBA Fantasy Basketball Stats." `www.hoopsstats.com`

[7] Zimmermann, Albrecht. (2016), "Basketball predictions in the NCAAB and NBA: Similarities and differences". *Stat. Anal. Data Min.* **9**, 5 (October 2016), 350-364.