Cluster Analysis of a Pitcher's Arsenal Using K-Means and PCA for Player Comparison

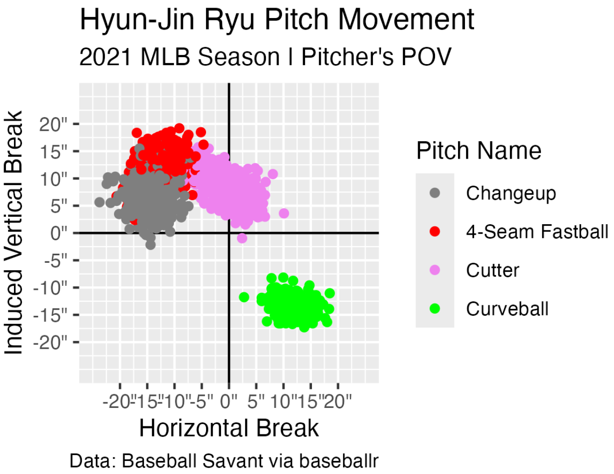
Andy Shin

October 23, 2022

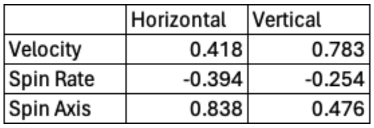
Pitch movement has become a critical metric for analyzing a pitcher's repertoire, as it reveals distinct clusters of pitch types when visualized on an x-y plane. In this context, the x-axis represents the degree of induced vertical break, and the y-axis represents the degree of induced horizontal break. Induced pitch movement describes the deviation of a baseball's trajectory from a straight path, primarily influenced by spin, speed, and aerodynamic forces as the ball travels toward the plate. This analysis disregards gravitational effects on the ball's movement, focusing exclusively on the impact of spin and velocity.

Observing clusters on the 2D movement plane sparked the idea that clustering algorithms could be applied to classify pitch types based on features closely related to movement. Using the k-means clustering algorithm, I found that it effectively captured the natural grouping of pitch types according to their movement characteristics. This approach not only highlights the inherent distinctions in pitch types but also offers a data-driven method for identifying patterns in a pitcher's arsenal. Furthermore, I believe that dimensionality reduction techniques, such as principal component analysis (PCA), could further enhance these clusters by refining the key features that differentiate pitches, allowing for even more insightful analysis.This approach enables more comprehensive comparisons between players, allowing teams to identify pitchers with similar profiles. This is because the two principal components obtained from PCA represent a combination of crucial features, capturing both pitch movement and additional pitching data on a single 2D plane. In contrast, the pitch movement plane reflects only the movement aspect of each pitch.

I chose Hyun-Jin Ryu for this cluster analysis, as he was once my role model when I wanted to become a baseball player. I selected comparable players based on pitch velocity and movement data obtained from MLB Statcast.



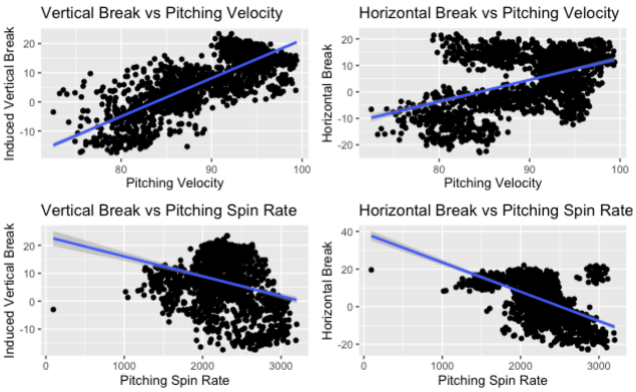
The pitch speed and spin rate affect the ball's movement because ball's aerodynamic is heavily influenced by these two features. I compiled a list of the top 100 pitchers from the 2021 season using Bleacher Report's rankings, and calculated the correlations between pitch speed/spin rate and horizontal/vertical movement. I set x-axis as pitch speed/spin rate and y-axis horizontal/vertical break. Below is a table of correlation:



	Horizontal	Vertical
Velocity	0.418	0.783
Spin Rate	-0.394	-0.254
Spin Axis	0.838	0.476

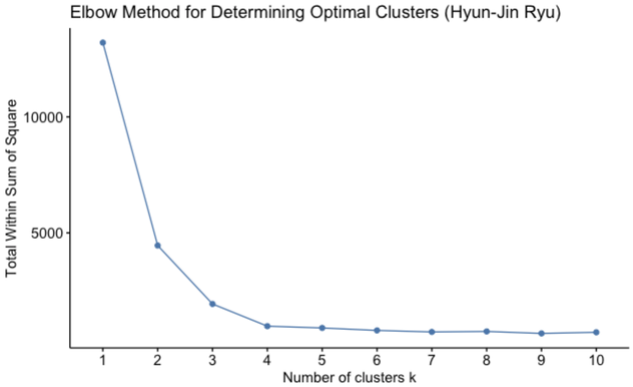
Pitching velocity has a strong correlation with vertical movement, and spin rate has a moderate correlation with horizontal break. For four-seam fastballs, higher velocities often correlate with increased backspin, which can result in less vertical drop, giving the illusion of the ball "rising" as it approaches the batter. This effect is due to the Magnus force acting against gravity, causing the ball to stay elevated longer.

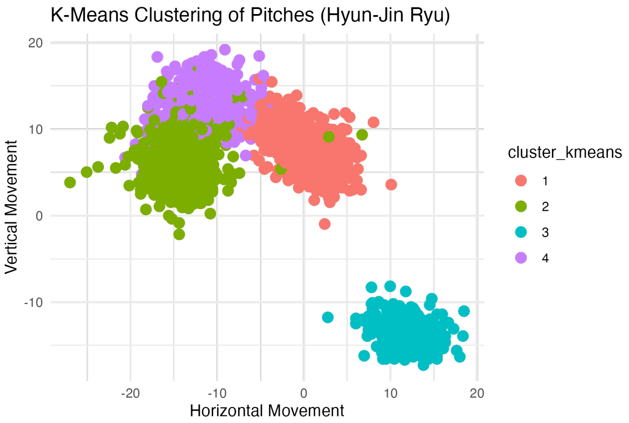
Analyzing the average spin rates across various pitch types reveals that breaking balls, such as curveballs and sweepers, exhibit higher spin rates among the 100 RHP players studied. These elevated spin rates contribute to significant horizontal movement, causing the pitches to break toward or away from the batter as they approach home plate. This pronounced lateral movement is a defining characteristic of these pitch types.



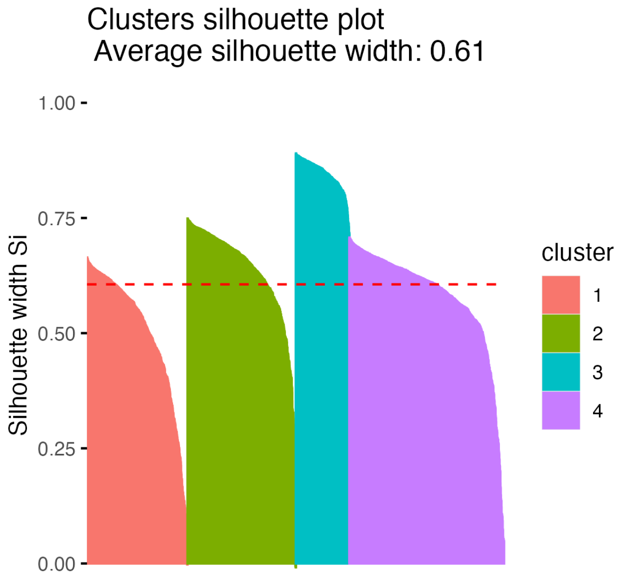
Moreover, the strong correlation between spin axis and horizontal movement highlights that the orientation of the ball's spin determines the direction and magnitude of its lateral movement. This implies that a spin axis tilted at an angle can cause the ball to break horizontally, as seen in sliders or curveballs.

This led me to conduct k-mean clustering algorithm with selecting features such as vertical/horizontal break, pitch velocity/spin rate, and spin axis. Hence I can experiment with different numbers of clusters to see how the groups form. I used Elbow Method to determine the optimal number of clusters for k-means.

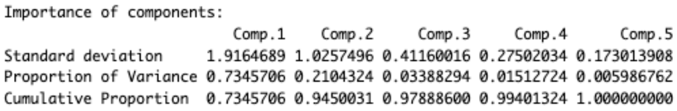




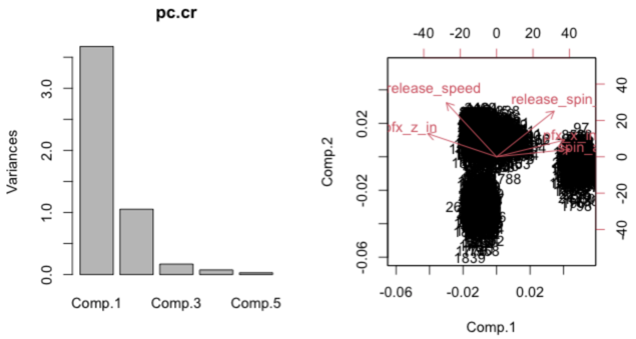
To assess the quality of cluster separation, I calculated the silhouette score and obtained a value of 0.61. This score suggests that the clusters are reasonably well-defined, as it indicates that, on average, samples are closer to their own cluster than to neighboring clusters. This is because silhouette scores range from -1 to 1, with higher values indicating better-defined clusters.

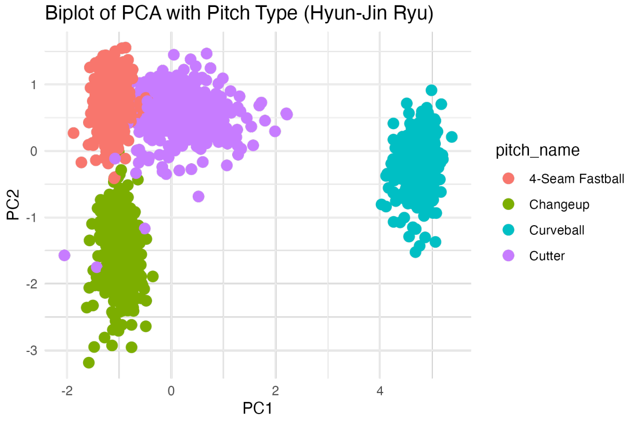


Since pitch movement provides crucial insights into how each pitch type forms clusters in a 2D space, it is valuable for analyzing a pitcher's arsenal. In a PCA analysis, I am interested in whether the principal components derived from multiple features, such as pitch speed/spin rates and axis and vertical/horizontal breaks, can reveal groupings, allowing clusters to form that could reflect similar characteristics in players with similar pitching styles. This approach could highlight how additional features contribute to defining the components and potentially reveal patterns across players.

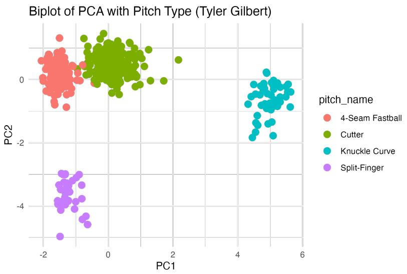


```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4      Comp.5
Standard deviation     1.9164689 1.0257496 0.41160016 0.27502034 0.173013908
Proportion of Variance 0.7345706 0.2104324 0.03388294 0.01512724 0.005986762
Cumulative Proportion  0.7345706 0.9450031 0.97888600 0.99401324 1.000000000
```

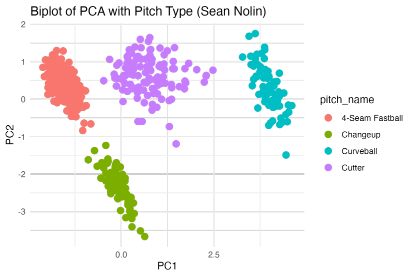




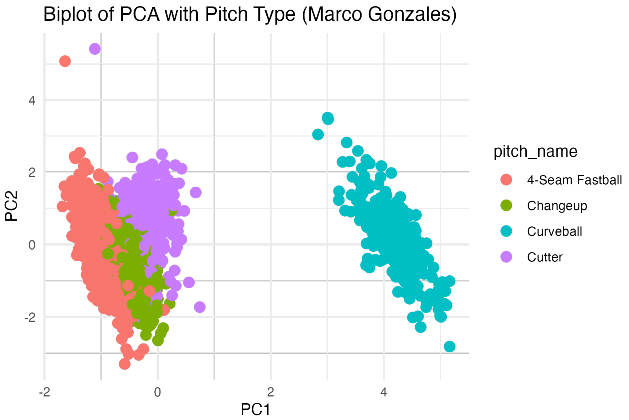
With two components explaining 94% of the variance, we can easily plot the data in a 2D space. As shown below, the points in the PCA plot form distinct groupings based on pitch type, indicating a clear pattern unique to each pitch in his arsenal. The MLB Statcast suggests that the pitchers similar to Hyun-Jin Ryu based on Velocity and Movement are 2021-Tyler Gilbert, 2021-Sean Nolin, and 2022-Marco Gonzales.



Biplot of PCA with Pitch Type (Tyler Gilbert)



Biplot of PCA with Pitch Type (Sean Nolin)



Biplot of PCA with Pitch Type (Marco Gonzales)

Surprisingly, the cluster for the 4-seam fastball in Gilbert's, Nolin's, and Gonzales' PCA space is located in a similar region to Ryu's, suggesting a common approach to this pitch type across these players. Additionally, the cutter clusters for all three players are closely aligned with Ryu's, indicating a comparable movement and velocity profile. Nolin's changeup and curveball clusters also appear in regions similar to those of Ryu, while Gilbert's knuckle curve—a slight variation of the curveball—follows a comparable region, hinting at a similar breaking profile. Interestingly, in Gilbert's PCA plane, the split-finger pitch occupies a position comparable to Ryu's changeup, suggesting that Gilbert's split-finger pitch might serve a similar tactical role to Ryu's changeup, likely used to disrupt timing and vary pitch velocity. This alignment across different players' pitch clusters reflects consistent pitch characteristics that could be useful for comparative analysis.

Reference:

- Major League Baseball. (n.d.). Hyun-Jin Ryu Statcast, Visuals & Advanced Metrics. Baseball Savant. https://baseballsavant.mlb.com/savant-player/hyun-jin-ryu-547943?stats=statcast-r-pitching-mlb

- Reuter, J. (2021, October 27). Ranking the top 100 MLB players of the 2021 season. Bleacher Report. https://bleacherreport.com/articles/2949847-ranking-the-top-100-mlb-players-of-the-2021-season
- RPP Baseball. (2019, July 1). Baseball spin axis, spin rate, spin efficiency explained. Rockland Peak Performance. https://rocklandpeakperformance.com/baseball-spin-axis-spin-rate-spin-efficiency-explained/