The Optimal NBA Roster-Building Strategy

Patrick Simpson Dr. Max Buot

Xavier University

April 29, 2020

Outline

1. What is Cluster Analysis?

2. Applying the Agglomerative Clustering Algorithm

3. Applying Hierarchical Clustering to NBA Data

Background

- Clustering is in our everyday life
- Cluster analysis methodology
- Early uses

Hierarchical Clustering

- Two different approaches:
    - Agglomerative
    - Divisive

Agglomerative Hierarchical Clustering Algorithm

Algorithm Agglomerative Clustering Algorithm

```
1: Starting with a dataset of $N$ observations, each begins in its own cluster.
   Form an $N \times N$ proximity matrix.
2: Identify the two most similar clusters.
3: Merge those clusters and recompute the proximity matrix (Note: it will
   have one fewer row and column than the previous matrix).
4: Repeat steps (2) and (3) until the proximity matrix is $1x1$.
```

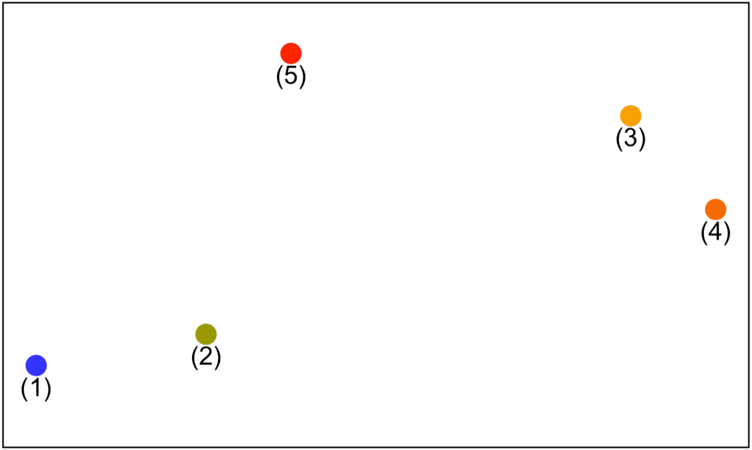
Outline

1. What is Cluster Analysis?

2. Applying the Agglomerative Clustering Algorithm

3. Applying Hierarchical Clustering to NBA Data

Sample Data



Step I

Algorithm Agglomerative Clustering Algorithm

```
1: **Starting with a dataset of $N$ observations, each begins in its own cluster. Form an $N$x$N$ proximity matrix.**
2: Identify the two most similar clusters.
3: Merge those clusters and recompute the proximity matrix (Note: it will have one fewer row and column than the previous matrix).
4: Repeat steps (2) and (3) until the proximity matrix is $1x1$.
```

Creating the Proximity Matrix

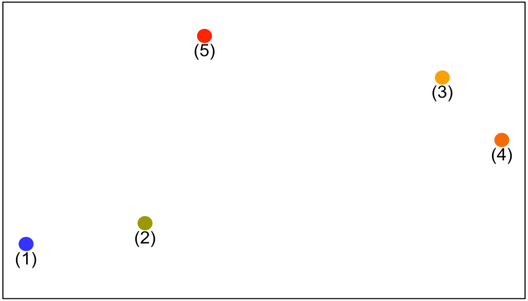
There are different measures of distance between two objects; in this model Euclidean Distance is used.

Definition

Let x and y be two points. The Euclidean Distance function between x and y can be expressed as: $d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

Proximity Matrix

	1	2	3	4	5
1	$\underline{\mathbf{0}}$	2.24	10.63	9.43	10.44
2	2.24	$\underline{\mathbf{0}}$	8.60	7.21	9.06
3	10.63	8.60	$\underline{\mathbf{0}}$	3.16	4.47
4	9.43	7.21	3.16	$\underline{\mathbf{0}}$	7.07
5	10.44	9.06	4.47	7.07	$\underline{\mathbf{0}}$



Step II

Algorithm Agglomerative Clustering Algorithm

```
1: Starting with a dataset of $N$ observations, each begins in its own cluster.
   Form an $N$×$N$ proximity matrix.
2: Identify the two most similar clusters.
3: Merge those clusters and recompute the proximity matrix (Note: it will
   have one fewer row and column than the previous matrix).
4: Repeat steps (2) and (3) until the proximity matrix is 1$x$1.
```

Merging the Two Most Similar Clusters

	1	2	3	4	5
1	**0**	2.24	10.63	9.43	10.44
2	2.24	**0**	8.60	7.21	9.06
3	10.63	8.60	**0**	3.16	4.47
4	9.43	7.21	3.16	**0**	7.07
5	10.44	9.06	4.47	7.07	**0**

Step III

Algorithm Agglomerative Clustering Algorithm

```
1: Starting with a dataset of $N$ observations, each begins in its own cluster.
   Form an $N \times N$ proximity matrix.
2: Identify the two most similar clusters.
3: Merge those clusters and recompute the proximity matrix (Note:
   it will have one fewer row and column than the previous matrix).
4: Repeat steps (2) and (3) until the proximity matrix is $1x1$.
```

Recomputing the Proximity Matrix

	12	3	4	5
12	**0**			
3		**0**	3.16	4.47
4		3.16	**0**	7.07
5		4.47	7.07	**0**

How do we account for a multi-observational cluster?

Measure of Dissimilarity

- Complete Linkage

Definition

Let G and H represent two clusters. The dissimilarity $d(G,H)$ between G and H is computed from the set of pairwise observation dissimilarities ij where one member of the pair i is in G and the other j is in H. The dissimilarity of G and H with complete linkage is computed as follows:

$$d_{CL}(G, H) = \max_{i \in G, j \in H} d_{ij}$$

- Average Linkage
- Single Linkage

Recomputing the Proximity Matrix (cont.)

	12	3	4	5
12	**0**			
3		**0**	3.16	4.47
4		3.16	**0**	7.07
5		4.47	7.07	**0**

$$d_{CL}(12,3) = \max_{i \in G, j \in H} d_{ij}$$

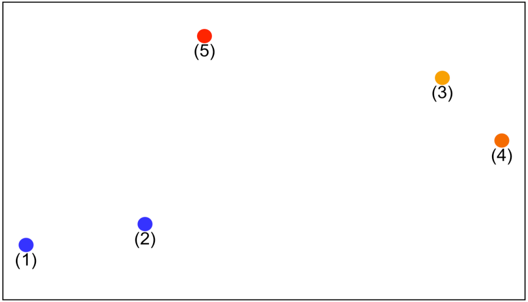
$$d_{CL}(12,3) = \max(d_{euc}(13), d_{euc}(23))$$

$$d_{CL}(12,3) = \max(10.63, 8.60)$$

$$d_{CL}(12,3) = 10.63$$

Updated Proximity Matrix

	12	3	4	5
12	**0**	10.63	9.43	10.44
3	10.63	**0**	3.16	4.47
4	9.43	3.16	**0**	7.07
5	10.44	4.47	7.07	**0**



Step IV

Algorithm Agglomerative Clustering Algorithm

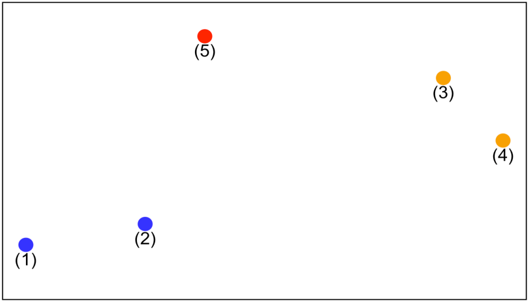
```
1: Starting with a dataset of $N$ observations, each begins in its own cluster.
   Form an $N$x$N$ proximity matrix.
2: Identify the two most similar clusters.
3: Merge those clusters and recompute the proximity matrix (Note: it will
   have one fewer row and column than the previous matrix).
4: **Repeat steps (2) and (3) until the proximity matrix is $1x1$.**
```

Applying the Algorithm

	12	3	4	5
12	**0**	10.63	9.43	10.44
3	10.63	**0**	3.16	4.47
4	9.43	3.16	**0**	7.07
5	10.44	4.47	7.07	**0**

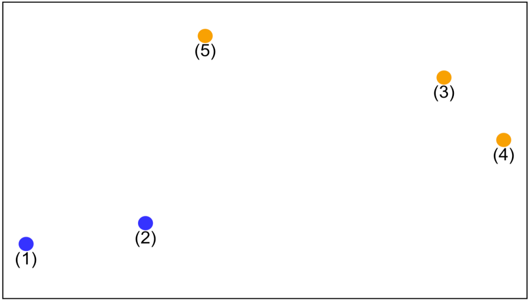
Applying the Algorithm (Cont.)

	12	34	5
12	**0**	10.63	10.44
34	10.63	**0**	4.47
5	10.44	4.47	**0**



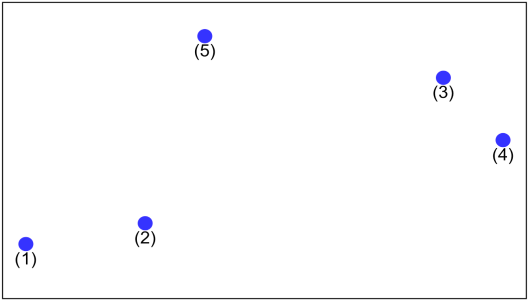
Applying the Algorithm (Cont.)

	12	345
12	**0**	10.63
345	10.63	**0**

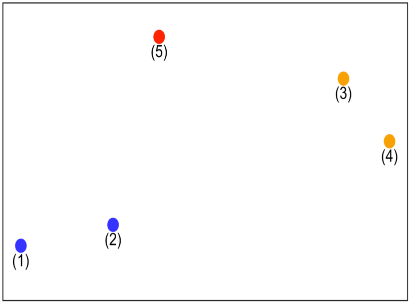


Applying the Algorithm (Cont.)

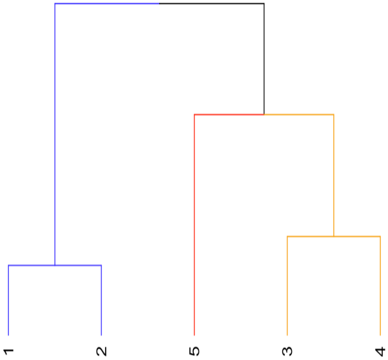
	12345
12345	0



Visualizing the Hierarchical Relationship



Observations



Dendrogram

Outline

1. What is Cluster Analysis?

2. Applying the Agglomerative Clustering Algorithm

3. Applying Hierarchical Clustering to NBA Data

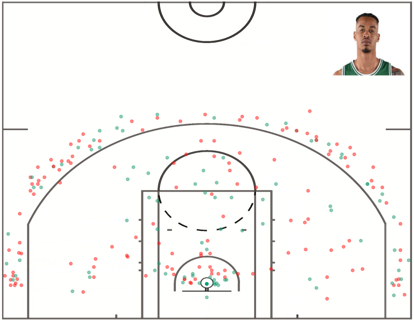
NBA Clustering

- Goal
  - To cluster players into archetypes based on skillsets

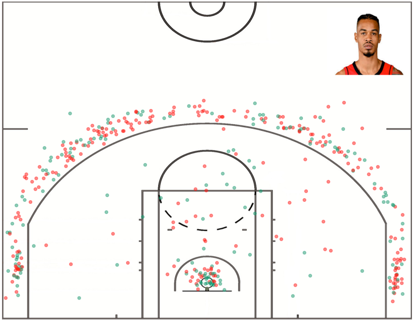
- Challenges
  - Must use player performance to determine ability. These are not the same.
  - Accounting for variables such as team philosophy, role in offense, level of teammates, etc.

Challenges (cont.)

It is unlikely Gerald Green's shooting ability improved in year 13.



Boston 2016-17 (47 gp)



Houston 2017-18 (41 gp)

The Data

- Data was collected via NBA.com, synergy sports, pbpbasketball.com, basketball-reference.com, and fivethirtyeight.com by scraping the sites.

- Tracking data did not become available until the 2013-14 season, so I only focused on the past 5 seasons.

The Data

- Data was collected via NBA.com, synergy sports, pbpbasketball.com, basketball-reference.com, and fivethirtyeight.com by scraping the sites.

- Tracking data did not become available until the 2013-14 season, so I only focused on the past 5 seasons.

Player	Avg. Sec Per Touch	Avg. Dribble Per Touch	Pts. Per Touch
LeBron James	4.52	3.32	0.32
Stephen Curry	3.73	3.45	0.36
James Harden	6.37	5.92	0.41

Table: 2018-19 Per Second Spectrum

Proximity Matrix

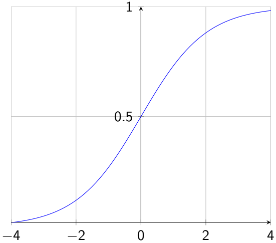
- For each season, an 8 component vector was created for each player.
- The Components:
  1. **3PT Shooting**
  2. **Mid-Range Scoring**
  3. **Inside Scoring**
  4. **Roll Gravity** (how effective a player is in the pick-and-roll)
  5. **Playmaking** (ability to create shots for teammates)
  6. **Self-Creation** (ability to get own shot)
  7. **Rebounding**
  8. **Defense**

3PT Shooting Component (Unadjusted)

- Classified 3pt attempts as contested, open, or wide open based on defender distance.
- For each category, calculated the players *shooting ability over expectation* (SAOE).
- Ex: $(SAOE)_{open} =$ proportion of open attempts (fgp-expected fgp)
- Also considered the location of shot.
- $unadjustedshootingcomponent = \sum_{i \in A}(SAOE_i)$
  - A = (Tightly Contested, Open, Wide Open, ATB, Corner)

3PT Shooting Component (Adjusted)

- Normalized the number of 3 point field goal attempts for each player.
- Passed the normalized variable through a sigmoid function to obtain a value, r.
- Then $SAOE_{adjusted} = (r)\ SAOE_{unadjusted}$ (given $SAOE_{unadjusted} > 0$).



$$S(x) = \frac{e^x}{1+e^x}$$

Name	Season	FG3A	Component
Kyle Korver	2014-15	449	0.2637
JJ Redick	2015-16	413	0.2134
Stephen Curry	2015-16	884	0.2089
Joe Harris	2018-19	386	0.2059
Kyle Korver	2013-14	392	0.1972
Stephen Curry	2014-15	646	0.1931

2018-19 Hierarchical Clustering Results

- There were 423 players who played at least 15 games.
- I created 15 clusters (avg. 28 player per cluster)
- 5 clusters had an average VORP > 0.
  - **Superstars**
  - **Supporting Stars**
  - **2-way Bigs**
  - **Playmaking Defenders**
  - **3&D Wings**

Archetypes

Archetype	PPG	RPG	APG	AVG. SALARY
Superstars	23.6	7.1	5.0	24.5M
Supporting Stars	14.2	4.6	3.8	16.6M
2-Way Bigs	15.2	9.5	2.4	15.8M
Playmaking Defenders	8.0	3.9	2.1	7.3M
3&D Wings	10.5	2.9	2.6	7.4M

Archetypes

Archetype	PPG	RPG	APG	AVG. SALARY
Superstars	23.6	7.1	5.0	24.5M
Supporting Stars	14.2	4.6	3.8	16.6M
2-Way Bigs	15.2	9.5	2.4	15.8M
Playmaking Defenders	8.0	3.9	2.1	7.3M
3&D Wings	10.5	2.9	2.6	7.4M

Superstars: LeBron James, Kevin Durant, Kawhi Leonard

Archetypes

Archetype	PPG	RPG	APG	AVG. SALARY
Superstars	23.6	7.1	5.0	24.5M
Supporting Stars	14.2	4.6	3.8	16.6M
2-Way Bigs	15.2	9.5	2.4	15.8M
Playmaking Defenders	8.0	3.9	2.1	7.3M
3&D Wings	10.5	2.9	2.6	7.4M

Superstars: LeBron James, Kevin Durant, Kawhi Leonard
Supporting Stars: Stephen Curry, Kyrie Irving, Kyle Lowry

Archetypes

Archetype	PPG	RPG	APG	AVG. SALARY
Superstars	23.6	7.1	5.0	24.5M
Supporting Stars	14.2	4.6	3.8	16.6M
2-Way Bigs	15.2	9.5	2.4	15.8M
Playmaking Defenders	8.0	3.9	2.1	7.3M
3&D Wings	10.5	2.9	2.6	7.4M

Superstars: LeBron James, Kevin Durant, Kawhi Leonard
Supporting Stars: Stephen Curry, Kyrie Irving, Kyle Lowry
2-Way Bigs: Joel Embiid, Anthony Davis, Rudy Gobert

Archetypes

Archetype	PPG	RPG	APG	AVG. SALARY
Superstars	23.6	7.1	5.0	24.5M
Supporting Stars	14.2	4.6	3.8	16.6M
2-Way Bigs	15.2	9.5	2.4	15.8M
Playmaking Defenders	8.0	3.9	2.1	7.3M
3&D Wings	10.5	2.9	2.6	7.4M

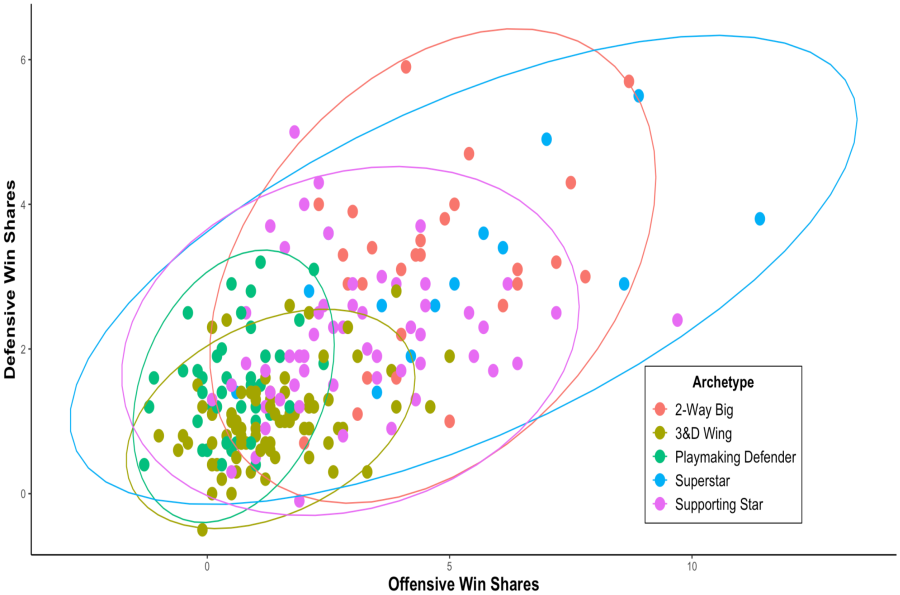
Superstars: LeBron James, Kevin Durant, Kawhi Leonard
Supporting Stars: Stephen Curry, Kyrie Irving, Kyle Lowry
2-Way Bigs: Joel Embiid, Anthony Davis, Rudy Gobert
Playmaking Defenders: Draymond Green, Andre Iguodala, Lonzo Ball

Archetypes

Archetype	PPG	RPG	APG	AVG. SALARY
Superstars	23.6	7.1	5.0	24.5M
Supporting Stars	14.2	4.6	3.8	16.6M
2-Way Bigs	15.2	9.5	2.4	15.8M
Playmaking Defenders	8.0	3.9	2.1	7.3M
3&D Wings	10.5	2.9	2.6	7.4M

Superstars: LeBron James, Kevin Durant, Kawhi Leonard
Supporting Stars: Stephen Curry, Kyrie Irving, Kyle Lowry
2-Way Bigs: Joel Embiid, Anthony Davis, Rudy Gobert
Playmaking Defenders: Draymond Green, Andre Iguodala, Lonzo Ball
3&D Wings: Klay Thompson, Robert Covington, P.J. Tucker

2018-19 Hierarchical Clustering Results



Future Research

- Linking clusters by season
- Predicting which clusters rookies will belong to
- Modeling career trajectories

Thank You

- Thank you to everyone who helped and supported me during my four years at Xavier!