NBA Player Clustering Using Gaussian Mixture Modeling

Noah Edwards-Thro

1/23/2022

Schedule

3/28 - Final Edits Part 1

4/4 - Final Edits Part 2

1. Introduction

This paper is for the purpose of completing an Honors Thesis in Mathematical Business at Wake Forest University. It was inspired by a paper presented at the 2020 MIT Sloan Sports Analytics Conference by Samuel Kalman and Jonathan Bosch called "NBA Lineup Analysis on Clustered Player Tendencies: A new approach to the positions of basketball & modeling lineup efficiency of soft lineup aggregates" (Kalman and Bosch 2020). Our work expands on their paper by analyzing how their clustering model performs on the most recent three years of NBA data and by investigating potential ways to achieve variable reduction and cluster reduction with similar results.

2. Methods

2.1 Reproduction vs. Replication

One important topic in statistics is the difference between reproduction and replication. Replication is the process of performing the same experimental setup with a different analyst. Essentially, given the same data and the analysis outline, another analyst could reproduce the same result. Reproduction, however, is the process of completing the analysis completely from scratch with a new sample, new experimental setup, etc. (Plesser 2018) An example of reproduction Kalman and Bosch (2020)'s analysis would be to use a new population (perhaps NCAA basketball data) and to arrive at the same nine clusters that they arrived at.

Our paper is an example of a replication of the analysis that Kalman and Bosch (2020) performed. To replicate their analysis, we endeavored to use as close to the same data and methods that Kalman and Bosch used as we could. The graphic below, created from the scifigure package (Patil, Peng, and Leek 2016), illustrates in what area our replication is different from their analysis and in what areas it is the same. Finally, it shows how our expansion is different from both the original analysis and our replication. In the figure, black icon mean that trait of the analysis is the same as the original while blue traits are different from the original analysis.

	Kalman and Bosch	Replication	Expansion
Population			
Question	?	?	?
Hypothesis			
Data	01100 10110 11110	01100 10110 11110	01100 10110 11110
Variables			
Analyst			
Code			
Results			
Claim			

2.2 Replication of Sloan Paper

2.2.1 Data

Kalman and Bosch manually scraped ten years (ranging from the 2009-2018 seasons) of player statistics covering advanced aggregation statistics, per possession statistics, and shot distribution statistics. Their data consisted of 5512 observations with 73 variables where each observation was a single season of a single player (if a player played all ten seasons during this range, he would show up ten times).

While we desired to stay as close to the process that Kalman and Bosch used as possible, we had knowledge of a package in R called nbastatr (Bresler, n.d.) that pulled data directly from Sports Reference LLC (the same website that Kalman and Bosch manually scraped their data from). Using the `bref_player_stats` function, we decided to pull the data from this package because of the speed and ease the package provided compared to manually scraping the data. We hypothesized that using this method would give identical (or near identical) data to the data that was manually scraped by Kalman and Bosch. Additionally, the shot distribution statistics were unable to download via the nbastatr package so we manually scraped them from Sports Reference LLC. Finally, we used the `player_profiles` function in the nbastatr package to download the heights (in inches) for all of the players before using the mutate-joins functions in the dplyr (Wickham et al. 2019)package to combine the data.

Our data consisted of 4760 observations and 136 variables. Notably, this is different than the 5512 observations seen in Kalman and Bosch's analysis (Kalman and Bosch 2020). Through some exploration, we found that the 752 discrepancy in observations has to do with players who are traded. In our data set, any player who is traded mid-season will still only show up as one row, with that players' team being "TOT" (signaling that the player played for multiple teams). Looking at the data on the Sports Reference LLC website, it seems that Kalman and Bosch likely had a separate row for each team that a player played on in a single

season (so if a player played games for three separate teams, he would have three separate rows for that season in Kalman and Bosch's data while he would only have 1 row in ours).

In Kalman and Bosch's analysis, they filtered the data so that only observations with more than 30 games in a season are counted, ending with 3,608 observations (Kalman and Bosch 2020). After using the same filter, we ended with 3676 observations, leading to a 68 discrepancy between our analyses. This difference is again likely due to the difference in data format with traded players (a traded player who plays 20 games for two different teams doesn't show up in their analysis but will show up as playing 40 games in ours).

Additionally, in our data validation, we found that a few observations transferred incorrectly, specifically as it relates to decimals, when using the nbastatR package. For most of the observations, a percentage would become a decimal, but in observations involving percentages under 1 percent, the number would transfer incorrectly. For instance, an offensive rebounding percentage of 12.4% would normally become .124 in my data set. However, in the 2009 season, J.J. Redick has an offensive rebounding percentage of 0.8%. Instead of coming into our data set as 0.008, the decimal in this observation (and others similar to it) did not move and came in as 0.800. We identified these errors and corrected them as part of the data validation process.

2.2.2 Variable Selection

We used the same variables that Kalman and Bosch used in their model (with the exception of using per 36 minute statistics instead of per 100 possession statistics for points and field goal attempts). Using per 36 minute statistics will lead to slight differences in scale versus using per 100 possession statistics but the important thing is that both are rate statistics.

The variables were a combination of offensive, defensive, and aggregate statistics and all statistics were calculated as rates except for player height. The drawback of these rate statistics is that they does not take into account how much a player is on the court. For instance, a player who plays ten minutes a night could have the same rate statistics as a player who plays 35 minutes a night, but we would consider these players very different in their abilities (largely due to the 35 minute player being able to play efficiently for 35 minutes). Below is a table displaying all of the variables included as well as a short description about the variable.

Table 1: Recreation of Table 1 from Kalman and Bosch (Kalman and Bosch 2020)

Variable	Description
Height	Player height, in inches
Offensive Rebound Rate	% of available offensive rebounds a player gets while on the floor
Defensive Rebound Rate	% of available defensive rebounds a player gets while on the floor
Assist Rate	% of teammate field goals that a player assisted while on the floor
Steal Rate	% of opponent possessions that end with a steal b the player while on the floor
Block Rate	% of opponent field goal attempted blocked by the player while on the floor
Turnover Rate	Turnovers committed per 100 offensive possessions
Points	Points scored per 100 offensive possessions
Usage Rate	% of offensive team possessions used by the player while on the floor
Player Efficiency Rating	Per-minute production standardized such that the league average is 15
Free Throw Rate	Number of free throws made per field goals attempted
Free Throw Percentage	Number of free throws made per free throw attempt
Field Goals Attempted	Number of field goals attempted per 100 possessions
2FG%	Number of two-point field goals made per attempt
3FG%	Number of three-point field goals made per attempt
2FG Assist Rate	% of two-point field goals that are assisted
3FGA%	% of field goal attempts that are three-point attempts
Corner 3FGA%	% of three point-field goal attempts that are from the corner
3FG Assist Rate	% of three-point field goals that are assisted

Variable	Description
Dunk Attempt Rate	% of all field goal attempts that are dunks
0-3 ft FGA%	% of all field goal attempts between zero and three feet from the basket
3-10 ft FGA%	% of all field goal attempts between three and ten feet from the basket
10ft-3p FGA%	% of all field goal attempts between ten feet from the basket and the three-point line

2.2.3 Gaussian Mixture Clustering

After attempting K-means clustering and being unsatisfied with the results, Kalman and Bosch pivoted to model-based clustering to cluster the players in their data set. Model based clustering, specifically finite Gaussian mixture modeling, uses an expectation-maximization (EM) algorithm to fit observations into clusters. Each distribution (cluster) has its own mean and variance (Dobilas 2021). This allows for clusters to not just take a circular shape (in two dimensions) but rather take an oblong shape. An advantage of model-based clustering is that it assigns "soft clusters," showing the probability that each observation will be in each cluster (Maklin 2019). The figure below (Carrasco 2019) displays a two-dimensional graphical representation of the clustering distributions produced by Gaussian mixture modeling. As you can see, each cluster has not only a different mean, but a different variance.
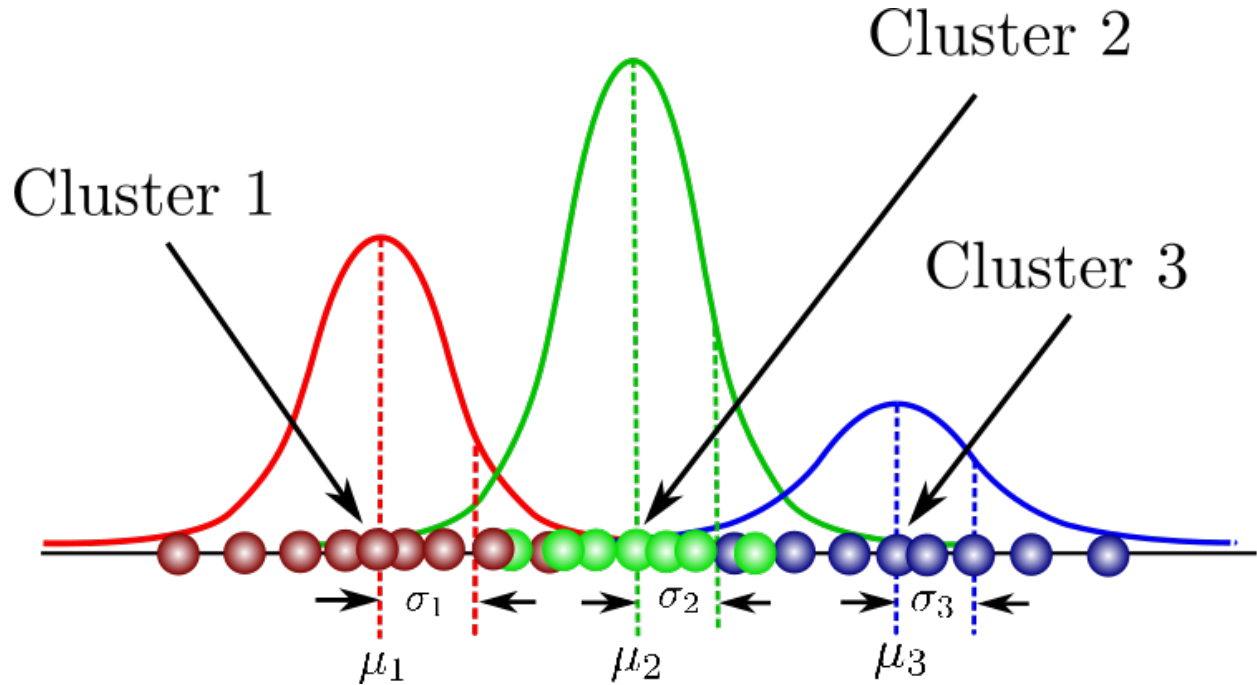
As done by Kalman and Bosch, we used the `mclust` (Fraley and Raftery 2006) package in R to implement the Gaussian mixture clustering.

2.3 Expansion of Kalman and Bosch's Paper

2.3.1 Data Expansion

While Kalman and Bosch's original paper only used data from the 2009-2018 seasons, we wanted to see what predictions the model would make on more recent seasons (2019-2021 seasons) given the great degree to which the NBA has changed over the past half decade alone. As before, we used the `nbastatr` package

to download the majority of the data and we manually scraped the shooting statistics necessary from Sports Reference LLC (LLC, n.d.).

Our three new seasons of data consisted of 1600 observations and 136 variables. As in the original ten seasons, our new three seasons of data only had one row per season for players traded in the middle of the season. To keep in line with the original analysis, we filtered the data to players with greater than 30 games, resulting in 1125 observations remaining. Additionally, we used the same data validation techniques that we used on the first ten seasons on the new three seasons.

A secondary reason that we wanted to add the three new seasons is we wanted to investigate how similarly or differently a new model with all 13 seasons data might cluster the players. For each of the three number of clusters and number of variable pairings that we tried (more on that later), we made a pair of models (one based on the original ten seasons of data and one based on all 13 seasons of data). One of the reasons that we made a pair of models for each one is it is difficult to test the accuracy of a Gaussian Mixture model. There is not necessarily a "correct" way to cluster the observations and in many ways, having a good model is based on expert knowledge of the data and being satisfied with the results. In an effort to come up with some test metric, we decided to test how the clustering predictions hold for each pair from one model to the other (i.e. does a player classify in the same cluster in both models). Below is a table displaying the six models (three pairs) that we created.

Pair	Original Ten Seasons	All 13 Seasons
1	Original Variables, 9 Clusters	Original Variables, 9 Clusters
2	Reduced Variables, 9 Clusters	Reduced Variables, 9 Clusters
3	Reduced Variables, 6 Clusters	Reduced Variables, 6 Clusters

2.3.2 Variable Reduction

The first component that we changed was the number of variables that were input into the model. We believed that for the most part, steal and block rates were mostly random and not necessarily indicative of position (or had a strong correlation with other variables - e.g. height and block rates both being a strong predictor for traditional center). We also removed the PER variable as we believed that most of the variability explained by PER could be explained by other variables. This belief was confirmed when we ran a redundancy analysis and found that 83.5% of the variability in PER could be predicted from the offensive and defensive rebounding percentages, assist percentage, turnover percentage, and points per minute variables. Next, we cut some of the variables related to percentage of shots being assisted or coming from very specific locations (% of 3s from the Corner). Finally, we cut free throw percentage and we combined the percentage of shots taken from 0-3 FT and 3-10 FT into a single variable.

2.3.3 Cluster Reduction

Kalman and Bosch never outlined in their paper why they chose 9 clusters as the optimal number, but we hypothesized that a slight reduction in clusters (to 6) would make more intuitive sense. We believed that some clusters would be better grouped together. For instance, we hypothesized that most of the players in the Stretch Forward and Three Point Shooting Guard clusters could be combined into one Three Point Shooter cluster. We also hypothesized that most of the players in the Traditional Center, Skilled Forward, and Mid Range Big clusters could be grouped into a Traditional Center (some of the players in Mid Range Big would shift to Traditional Center) cluster and a Skilled Big cluster. Finally, we hypothesized that most of the players in the Ball Dominant Scorer, High Usage Guard, and Floor General clusters could be better categorized into a Ball Dominant Scorer (some High Usage Guards would end here) cluster and a Complimentary Guard cluster. For this model, we used the reduced variable set as well.

3. Results

To analyze our clustering methods, we looked at a few different features. To start, we looked at the proportion of players in each cluster based on the original ten seasons and the new three seasons. We wanted to see how this proportion matched with that of Kalman and Bosch's analysis. Following this, we took a deep dive into each cluster to make sure that our cluster characteristics matched that of Kalman and Bosch. While Kalman and Bosch showed the Three Point Shooting Cluster in their analysis, we opted to show the Traditional Center cluster to give readers a view of a different cluster. Following this, we investigated if the the clustering method could accurately capture the evolution of individual teams and the league as a whole. Next we analyzed cluster consistency between our model with only ten seasons and our model with all 13 seasons.

Next, in our expansion of Kalman and Bosch's analysis, we looked at how the proportion of players in each cluster in the original model compared with the proportion of players in each cluster in the model with reduced variables. To go along with this, we investigated the percentage of players in each cluster from the original model that stayed in that cluster for the reduced variable model, another measure of cluster consistency. Finally, we reduced the number of clusters and investigated trends in how players in the original nine cluster model were clustered in the new six cluster model.

3.1 Reproduction of Kalman and Bosch's Model

3.1.1 Cluster Breakdowns

In their analysis, Kalman and Bosch found 9 clusters and assigned them labels based on player types within those clusters. Our initial reproduction of their model had clusters that matched theirs, though the proportions for each cluster were different than their proportions by an average of 3.37% per cluster (the average cluster had 11.1%). Not unexpectedly, we found that the difference in proportion of clusters in the new three seasons was noticeably greater than that of the original 10 seasons, suggesting the type of players in the NBA has evolved significantly. Below is a graph showing how the proportion of our clusters compared to theirs.

Figure 3: Distribution of Players Across 9 Clusters
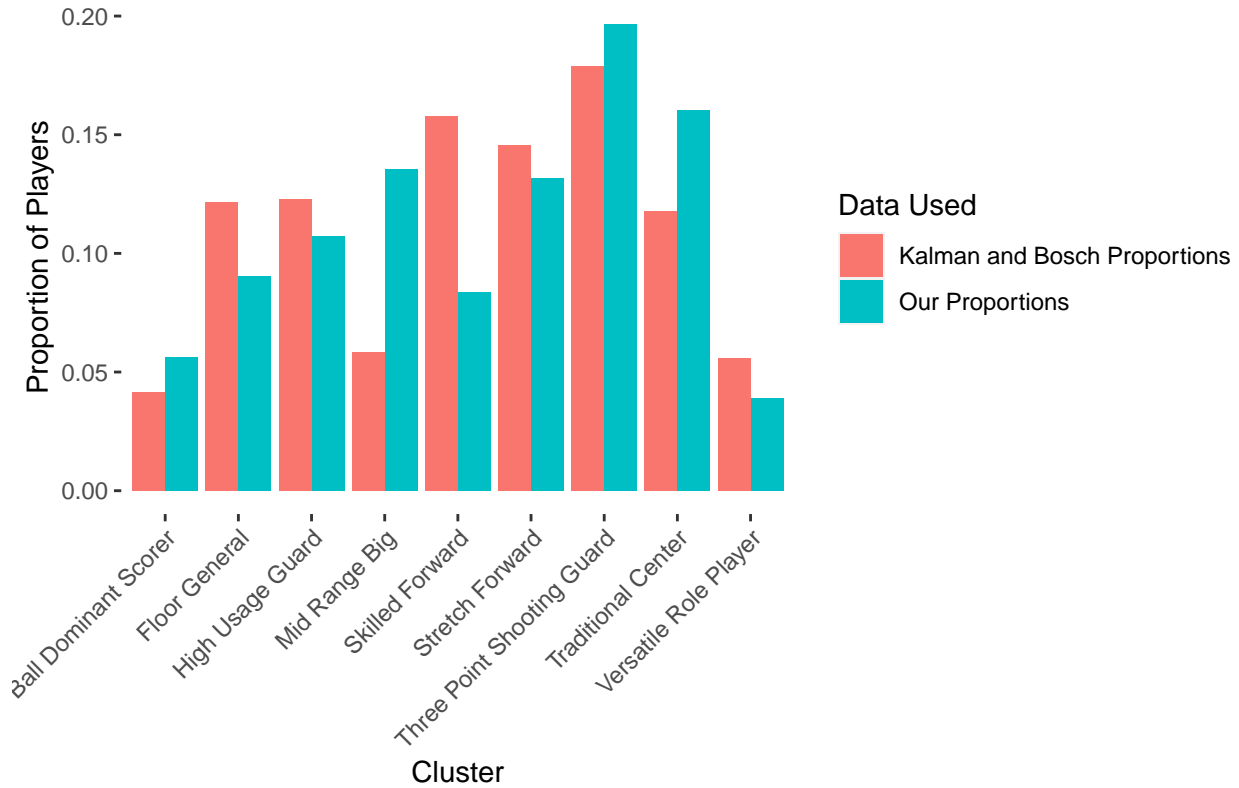
In our analysis, Three Point Shooting Guard was the highest cluster, and while it would have been the second highest cluster with just the seasons that Kalman and Bosch used, the three new seasons that we added had an over 5% higher proportion of players clustered into the Three Point Shooting Guard cluster. Meanwhile, the Traditional Center cluster had the highest proportion of players in the original ten seasons but then saw its proportion of players plummet with the new three seasons worth of data. These trends sparked an interest in us to look at overall league trends in cluster proportions, something we will do in a later section.

Figure 3: Distribution of Players Across 9 Clusters
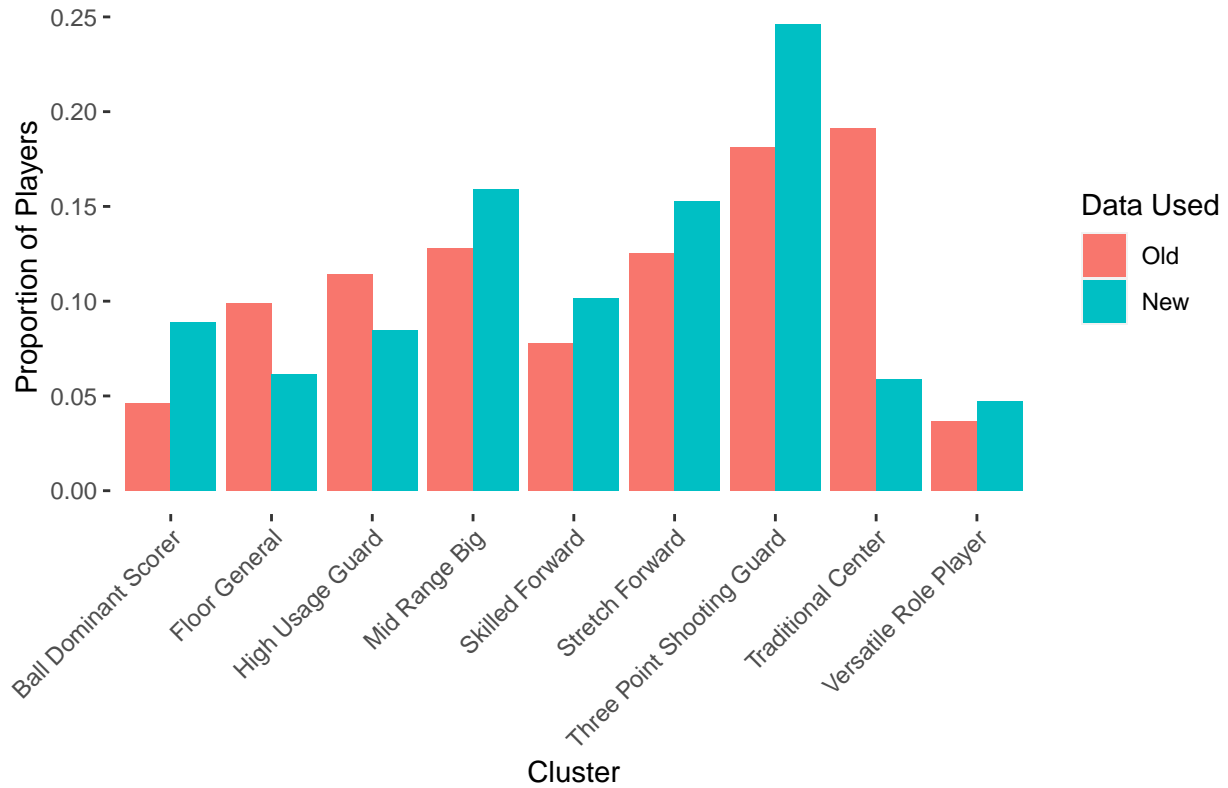
3.1.2 Traditional Center Cluster Analysis

While it had the highest proportion of players in the original ten seasons, the proportion of players in the Traditional Center cluster plummeted in the new three seasons, making it a particularly interesting cluster. Below is a graphical representation of the Traditional Center cluster using box plots. The Traditional Center was very high in Offensive Rebounding, Block Rate, and Dunk Rate. Additionally, the Traditional Center was low in Assist Rate, Free Throw Percentage, Percentage of Shots Taken from 3.

Traditional Center Cluster
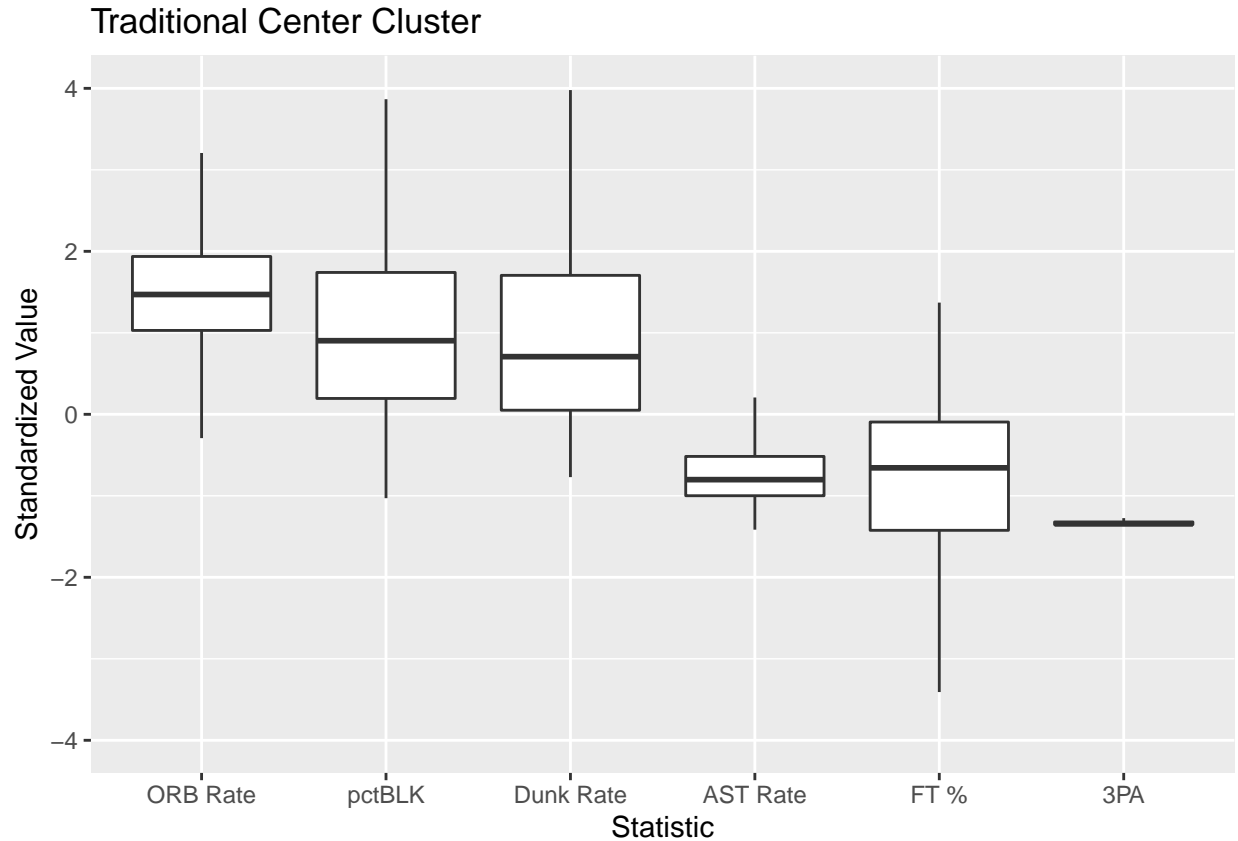
3.1.3 Golden State Warriors

We wanted to investigate how the clustering told the story of individual teams' evolution over the course of the 13 year period. The figure below displays the a heat map of how the Warriors have gone about team building the past 13 years. We picked the Golden State Warriors to show because they historically have been averse to making mid-season trades, meaning that all of their players will show up as playing for their team and will not have a "TOT" team name in our data set. The most noticeable part of this figure is how the Warriors moved from Traditional Centers to Mid Range Bigs. While they are most recently known for their "Death Lineup" and three point shooting, it is not that the Warriors don't have big men, but rather that they have transitioned from Traditional Centers to Mid Range Bigs. The Andrew Bogut/Festus Ezeli pairing that marked the early 2010's has been replaced by the more mid range oriented Kevon Looney/Eric Pascall pairing. The three year period of Kevin Durant can also be seen by an increase in the Ball Dominant Scorer cluster. Finally, it is interesting to note that the Warriors have more Versatile Role Players than most teams with players such as Draymond Green and Shaun Livingston often being classified as Versatile Role Players.

Golden State Warriors Cluster Heat Map
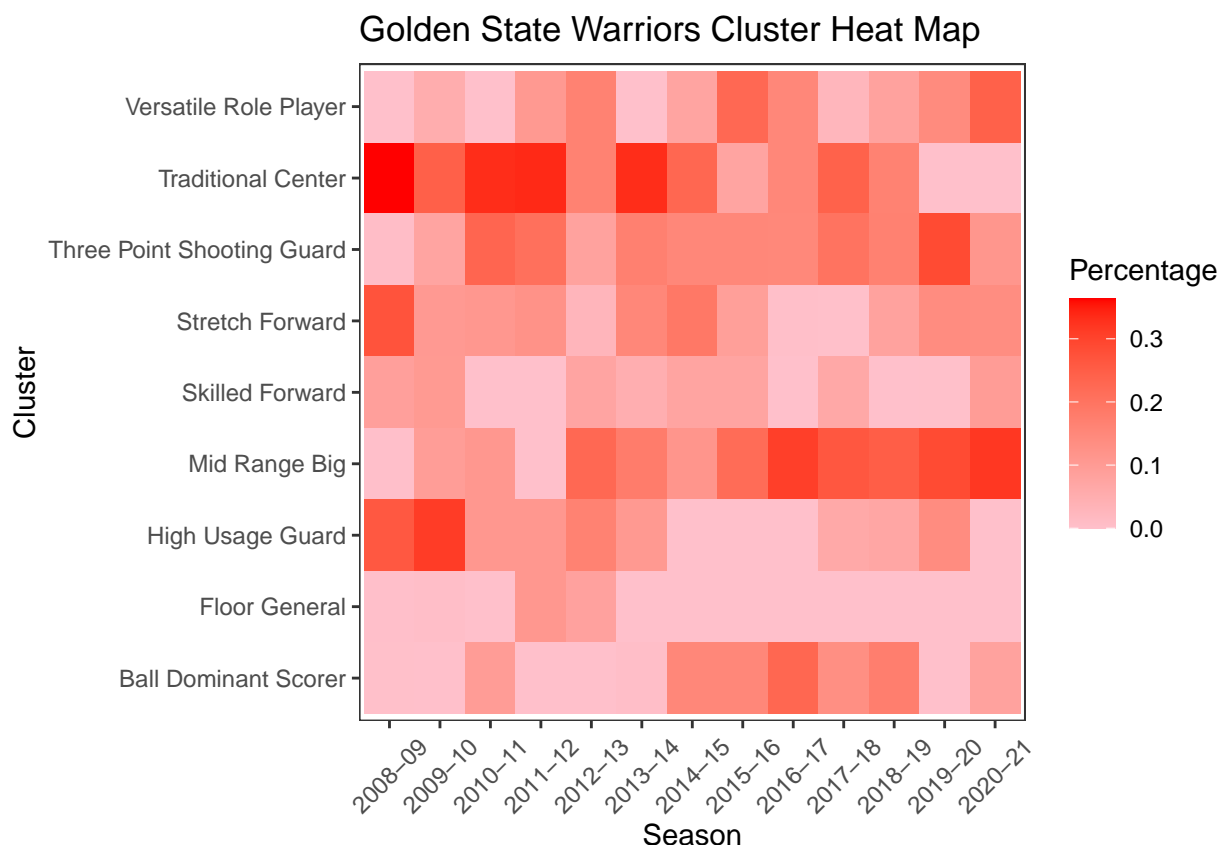
3.1.4 League Trends

One of the things that was not investigated by Kalman and Bosch (at least in their paper), was the league wide trend of shooting more threes and how that shows up in the clustering algorithm. We sought to investigate this and other potential trends could be seen through the use of a heat map of the clustering results across the 13 seasons worth of data. This heat map uses the probabilities that a player was in each cluster so it accounts for the 25% that a player is in one cluster even though they are primarily classified in a different cluster. As seen in the heat map below, the Traditional Center cluster had the highest proportion at the beginning of this time frame but gradually decreased as the NBA moved towards a more perimeter oriented game. At the same time, the Three Point Shooting Guard cluster gradually increases before exploding over the last five years where now over 25% of the players in the most recent completed season classify as a Three Point Shooting Guard. The transition away from a game dominated in the paint can also be demonstrated by slight increases in the proportions of the Stretch Forward, Mid Range Big, and Skilled Forward clusters. Finally, we thought it was particularly interesting that in the last few years, the Ball Dominant Scorer cluster has increased with slight declines in the Floor General and High Usage Guard clusters.

Proportion of Cluster Across the NBA Over 13 Seasons
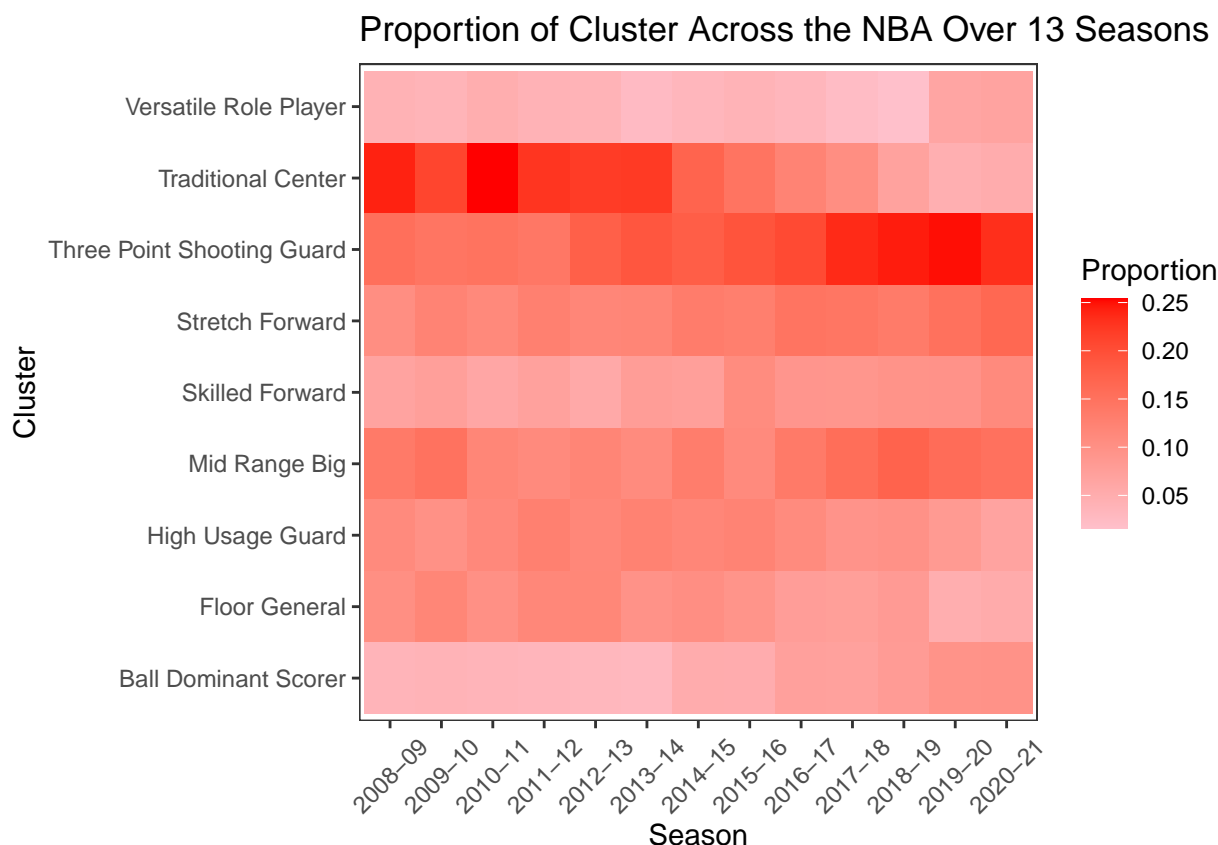
3.1.4 Cluster Consistency

As we mentioned in our methods section, one of the challenging things about clustering players is that there is not necessarily a way to tell if the clustering model is effective or not. We chose to go about this problem by using the same clustering method over all 13 seasons (instead of just the original ten) and seeing how many players remained in the same cluster. When we did this we found that 62.95 percent of the players remained in the same cluster.

3.2 Expansion of Kalman and Bosch's Analysis

3.2.1 Variable Reduction

While working with the data from the Kalman and Bosch analysis, we hypothesized that the data could achieve similar clustering outcomes using a smaller number of variables. We hypothesized this because we believed that some variables were either marginally relative to the clustering process (such as those relating to defense), were strongly correlated with other variables, or made intuitive sense to combine (such as some of the shot distance data).

Below is a figure displaying the proportion of players in each cluster for the first model using all the original variables and the second model, using the reduced variables. At first glance, it appears as if the results of the model match up decently well. There seem to be a few clusters that have a similar proportion of players for in both models, specifically the Traditional Center, Skilled Forward, Versatile Role Player, and Floor General clusters. However, a closer inspection reveals that many of the players did not actually remain in the same clusters that they were assigned to in the original model.

Original Variables vs Reduced Variables Model Comparison
Model Based off of Data from 2009–2018
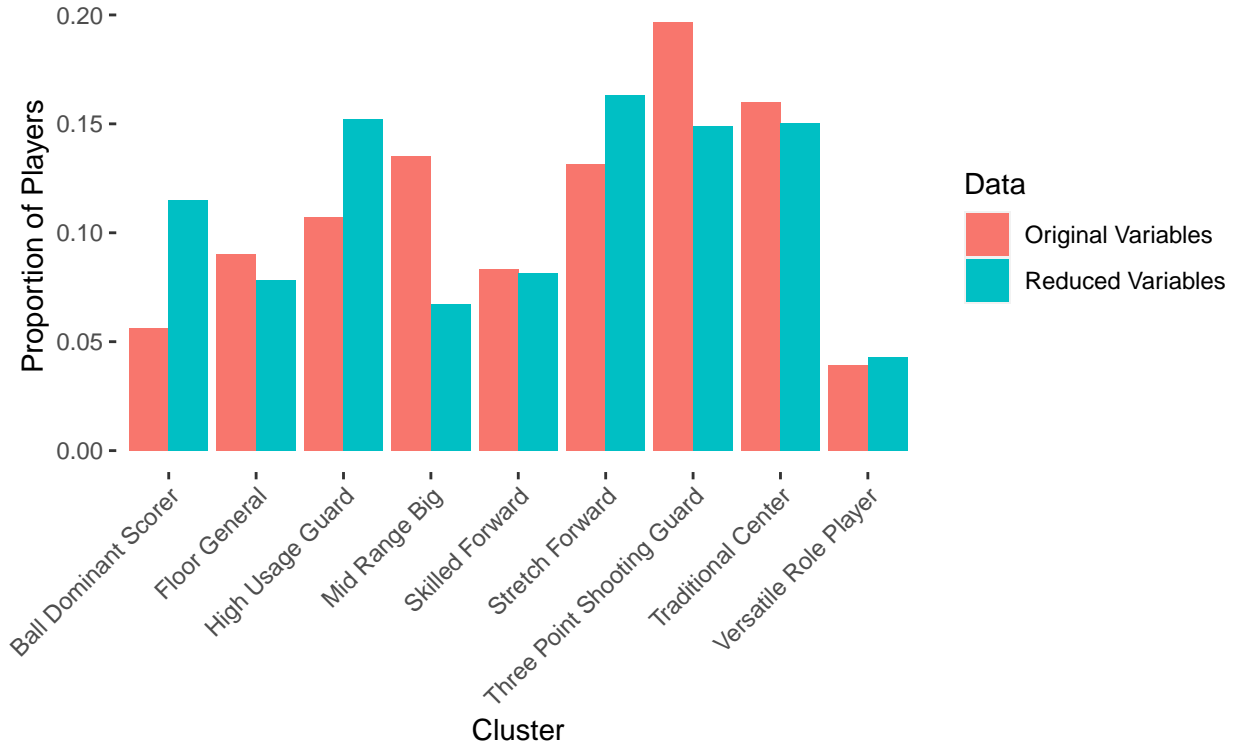
The table below shows the percentage of players for each cluster in the original model that remained in their cluster in the model with the reduced variables. There was a wide range of percentage of players retained as the Traditional Center cluster retained over 92% of its players while the Mid Range Big cluster retained just over 20%. Overall about 54.7% of players retained their original cluster, a number lower than we was expecting to see. One thing that this table reveals is that there are certain clusters (such as the Traditional Center and Ball Dominant Scorer clusters) that have a very strong identity. Once a player is put into one of these clusters, they are not likely to flip to another cluster in a different model.

Cluster	Percentage
Ball Dominant Scorer	74.81
Floor General	31.41
High Usage Guard	60.31
Mid Range Big	20.46
Skilled Forward	24.69
Stretch Forward	68.04
Three Point Shooting Guard	52.22
Traditional Center	92.59
Versatile Role Player	58.51

Finally, we wanted to investigate how the cluster consistency in this pair of models (the model with 10 seasons worth of data and the model with 13 seasons worth of data) with reduced variables compared to the cluster consistency of the original pair of model. When we did this we ound that 54.01 percent of the players remained in the same cluster. This is considerably lower than the 62.95 percent of the players remained in the same cluster in the original pair of models. We hypothesize that this lower cluster consistency is due to the variable reduction and that the smaller number of variables is unable to capture all of the nuances that

keep a cluster consistent from one model to another.

3.2.2 Cluster Reduction

Following the variable reduction, we wanted to see if the number of clusters could be intuitively reduced. This six cluster model created clusters of Skilled Big, Complimentary Guard, Three Point Shooter, Ball Dominant Scorer, Traditional Center, and Versatile Role Player. The table below outlines each of the clusters and the strengths and weaknesses that characterize each cluster.

Cluster	Description	High Stats	Low Stats
Skilled Big	A big who does plays both inside and outside of the paint. Has a higher usage and scores more than the Traditional Center. Has a very high 2 point Field Goal %.	Def Reb. % 2P FG %	3P FGA
Co mplimentary Guard	Secondary creator to the High Usage Scorer. Has the highest assist rate among the clusters and a strong 3 point shooting percentage, though this cluster doesn't take many 3s. Weak in height and defensive rebounding.	Assist Rate 3P FG%	Off. Reb % Height
Three Point Shooter	A guard or wing who takes a lot of 3's and is highly efficient from 3. Has a very low FT rate and does not crash the offensive glass.	3P FG% 3P FGA	Off. Reb % FT Rate
Ball Dominant Scorer	A high usage player, typically a guard or wing, who is proficient at scoring and creating for others. Does not crash the offensive glass and shoots a smaller proportion of their shots from 3 as compared to other players.	FGA (10 FT-3P) Point Per Minute	% of FGA from 3

Cluster	Description	High Stats	Low Stats
Traditional Center	Interior post center who shoots almost exclusively inside ten feet. High offensive rebounding rate and FT rate. Does not have a very high assist rate.	Off. Reb % FT Rate	3p FG% Assist Rate
Versatile Role Player	Group of players who act as chameleons on the court and can do whatever is needed by their team. Often serve as connectors on the team, though they have lower assist rates than the C o mplimentary Guard.	Assist Rate Usage Rate	Off. Reb % FGA (0-10 FT)

The figure below shows how players from the original nine clusters fit into the six cluster model. The Ball Dominant Scorer cluster retained almost all of the Ball Dominant Scorers from the original model and added some players previously clustered as High Usage guards, Skilled Forwards, and Three Point Shooting Guards. It might appear that the Ball Dominant Scorer cluster has become to broad and perhaps it has, but it also captures some elite scorers that were previously not in this cluster, such as Klay Thompson and CJ McCollum. In the original model, Thompson was classified as a Three Point Shooting Guard in the majority of his seasons and McCollum was classified as a High Usage Guard. One can debate whether or not we would classify these players as High Usage Scorers based on expert knowledge but it is not surprising that they are classified as such with a reduction in clusters. The Complimentary Guard cluster is mostly a collection of Floor Generals, Three Point Shooting Guards, and High Usage Guards, about what we expected it would be. The Skilled Big cluster too is largely what we thought it would be - a mix of Mid Range Bigs, Skilled Forwards, and a few Traditional Centers. The Three Point Shooter cluster is about what we predicted it would be, though more Mid Range Bigs ended up being classified in this cluster than we hypothesized. This suggests that a better name for this cluster might be "Floor Spacer" rather than "Three Point Shooter." The Traditional Center cluster is made up almost entirely of Traditional Centers from the original model, again demonstrating a strong cluster identity for Traditional Centers. Finally, the Versatile Role Player cluster retained the majority of the Versatile Role Players from the original model and added at least a few of players from almost every other cluster.

9 Clusters vs. 6 Clusters Model Comparison
Models Based off of Data from 2009–2018
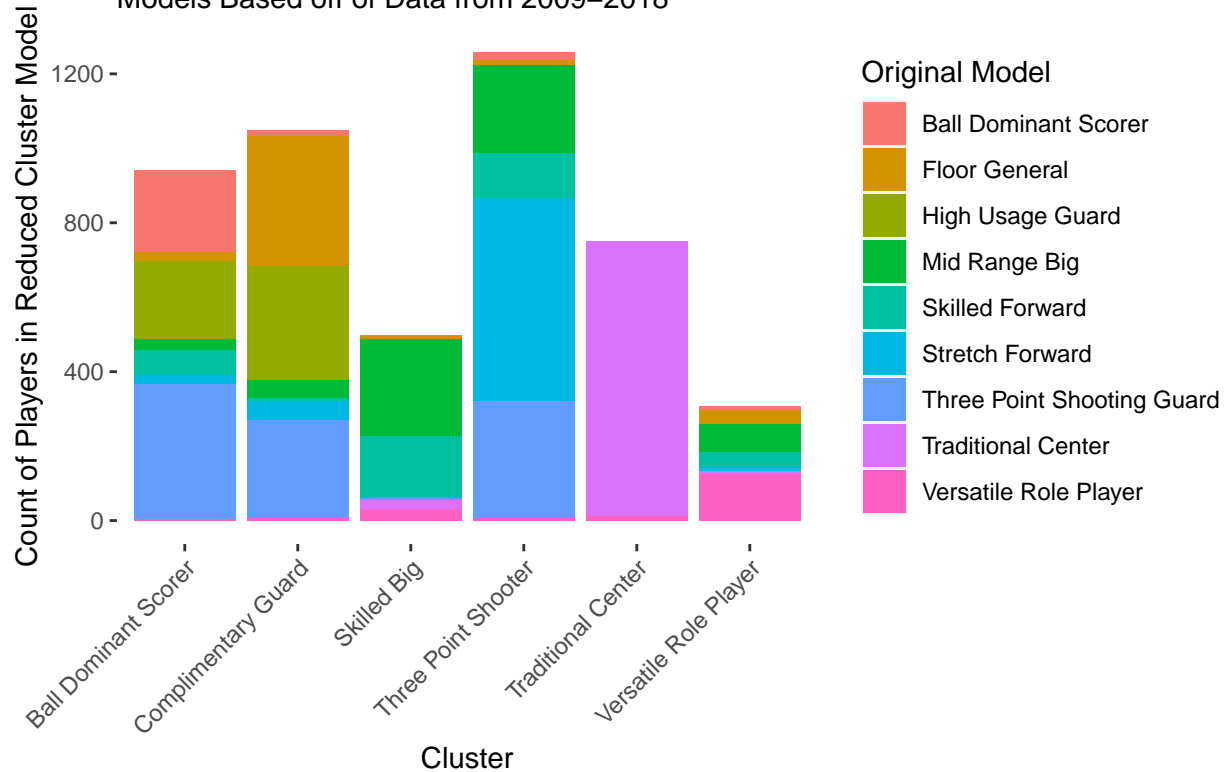
As we did with the other model pairs, we wanted to investigate how the cluster consistency held up with a model based on the original 10 seasons and another model based on all 13 seasons. When we did this for the 6 cluster models with reduced variables, we found that 54.2 percent of the players remained in the same cluster. This is remarkably similar to the 54.01 percent of the players that remained in the same cluster in the reduced variable model with nine clusters. We believe that this further demonstrates that the reduction in variables is the main driver in why these two pairs of models have a significantly lower cluster consistency than the original model and its pair.

4. Discussion

4.1 Advantages of Mixture Modeling

One advantage of mixture modeling, specifically finite Gaussian mixture modeling, is that it assigns "soft" clusters instead of "hard" clusters, giving each observation a probability for each cluster. This allows the observer to see some of the nuances of a player instead of painting him as entirely as a single cluster. For instance, a player might be primarily a Three Point Shooter but perhaps in some lineups, the player shifts to a more Ball Dominant Scorer type. Mixture modeling allows us to capture this by assigning some probability to both clusters. This shows up particularly well with some of the heat maps for a particular team or the league as a whole because it allows us to capture the 25% or 40% of a player that is in one cluster even though they are primarily listed in another cluster.

4.2 Importance of Input Data

One of the most important things that we saw through our analysis was how the input data impacted the model. For each of the three sets of models, we created one model based on the original ten seasons of data that Kalman and Bosch used and a second model based on 13 seasons worth of data. We believe that one of the reasons that the cluster consistency was not very high for any of the three pairs of models is that the new three seasons worth of data are significantly different than the previous ten seasons. The NBA has changed a lot in the past ten years as it moves away from a game dominated in the paint to a game dominated on the perimeter. That change has been gradual throughout the past ten years but has been accelerated significantly over the past five years. While the data over the most three recent seasons is still NBA data, the data itself looks very different (specifically as it relates to shot location and shot percentage data) than it did from the previous ten seasons. This change in data likely forces the model that has all 13 seasons of data to cluster things differently than the model that only has the original ten seasons worth of data, leading to a low cluster consistency than we expected seen throughout all of the pairs of models.

4.3 Strength of Cluster Identity

As demonstrated earlier, a few of the clusters (such as the Traditional Center cluster) had a strong cluster identity in that they appeared with similar characteristics no matter what data, variables, or number of clusters was used. This idea of cluster identity is interesting, especially as positions in the NBA become more fluid. Many people believe that the NBA is becoming a "positionless game" where it does not really matter what position you classify a player as. However, these strong cluster identities suggest that there are in fact indicators that separate players into specific positions, even if they are not the traditionally five used position that have been used historically. We do not want to corner players only into a single cluster (the whole point of "soft" clusters with Gaussian mixture modeling), but it is important to identify that there are real differences in skill sets between positions.

4.4 NBA Becoming More Heliocentric

One result that we found that we thought was interesting was the slight growth in the Ball Dominant Scorer cluster in the past 13 seasons. Looking at the heat map in Section 3.1.4, we can see small increases in the proportion of players who are classified as Ball Dominant Scorers. It appears this trend has grown even more rapidly over the past five years. While the NBA has always been a star dominated league, the development of offenses around superstar players such as James Harden, Luka Doncic, and Trae Young has changed the league. These starcentric offenses developed in an effort to be more efficient and we believe that over time, this trend will only continue to grow. The NBA has taken some measures to reduce the efficiency of this style of play, such as changing some of the rules relating to fouling that Harden and Young have so famously taken advantage of. Despite the NBA's effort, we believe that the league as a whole (especially as it relates to offense) will continue to gradually grow more starcentric. Further research should be done in the future to investigate this trend and potential ceilings on this trend.

5. Conclusion

We concur with Kalman and Bosch that NBA players can be clustered according to stylistic tendencies and efficiency. NBA players fit into more than just the traditional five positions. While we hypothesized that NBA players were better clustered into 6 clusters, we found that they are better clustered into the original nine clusters that Kalman and Bosch used. However, as the NBA continues to evolve, we believe that a reduction in clusters would be optimal.

Further work on NBA player clustering should be done using particular player skills and not just rate and aggregate statistics. Sports Info Solutions has started doing some of this work analyzing the value of

particular skills such as shooting off of pin down screens, creating advantages, and filling defensive space. The more granular that we can break players down, the more accurate and informative the clustering will be.

Using these new clusters, the casual NBA fan can better understand the game and how players compare to each other. NBA front offices can use these clustering techniques to better compare players and identify which players are over or undervalued. We hope that our paper furthers the work begun by Kalman and Bosch and that others will expand on our work in the future as more granular data becomes available.

References

Bresler, Alex. n.d. *nbastatR : R's Interface to NBA Data.* http://asbcllc.com/nbastatR/index.html.

Carrasco, Oscar Contreras. 2019. "Gaussian Mixture Models Explained." https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95.

Dobilas, Saul. 2021. "GMM: Gaussian Mixture Models - How to Successfully Use It to Cluster Your Data?" May. https://towardsdatascience.com/gmm-gaussian-mixture-models-how-to-successfully-use-it-to-cluster-your-data-891dc8ac058f.

Fraley, Chris, and Adrian E. Raftery. 2006. "MCLUST Version 3: An r Package for Normal Mixture Modeling and Model-Based Clustering." https://doi.org/10.21236/ada456562.

Kalman, Samuel, and Jonathan Bosch. 2020. "NBA Lineup Analysis on Clustered Player Tendencies: A New Approach to the Positions of Basketball & Modeling Lineup Efficiency of Soft Lineup Aggregates." In.

LLC, Sports Reference. n.d. *NBA Player Stats: Shooting.* https://www.basketball-reference.com/leagues/NBA_2020_shooting.html.

Maklin, Cory. 2019. "Gaussian Mixture Models Clustering Algorithm Explained." https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e.

Patil, Prasad, Roger D. Peng, and Jeffrey Leek. 2016. "A Statistical Definition for Reproducibility and Replicability." *bioRxiv.* https://doi.org/10.1101/066803.

Plesser, Hans E. 2018. "Reproducibility Vs. Replicability: A Brief History of a Confused Terminology." *Frontiers in Neuroinformatics* 11 (January): 76. https://doi.org/10.3389/fninf.2017.00076.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.