STATS 551 Final Project

Using a Bayesian Hierarchical Model to Predict NBA Scores

Soncrant, Andrew
asoncran@umich.edu

Spooner, Taylor
spoonert@umich.edu

1 Motivation and Purpose

Statistical modeling for sports is a growing topic and one that interests both of the authors greatly. To explore this field of research we focused on the results from 2010 paper by Gianluca Baio and Marta Blangiardo [1]. This paper used Bayesian hierarchical modeling to predict the results of soccer games by estimating the number of goals scored by the home and away teams using an attacking and defending rate parameter. The goal of our project is to extend the results of this paper to a different sport: NBA basketball. We chose basketball as a sport because 1) we aim to see how the models and ideas from [1] translate to a much higher scoring game, where the value of a goal/basket is not as high and 2) while the NBA is like soccer in that each team plays every other team both at home and away, the NBA has additional games as well. The NBA schedule consists of more games than just the home and away series with every other team and those games are determined by the division of each NBA team. Finally, the third reason we chose predicting NBA basketball is because from what we can tell, while there has been research done to predict outcomes of basketball games [4, 7], we have not seen much research on predicting NBA games using a Bayesian setting, especially in the method that we will use.

In addition to extending the ideas of Baio and Blangiardo's paper by changing the sport of interest, we will also be using informative priors and clustering techniques for our mixture model which will be described further in Section 3. Our final goal of this project is to use Bayesian methods to quantify the scoring/defending abilities for NBA teams in our model by analyzing the posterior distributions. We will then use our model to predict the outcomes of NBA games and by simulating the season many times. We plan on comparing our projected wins for each team with the number of wins actually observed.

2 The Data

We will be using game by game data from both the 2015-2016 and 2016-2017 NBA regular seasons to perform our analysis. The data was collected from the website `hoopsstats.com`, an online source for a wide range of NBA data [6]. There was no way to download all of the data efficiently, so we created a webscraping script that downloaded, cleaned and saved the data to two `.csv` file, one for each season.

Due to computational constraints, we will not be using the entire season nor will we use all of the teams. Instead we focused on only all of the teams in the "Western Conference" and the games played only amongst those teams. Since these teams are all fighting for playoff positions against each other and play the same number of games against other Western Conference teams, we felt this data will still yield meaningful results.

There are a total of $T = 15$ teams in the NBA's Western Conference that played a total of 390 games against each other over the course of the regular season. Since we have collected data for every game played during the season, we have two instances of the population of data and not a random sample. Thus we do not have to account for the data gathering procedure in our analysis. For each game (a row in the dataset corresponding to the respected season) we have the following information:

- The two teams that played in the game, indicating which team was considered the home team and which was the away team. Additionally, we gave each team an index number to more easily identify teams.
- The final scores for both teams. The number of points for each team will be our response variable in our regression models.

One potential pitfall of this data comes from the fact that we are using two separate seasons. We are going to use the 2015-2016 season to inform our priors for the next season. While we do not believe that NBA seasons are completely independent of each other, we do have some reservations about using a different season as teams gain and lose players that may change the team's ability. Some initial analysis between the two seasons shows that while there are some teams that do make rather large changes in the average points they scored per game, average points they let up and the number of wins, the median changes in these numbers were 2.8 points, 2.8 points and -1 win respectively.

3 The Model

As mentioned in Section 1, our main motive is to extend the results from Baio and Blangiardo's [1] work in soccer to a higher scoring sport, like basketball, and to investigate how adding informative priors changes our predictive ability. Because of that, the formulation of the model comes from those authors' work.

We begin by defining each game's outcome as $\mathbf{y} = (y_{g1}, y_{g2})$ as the number of points scored by the home team, y_{g1}, and the away team, y_{g2}, for each game in our dataset $g = 1, \cdots, 390$. We will make the assumption that each game is independent of each other. Thus, given the scoring rate, θ_{gi} for the home ($i = 1$) and the away team ($i = 2$) we model our likelihood as:

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

After some initial analysis we find that the distribution of home and away scores shows signs of over-dispersion. However, as noted in the works of Baio and Blangiardo [1], Karlis and Ntzoufras [2] and Millar [5], we can model these scoring rates by using a log-linear random effect model:

$$log\theta_{g1} = intercept + home + \text{off}_{h(g)} + \text{def}_{a(g)}$$

$$log\theta_{g2} = intercept + \text{off}_{a(g)} + \text{def}_{h(g)}$$

where the *home* parameter adds a "home-court advantage" to the scoring of the home team and $h(g)$ and $a(g)$ uniquely identify the home/away team in the gth game. Unlike the work in [1] we added the *intercept* parameter to signify the average number of points an average team would score in any given game. This was needed because, unlike soccer, basketball is much higher scoring. Since the team effect parameters where centered (discussed below), in order to predict basketball point totals, the average points in a game needed to be included in the modeled.

In the above cited work, it is discussed that the use of this Poisson-logNormal can be used to analyze over-dispersed count data. Furthermore, we model the scoring rates by summing the team's

scoring effect and the opposing team's defensive effect. A good defensive team would thus have a negative defensive effect as they would be subtracting points from their opponents. As further suggested by Karlis and Ntzoufras [2], to better compare these offensive and defensive effects across teams, we will start by setting priors for team $t = 1, \cdots, 15$ (shown with the mean and variance parameterization):

$$\text{off}_t \sim N(\mu_{off}, 1/\tau_{off}) \qquad \text{def}_t \sim N(\mu_{def}, 1/\tau_{def})$$

and then center our draws such that:

$$\sum_{t=1}^{15} \text{off}_t = 0 \qquad \sum_{t=1}^{15} \text{def}_t = 0$$

The constraint is accomplished by subtracting the mean of the draws of the offense/defensive parameter from each team's individual draw. This identifiability constraint is needed for the team-specific effects to converge.

3.1 Non-Informative Priors

We will first use the 2015-2016 data to run this above model using non-informative priors. That is:

$$\begin{aligned} intercept \sim N(0, 10000), \qquad & home \sim N(0, 10000) \\ \mu_{off} \sim N(0, 10000), \qquad & \mu_{def} \sim N(0, 10000) \\ \tau_{off} \sim \text{Gamma}(0.1, 0.1), \qquad & \tau_{def} \sim \text{Gamma}(0.1, 0.1) \end{aligned}$$

We are also going to run this same, non-informative model using the 2016-2017 data in order to compare our more informative models with the basic non-informative model.

3.2 Informative Priors

Using the model defined in 3.1, we will obtain posterior estimates of the mean and precision of the *home* parameter and off_t and def_t for every team in $t = 1, \cdots, 30$. Let μ_{home} and τ_{home} be the posterior mean and precision of the *home* parameter and $\mu_{off(t)}$, $\tau_{def(t)}$, $\mu_{def(t)}$ and $\tau_{def(t)}$ be the means and precisions for the t-th team. All of these values are just numbers, thus, we lose a level of hierarchy and our informative model has the same likelihood, but priors:

$$intercept \sim N(\mu_{intercept}, 1/\tau_{intercept}), \quad home \sim N(\mu_{home}, 1/\tau_{home}),$$

$$\text{off}_t \sim N(\mu_{off(t)}, 1/\tau_{off(t)}), \quad \text{def}_t \sim N(\mu_{def(t)}, 1/\tau_{def(t)})$$

Our goal with informative priors is to really get into the mindset of Bayesian analysis. We both have found the idea of using our prior thinking and experiments to update and improve our analysis is a powerful tool in Bayesian analysis. For example, if a team is very good offensively in the season before, we believe that we can better estimate their offensive ability in the next season than having no information on that team.

After running this informative model, we found that the variance from the posterior estimates was too small and the estimates from the model were almost exactly the same as the the 2015/2016 results (section 5). Because of that, we decided to use half informative priors and replace all of the variances in the above model with .1 except for $\tau_{intercept}$ which was replaced with 1 because it is on a larger scale. Therefore, our "informative" model is described by:

$$intercept \sim N(\mu_{intercept}, 1), \quad home \sim N(\mu_{home}, .1),$$

$$\text{off}_t \sim N(\mu_{off(t)}, .1), \quad \text{def}_t \sim N(\mu_{def(t)}, .1)$$

4 Mixture Models

Both [1] and [2] note that the type of hierarchical modeling in section 3 leads to over-shrinkage in the offensive and defensive effects. The over-shrinkage occurs in two ways: first by penalizing/dragging down the better teams due to a low mean prior distribution and the second is awarding/bringing up bad teams because the prior mean is higher than the team's actual abilities.

Another motivation for the informative model that we are planning on building (described in 3.2) is to account for this. However, both [1] and [2] have also found great success by defining a mixture model.

We will define two groups of teams according to their ability (good and bad) for both offense and defense ability. For each team, t, two latent variables, $grp^{off}(t)$ and $grp^{def}(t)$, will be created to signify if that team is in group 1 (good), of 2 (bad). The probability that a team will fall into these groups will be represented by $\boldsymbol{\pi_t^{off}} = (\pi_{1t}^{off}, \pi_{2t}^{off})$ and $\boldsymbol{\pi_t^{def}} = (\pi_{1t}^{def}, \pi_{2t}^{deff})$

Once we have determined these probabilities, we will represent our offensive and defensive effects by:

$$off_t = \sum_{gr=1}^{2} \pi_{gr,t} \times \text{nct}(\mu_g^{off}, \tau_g^{off}, \nu) \quad def_t = \sum_{gr=1}^{2} \pi_{gr,t} \times \text{nct}(\mu_{gr}^{def}, \tau_{gr}^{def}, \nu)$$

Where nct() is a non-centralized t-distribution with ν degrees of freedom [1]. The distributions for hyperparameters μ_{gr}^{off}, τ_{gr}^{off}, μ_{gr}^{def}, τ_{gr}^{def} with $gr = 1, 2$ and the probability vectors $\boldsymbol{\pi^{off}}$ and $\boldsymbol{\pi^{def}}$ will be determined in two ways.

Non-informative model: First we will set up a non-informative model using the past season data. The vector probabilities will both have a Dirchlet prior with non-informative parameters $(1, 1)$. The distributions for the hyperparameters will be chosen using sensitivity analysis, finding group cutoffs that give reasonable estimates.

Informative Model: Using the posterior estimates that each is in group gr, we will run the same mixture model on the 2016-2017 data, but this time using a Dirchlet prior with parameters (π_{1t}, π_{2t}) for team t. Additionally, the posterior estimates for the hyperparameters will be used to better estimate the mean and variance effects for each group. However, because the mixing deemed ineffective (discussed below), the results for this model are withheld.

Outside the Model: We can also model teams' average points scored and points allowed over the entire season as coming from a mixture of bivariate Gaussian clusters. We perform standard EM clustering with k=2 clusters (good and bad teams).

5 Results

5.1 Comparison Between Non-Mixture Models

We begin by analyzing the posterior results from the Poisson model using both non-informative and informative priors with 4 chains sampled 25,000 times using the first 12,500 samples as a burn in period. Looking at the diagnostic traceplots for 4 chains, the *home* and *intercept* parameters along with off and def effects for each team all converged. Additionally, the $\hat{R}$ values were all 1 and the effective sample size were all large. Therefore, we feel confident that we sampling from the posterior distribution.

We now look at the posterior mean and 95% credible intervals for the parameters. Remember that to predict points, we exponentiate the sum of all the parameters. The *intercept* parameter, which corresponds to the average number of points scored by an NBA team in a given game, and the *home* parameter, the point advantage the team playing at home gets, are shown in Table 1. The

Table 1: Posterior estimates for the *intercept* and *home* parameter for both the non-informative and informative models.

Model	Intercept			Home		
	2.5%	Mean	97.5%	2.5%	Mean	97.5%
Non-Informative	4.63	4.64	4.65	0.023	0.034	0.047
Informative	4.57	4.64	4.71	0.021	0.035	0.048

interval widths for both parameters are the same. Additionally, the mean estimates are very close to being the same. Surprisingly, we see that the posterior estimate for the *home* parameter is a lot smaller than we expected. Holding all of the other parameters constant in the θ_{g1} equation, the *home* parameter would contribute about $e^{0.034} = 1.03$ points, on average, to the home team's score.

Switching our focus to Figure 1, we compare the posterior estimates of the offense and defensive team effects. The widths of the posterior intervals are about the same for both the non-informative and informative models. We can see that for the really good and really bad teams, the mean estimates go further to the extreme (good teams get better, bad teams get worse). As mentioned in Section 4, a drawback in hierarchical models is over-shrinkage. Even though we don't see too much evidence of this in our model (discussed below), it seems that by adding prior information has allowed for the teams at the top and bottom to better estimate their true abilities.

In Figure 2, our models help us visualize one important aspect of sports: teams change from year to year. While the informative model does not provide drastically different results from the uninformed model, the models run on 2016-2017 are different from 2015-2016. It is worth noting that these results tend to agree with domain knowledge. Thus, more than anything Figure 2 helps us to validate our models before we perform any prediction.

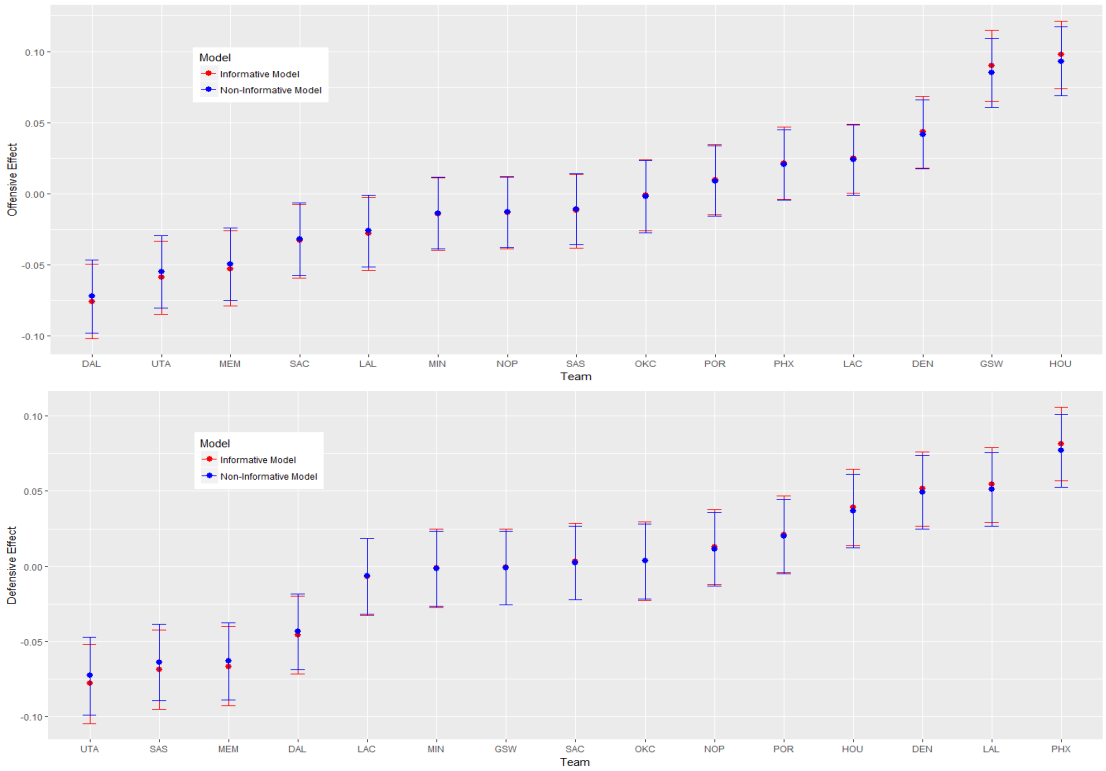

Figure 1: Both figures show the mean and 95% confidence interval for the team effect parameters. The top figure shows the offensive effect (with larger indicating a better offensive team) and the bottom figure shows the defensive effect (with smaller indicating a better defensive team).

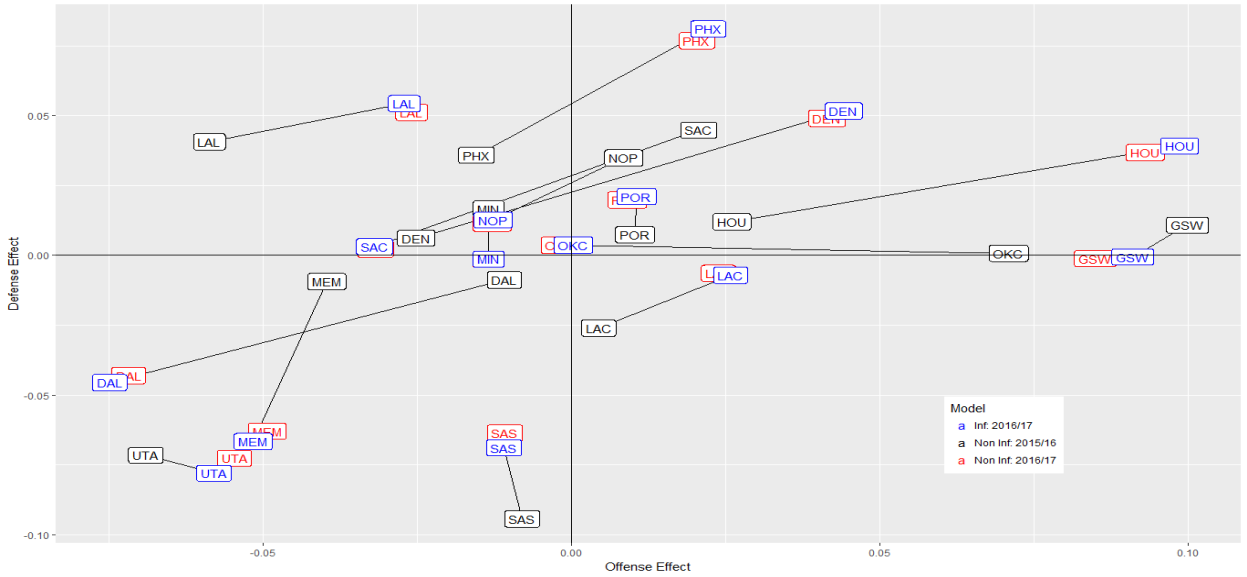

Figure 2: Offense and Defense Effect for the non-informative and informative models. The black team label represents the mean posterior estimate of the offensive and defense effect from the 2015-2016 season (NI 2015/16) while the red team label represents the same but for the 2016-2017 (NI 2016/17) season using non-informative priors. The blue team label is the posterior estimates for the informative model (I 2016/17). A line segment connects NI 2015/16 and I 2016/17 to see the change.

We end our comparison of the two models by predicting the outcome of each game in our 2016/2017 data. To predict, we simulate 50,000 versions of the same season, each time random draw from our posterior model. For each game we predicted the home score and away score, then use those predicted scores to determine the winner of the game. Tables 2 shows the results from our predictions. In Table 2 we see the prediction results for each team from the simulations. First we notice that both models almost perfectly predict the average points per game both scored and let up by each team. However, what is not predicted as well is the number of wins for each team. More detailed, the model under predicts the number of wins for good teams while over predicting the number of wins for bad teams. The prediction accuracy for the Non-Informative model after 50,000 season simulations was 0.596 and the prediction accuracy for the Informative model was 0.605.

To further look into the predicted wins and how we may improve our model we look at Figure 3 which shows the distribution of observed points and predicted points for both models. We saw in Table 3 that while we are able to predict points really well, the number of wins for each team was too low for really good teams and too high for the bad teams. After simulating many seasons, our posterior predicted points for both models are concentrated heavily around the mean number of points; though in any given game the number of points can vary greatly.

As mentioned, we don't suffer from the same over-shrinkage as seen in [1] (in the points scored), but we do tend to shrink the more extreme teams towards the mean in terms of games won. This is most likely due to the asymmetric observed distribution of points scored/allowed for very good and very bad teams. Our off/def posterior distributions for all teams are symmetric, implying a below-average game is just as likely as an above-average one. In reality, teams like Golden State and Houston are very unlikely to have bad games, and vice-versa for bad teams like Phoenix. As a result, our model overestimates the number of bad games for good teams and good games for bad teams (if only slightly), resulting in this over-shrinkage. Despite this slight skewness, the 95% prediction intervals for average wins per season covers the observed number of wins in 2016-2017 for every western conference team.

Another possible explanation is that in 2016-2017 Golden State and Phoenix were historically

Table 2: Posterior predictions for the non-informative model and the informative model after 50,000 simulated seasons. Predictions were made for the total number of points scored and allowed. The table shows the predicted wins (when points for was more than points allowed) and the average points per game. O represents the observed values for each team while NI and I are the predictions for the non-informative and informative models respectively.

Team	Wins			Avg. Points For			Avg. Points Ag.		
	O	NI	I	O	NI	I	O	NI	I
Golden State	42	38.43	38.96	116.23	115.63	116.24	105.33	105.33	105.35
Houston	36	34.29	34.58	116.62	115.95	116.48	109.50	109.22	109.45
San Antonio	36	33.65	34.24	104.96	105.05	104.97	98.75	99.24	98.69
L.A. Clippers	31	30.56	30.80	108.60	108.43	108.54	104.88	104.93	104.83
Utah	31	28.79	28.97	100.23	100.62	100.23	98.06	98.58	98.05
Oklahoma City	29	24.96	25.12	105.83	105.84	105.96	106.67	106.65	106.66
Memphis	28	27.61	27.62	100.60	100.94	100.63	99.33	99.79	99.45
Portland	28	24.61	24.51	106.88	106.82	106.88	108.08	107.93	108.05
Denver	24	25.06	24.96	110.48	110.16	110.35	111.29	110.93	111.20
Sacramento	21	20.84	20.57	102.42	102.65	102.52	106.63	106.62	106.73
New Orleans	20	22.39	22.17	104.56	104.66	104.59	107.56	107.46	107.58
Dallas	19	21.94	21.75	98.42	98.93	98.55	101.65	101.98	101.75
Minnesota	18	23.99	23.85	104.37	104.47	104.35	106.00	106.02	106.02
L.A. Lakers	16	15.08	14.48	102.58	102.76	102.51	111.96	111.58	111.92
Phoenix	11	17.78	17.40	108.04	107.86	107.98	115.12	114.52	115.03

extreme teams. Such outlier teams are tough to account for in any model, and though unable to account for these phenomena here, we discuss potential future work in section 6.

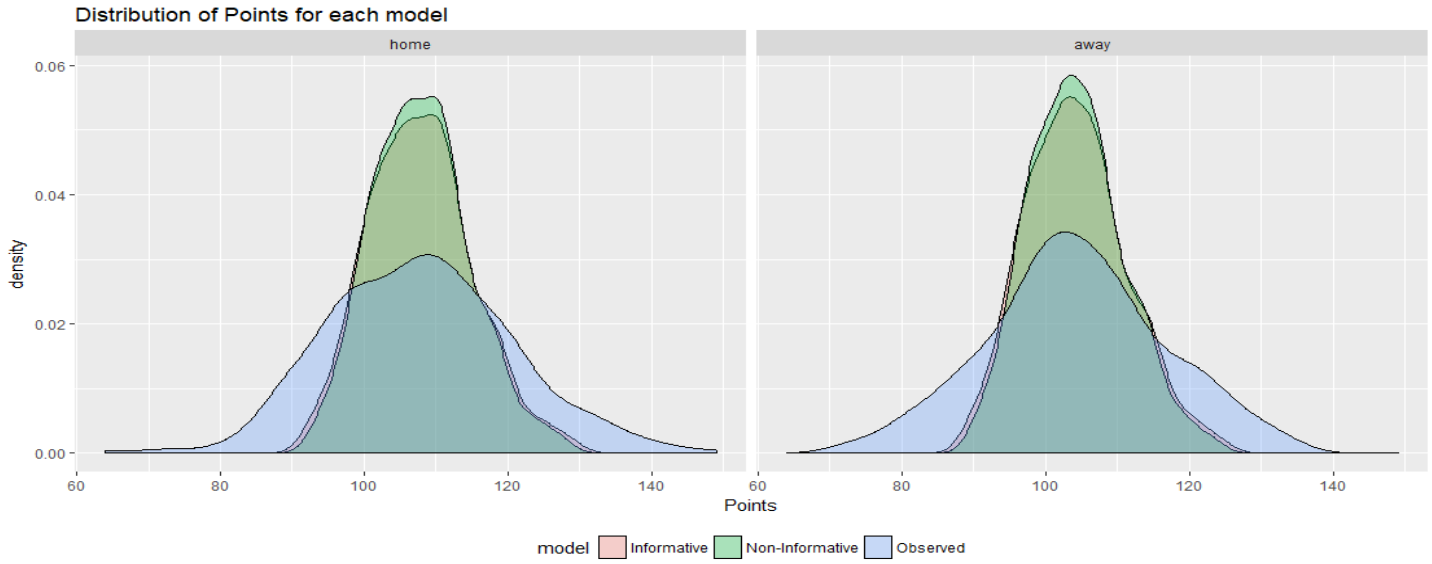

Figure 3: The observed point distribution for all teams when they were playing at home and away along with the predicted points scored by both the non-informative and informative models. Both models are concentrated around the mean.

5.2 Mixture Model Results

Here we discuss the results of our mixture-model and it's failure to yield significant improvement. First we discuss clustering of team summary statistics, then how these clusters relate to the hierarchical model results.

As seen in Table 2, the hierarchical model developed for basketball does not suffer from the same problem of over-shrinkage seen in [1]. In fact, our model does a fairly good job of predicting average points for and against teams. Thus, it is of little surprise that the mixture model implemented in Section 4 offers marginal improvement over the non-mixture models. We corroborate this result by analyzing summary statistics of the Western conference teams from the 2015-2016 season. First we calculate the average points scored for and against each team. Also mentioned in Section 4, we model these summary statistics as coming from a bivariate normal distribution. We then perform EM clustering on these teams to try and identify two clusters (or groups) of teams. Ideally, we will be able to identify two classes of teams - good and bad. The results however, indicate that our data is dominated mostly by one large, far-reaching cluster. The second cluster is motivated largely by outlier teams. This helps explain why clustering within the Hierarchical Model yields little improvement, as separation between classes is tenuous.

Figure 5 corroborates the results from clustering outside of the model. We see in the upper plot that the two clusters resulting from our model have little separation (similar to what we saw above, with one dominant cluster). As expected, little separation between classes leads to marginal difference in parameter estimation (lower plot). It's true that the bad teams are made slightly worse, and good teams are made slightly better, but these differences don't translate to an increase in prediction performance.

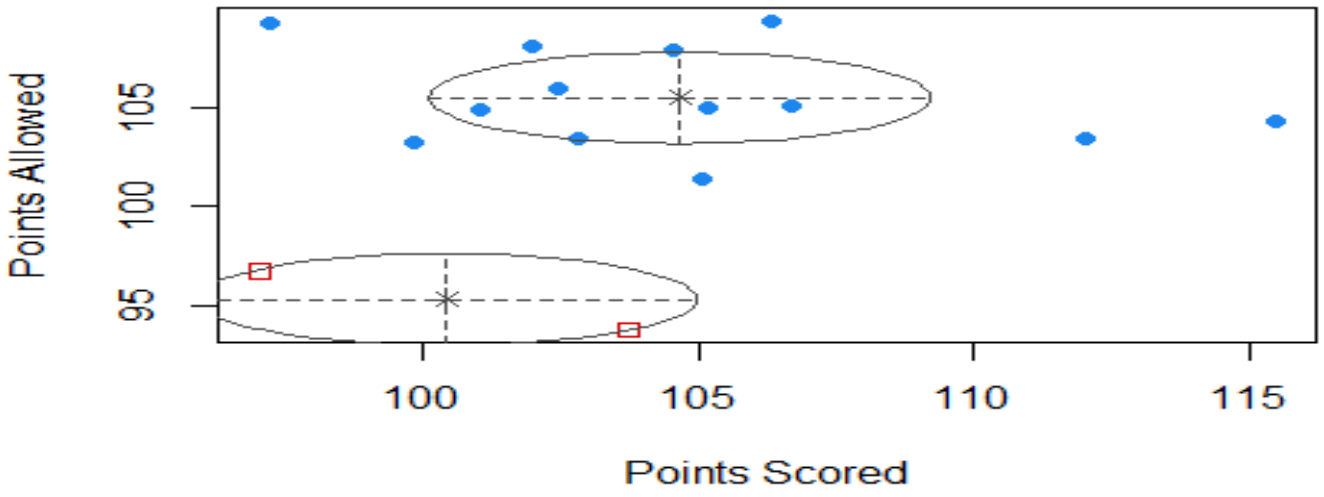

Figure 4: Results of clustering the team summary statistics with k=2 clusters.

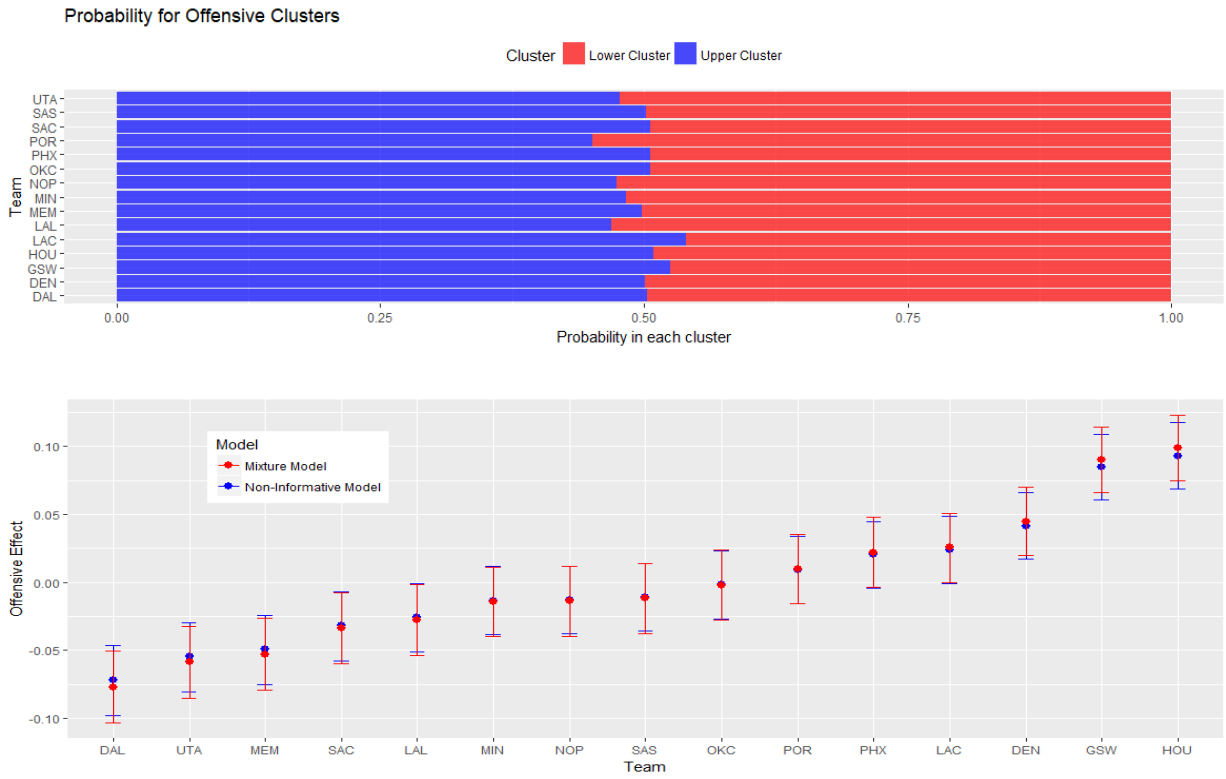

Figure 5: Mixture Model results for offensive effects and cluster probabilities. The top figure shows the probability each team is in the Lower and Upper cluster, however, all teams are close to being 50/50. The bottom figure shows the mean and 95% credible intervals for the offensive effect compared to the non-informative model. We see a similar pattern in defensive effects (withheld for space considerations).

6 Conclusion

6.1 From Soccer to Basketball

Throughout this paper, we follow the model proposed in [1]. Baio and Blangiardo, however, performed their analysis on soccer data. While soccer and basketball are inherently different sports — basketball is higher scoring, scores have very high variance, and the scores within a game are at least somewhat correlated — we were happy with the performance of our model, especially as a first attempt at modeling basketball data in this way. With that being said, these issues are things that need to be taken into account more rigorously in the future. The success of this model leads us to believe that a more thorough analysis would yield even more powerful results.

6.2 Potential difficulties with the data analysis

We just mentioned some major differences between soccer and basketball, and thus some major flaws with our model. In Section 2 we mentioned our concerns about using two separate seasons as we are aware that teams can have major changes from one season to the next. Additionally, we are making the assumption that each game is completely independent of the next. While this assumption is probably fine for our purposes, we do not know how true this is in reality. Teams play back-to-backs or take long road trips which begin to take effect on the performance of the team. Also, injuries could happen in a game that could directly effect the performance of subsequent games. Finally, unlike the

soccer data used in [1], not every NBA team has the same schedule. Thus, some teams may have inherently harder schedules than others which could change the estimates for the performance of the team.

Another potential difficulty is that through some preliminary data visualization, we noticed that the home score and away score are correlated with a strength of about 0.35. Thus, how many points a team will score in a game does not only depend on their ability to score and their opponent's ability to defend but also a correlation between the two.

6.3 Future Work

Due to time constrictions, there are other aspects of this analysis that we hope to complete in the future. The first being to better model the correlation between teams. Karlis and Ntzoufras suggest using a Bivariate Poisson Model to account for this correlation in scoring. It was also brought to our attention that modeling offensive and defensive parameters as bivariate normal could help account for the correlation between home and away scores.

Furthermore, in our results we mentioned how our model is under-predicting the wins for some teams while over-predicting wins for other. We hope to further research this phenomenon and correct for this. One way we have thought of doing it is by including further levels to our hierarchical model and including box score statistics/predictors.

Finally, we believe a "dynamic" model that could update as the season is going on would give better predictive results. By including the current state of the team (injuries, win streaks, etc.), we believe our model could better predict a single game.

References

[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.`

[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.

7 Appendix

In the submitted zip file with this final report are three folders: Data, Code and Models. In this section we will explain the files in each folder.

7.1 Data

The main data that was used for this report can be gathered by running the jupyter notebook file, `data_gather.ipynb`, located in the Code folder. Running this script will create four csv files in the Data folder. The two resulting files that were used to perform the analysis are called `stats551data_1516 _updated.csv` (which contains every game in the entire NBA season from the 2015-2016 season) and `stats551data _updated.csv` (which contains every game in the entire NBA season from the 2016-2017 season).

When running the script to obtain the data and other scripts throughout the project, a data file named `teamsdf.csv` is needed. We have included this file in our submission. The csv file contains the list of every NBA team with their team abbreviation and team index. We created the data file by hand and used it to join tables together.

Additionally, while predicting seasons, we simulated 50,000 seasons which also took a while to run. Because of this, we create more csv files along the way. These files are included in the submission.

7.2 Code

The first two files in the Code that should be run to recreate this analysis are `data_gather.ipynb` to obtain the data from the web and `west_teams.R` to subset the data into only the Western Conference teams. We discussed above how we needed to scale our data down which is why we created the script `west_teams.R`.

Each model that was ran has its own R script and STAN model with it. Each R script prepares the data, runs the STAN model, saves the model, reloads the model and then does different sort of analysis. The non-informative models were ran using the R script `initial_model.R` with the STAN file `init_model.stan`. This script was used to run both the non-informative models for the two seasons by changing the data that was read in. The informative model can be run using `inform_model.R` and `informative_model.stan`. Finally the two mixture models have corresponding R and STAN files that all start with `mix_model_*`.

To predict seasons we wrote a separate R file titled `predict_seasons.R`. In this R file there are functions to simulate a season and put the data into one of two forms that are needed for different reasons. The code compares the non-informative and informative models, including calculate the predicted wins, prediction accuracy and create plots/tables, can be found in `compare_models.R`.

Finally, the other clustering was performed in the R file `cluster_plot.R`.

7.3 Models

The Models folder contains the saved `.rds` (STAN models). Because some of the models took a long time to run, we found it easier to run the model once and save it so we could reload it and perform analysis on the models. There are 5 saved models submitted with this project, however, the original code to run and save the models is also included.

8 Breakdown of Work

In this section we discuss how the work was split up across the two group members. Overall, this project involved a lot of trial and error in creating models and the aspects of the project were mostly worked on together. The background research and writing of results were done in equal parts by both group members.

8.1 Andrew Soncrant

Worked heavily on the mixture model and clustering.

8.2 Taylor Spooner

Wrote the script to gather the data. Worked more on the informative/non-informative models.