Clustering NBA Players

Zongkai Tian

PPT

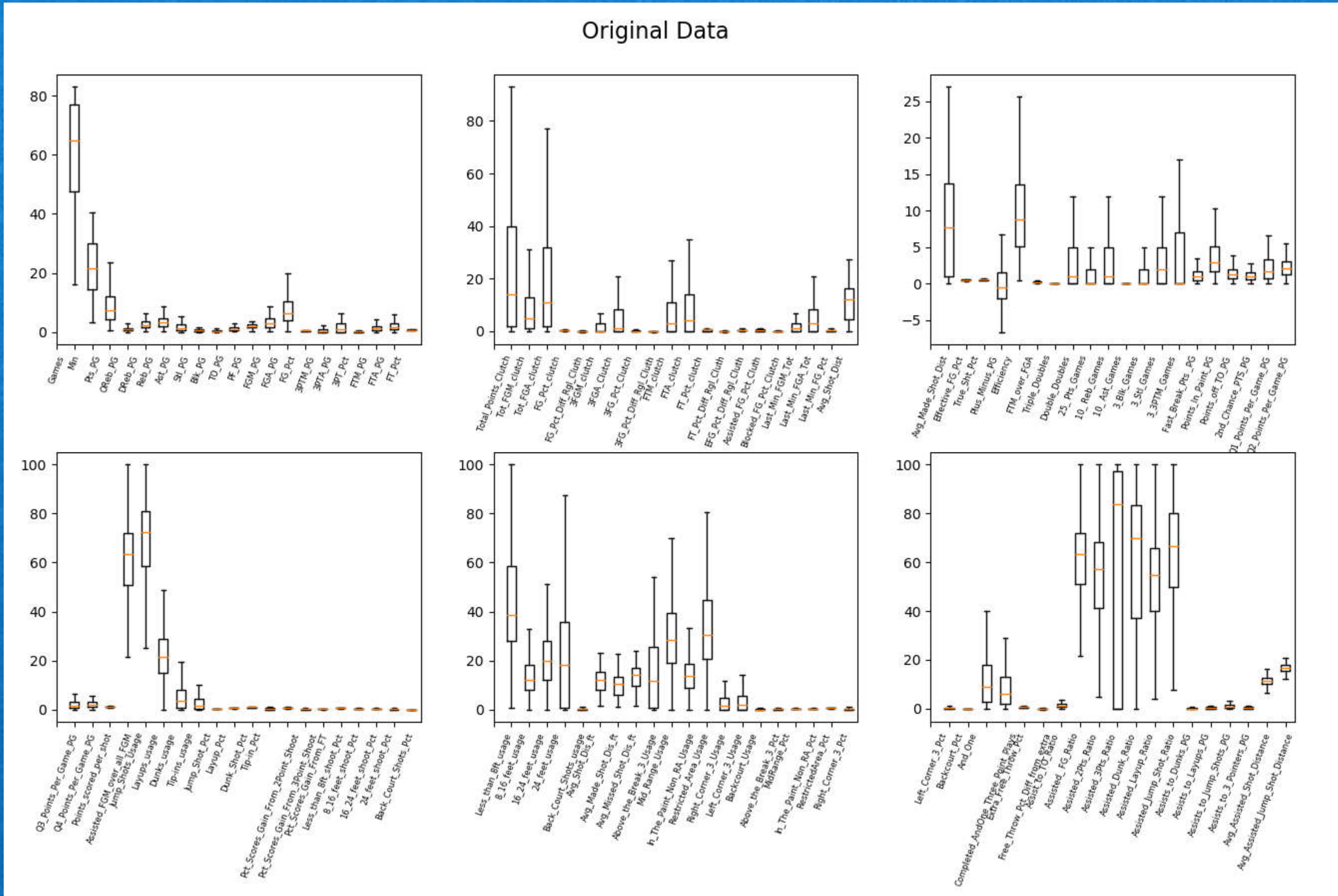
Purpose

- Group players based on their performance and playing style.
- When a player leaves the team, manager can acquire a replacement player from the same cluster without affecting team strategy.

Data

- Scraped from 2010-2011 regular season
- 411 Players
- 119 features (all numerical)
  - "Counting" features (i.e. Games Played, Points, Rebounds, Efficiency, Avg Shooting Distance, etc)
  - "Percentage" features (i.e. Field Goal Percentage, 8-16 ft. Usage, Assisted 3-Pts Ratio)

Original Data



PPT

Pre-processing Data

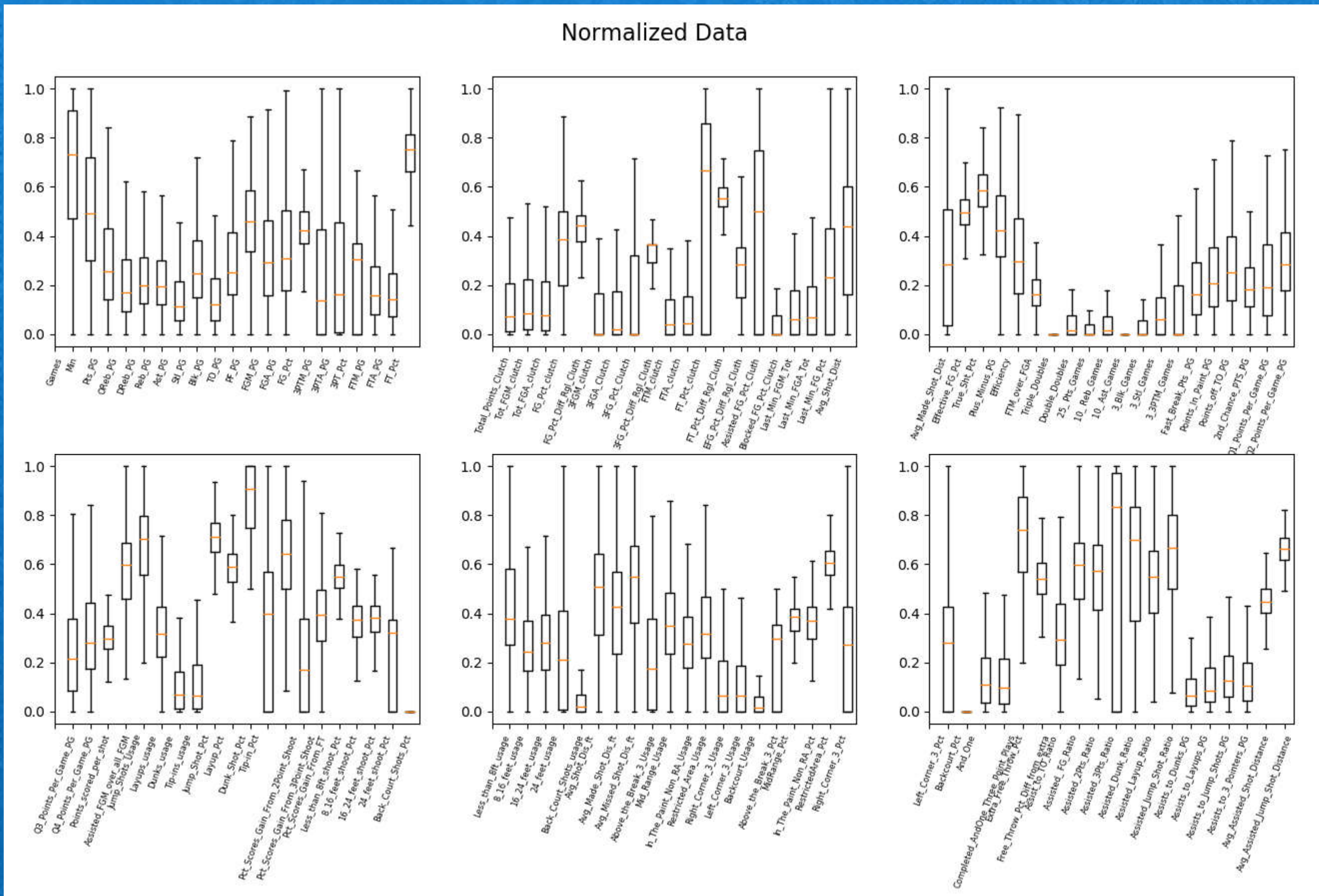
- Scaling percentage data to range [0 – 100]

- Normalizing all features by

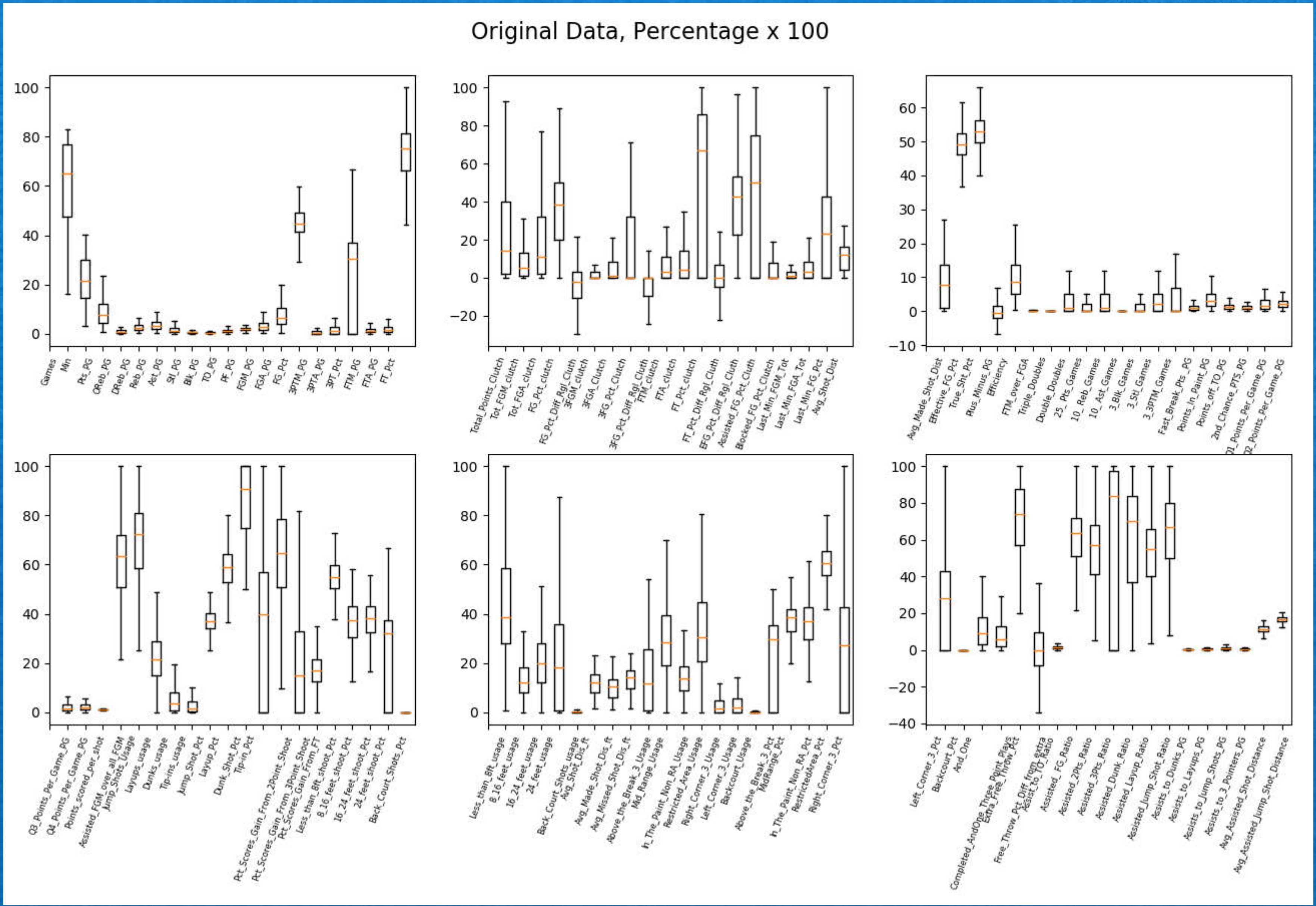
$$\bar{X} = (X - X_{min})/(X_{max} - X_{min})$$

  - Range is [0 – 1]

Normalized Data

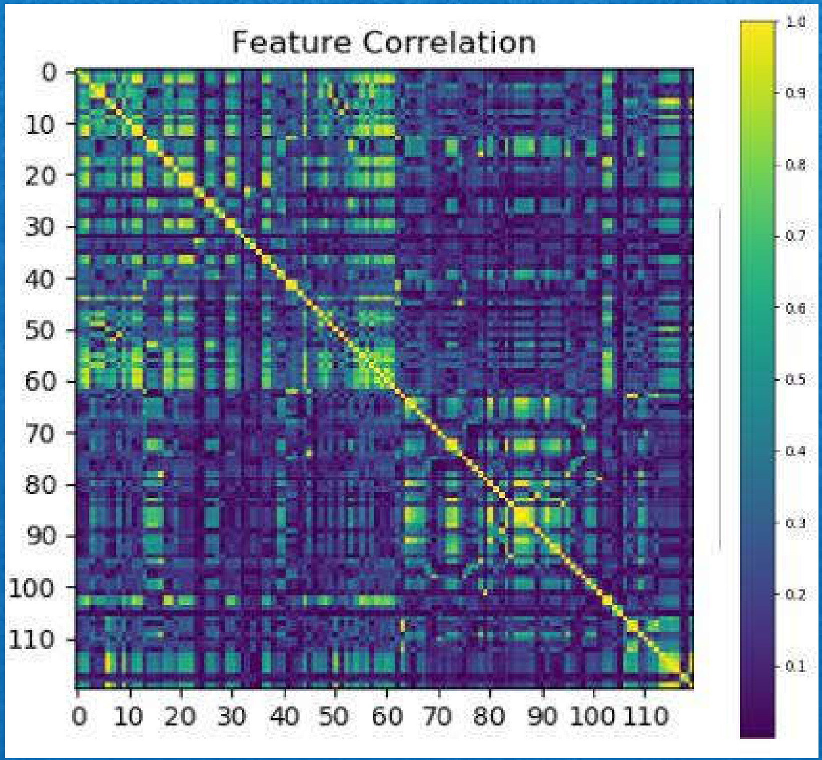


Scaling Percentage by 100



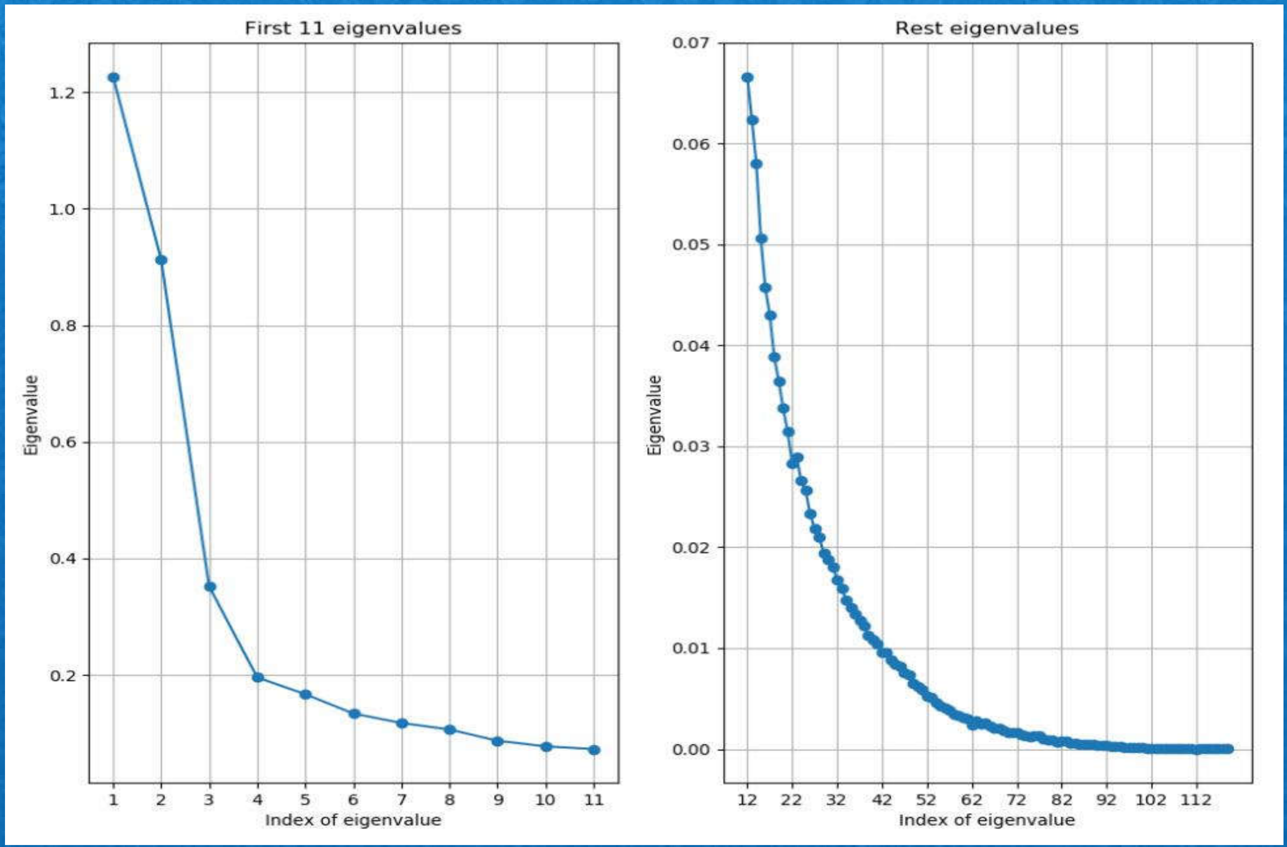
Dimensionality Reduction

- **PCA** - Feature Correlation Coefficient
  – Yellow box indicates correlated features



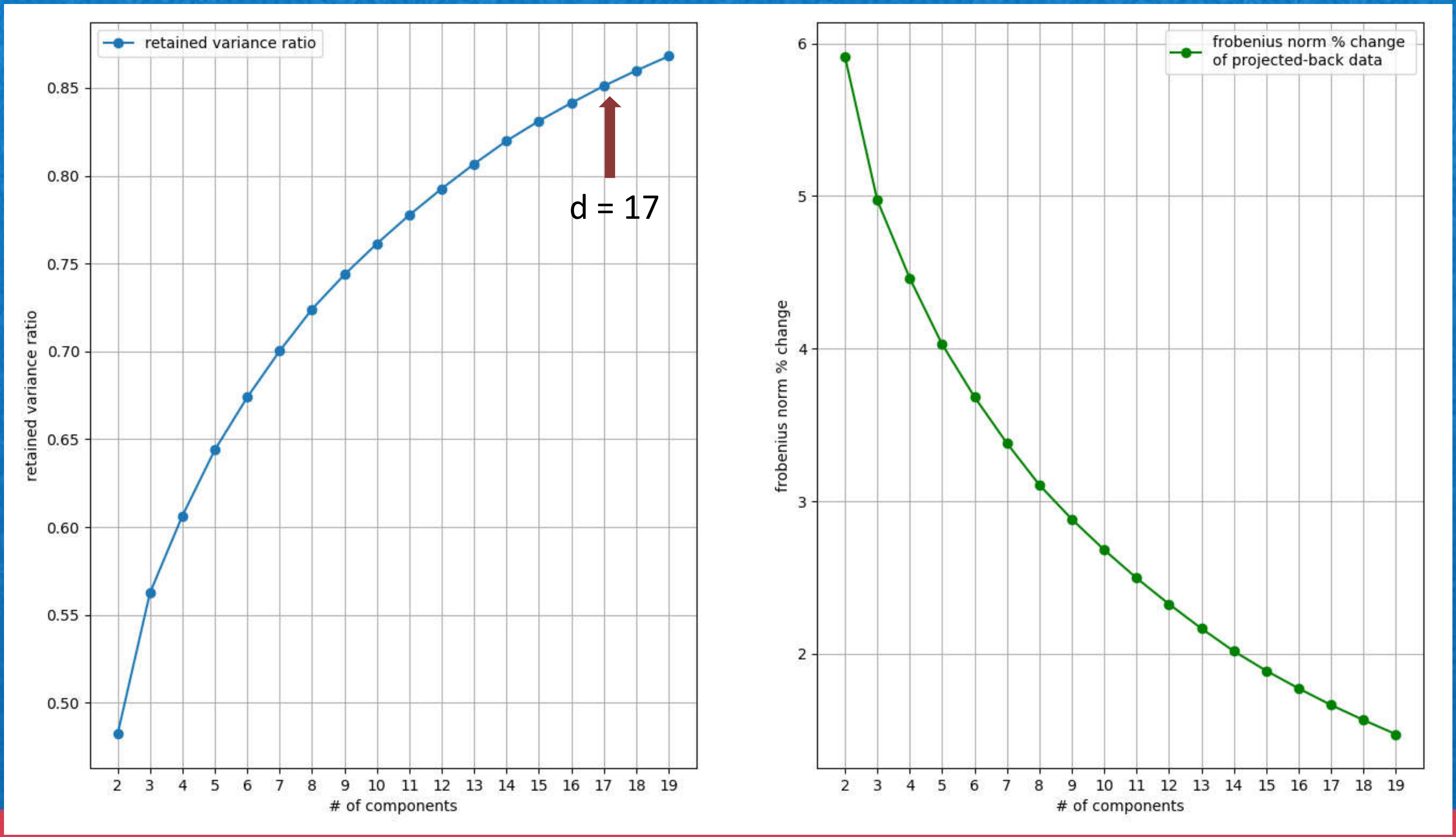
Dimensionality Reduction

- **PCA** – Eigenvalues
  - First 2 eigenvectors have much high eigenvalues



Dimensionality Reduction

- **PCA** – Deciding # of components
  - 85% retained variance as cut-off



Dimensionality Reduction

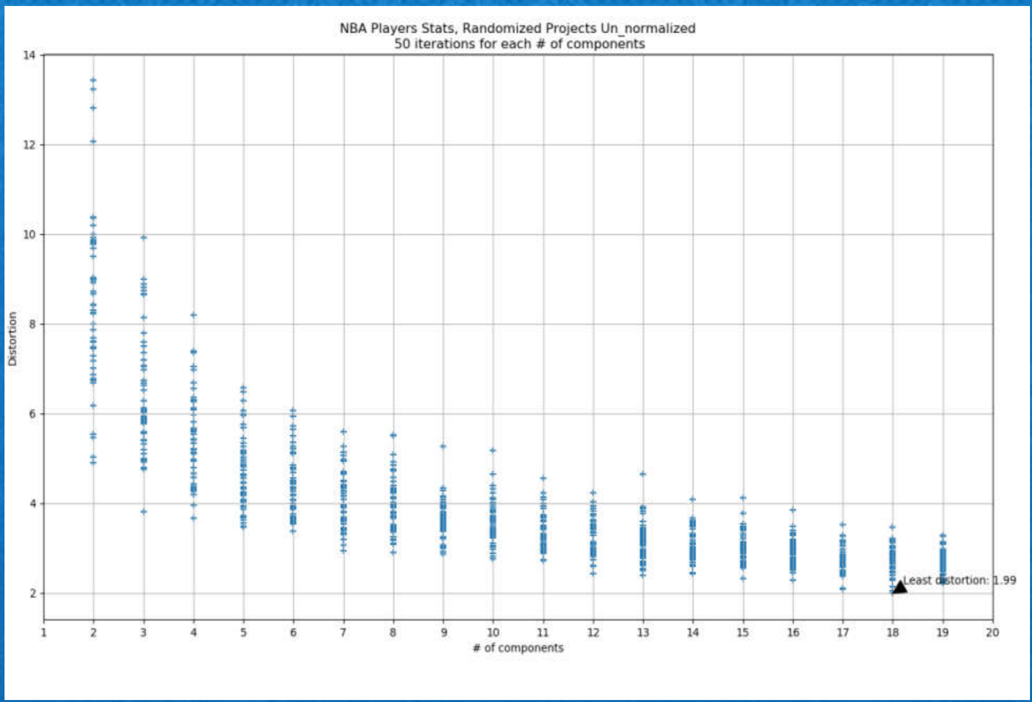
- Random Projection –Theoretical Analysis
  - If lower dimension d satisfies

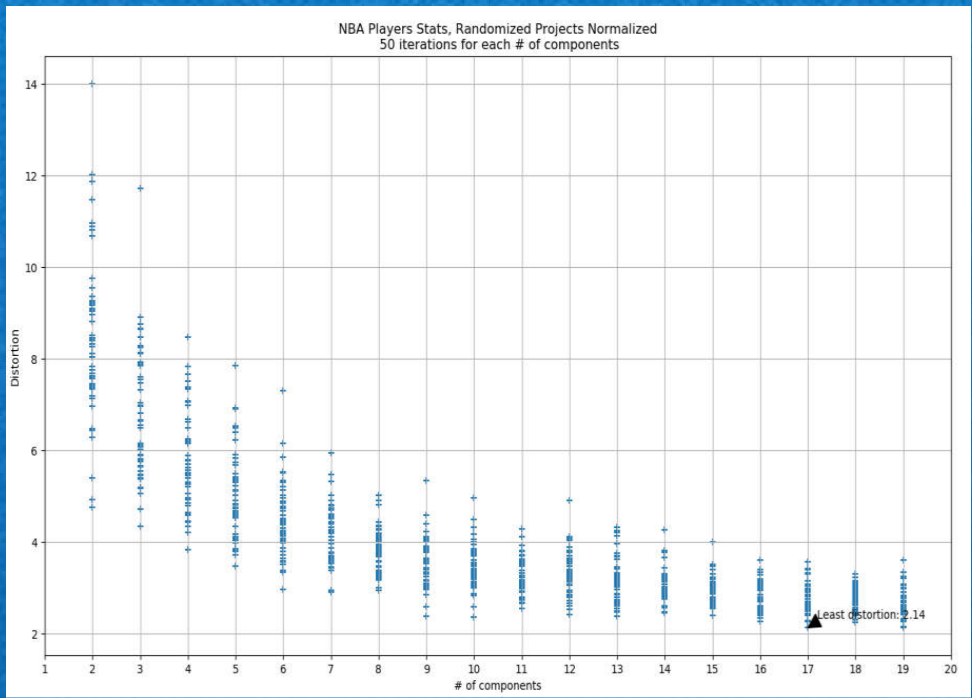
$$d > O(\frac{logn}{\epsilon^2}) = O(\frac{log411}{\epsilon^2}) = O(\frac{2.61}{\epsilon^2})$$

  - Then distortion D is at most $\frac{1+\epsilon}{1-\epsilon}$
  - If D=2 is required, $\epsilon = \frac{D-1}{D+1} = 0.33$
  - Then $d > O(\frac{2.61}{0.33^2}) > O(24)$

Dimensionality Reduction

Random Projection – Finding # of components of least distortion. (50 iterations on each #)





Original Data
D = 1.99

Normalized Data
D = 2.14

Dimensionality Reduction

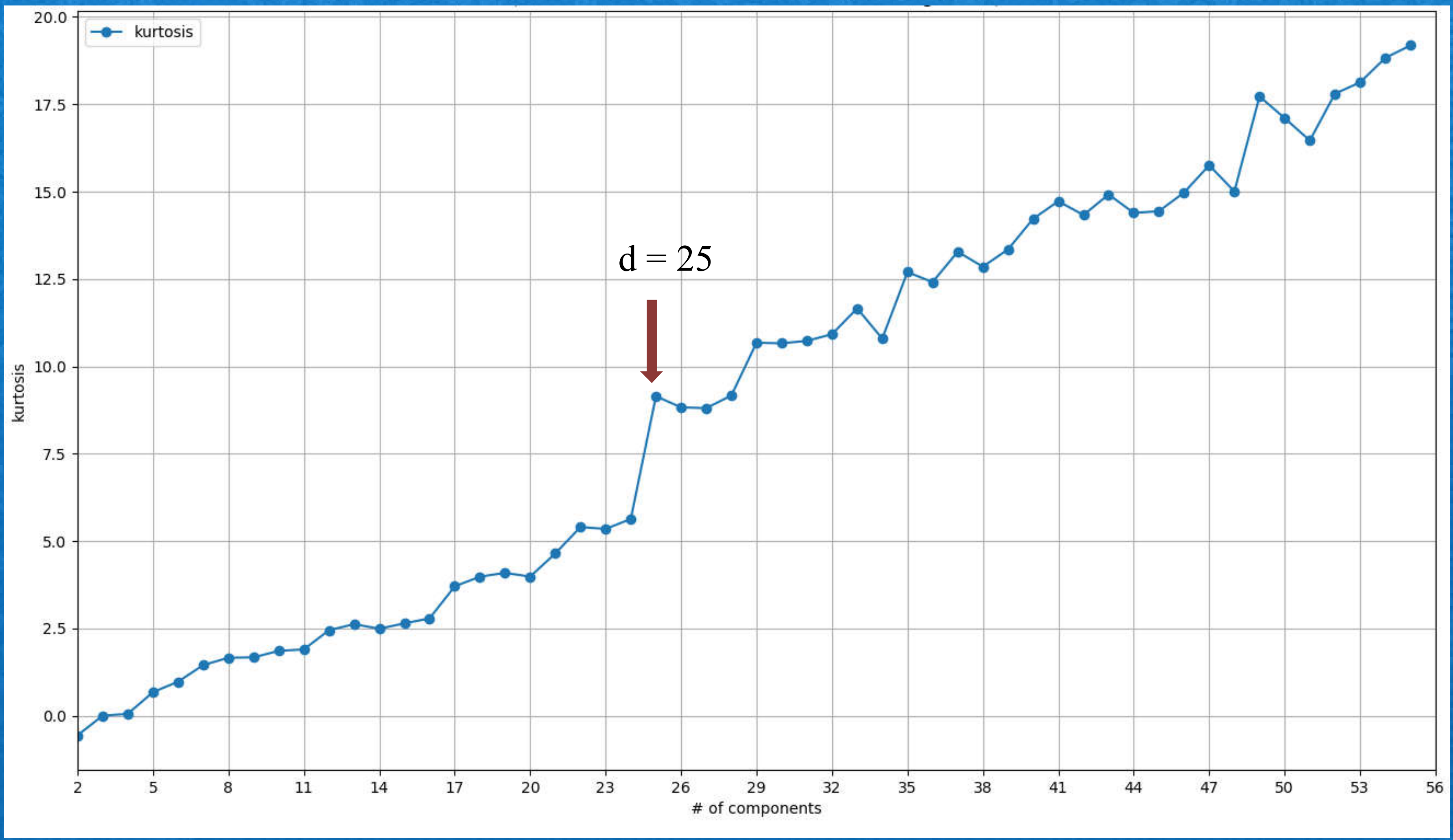
ICA

- Hypothesis: Observed data is a mixture of independent non-Gaussian like distributions
- Reflects a player's variant abilities (Basketball Genes)???

- Measure if a distribution is Gaussian-like
  - Kurtosis: $\mathbb{E}[y^4] - 3(\mathbb{E}[y^2])^2$
  - Kurtosis = 0 for ~N(0,1)

Dimensionality Reduction

ICA



Clustering Algorithm

K-means

- Finding k partitions that minimize cost function

$$cost := \sum_{j=1}^{k} \sum_{x_i \in p_j} ||x_i - \mu_j||^2$$

- K-means++
  - Pick successive centers from points far from selected centers (with high probability)
  - Used to initialize centers for k-means
  - <O(logk)OPT

Clustering Algorithm

K-means

- Deciding value of k, intuitively
  - Small k ➔➔ a long list for manager to select players ➔➔ infeasible
  - Larger k and smaller cluster is preferred
  - Ideally 6-10 players per cluster

➔ k = 40 – 70

Clustering Algorithm

K-means

- Deciding value of k, mathematically
- Silhouette Score (S-score)

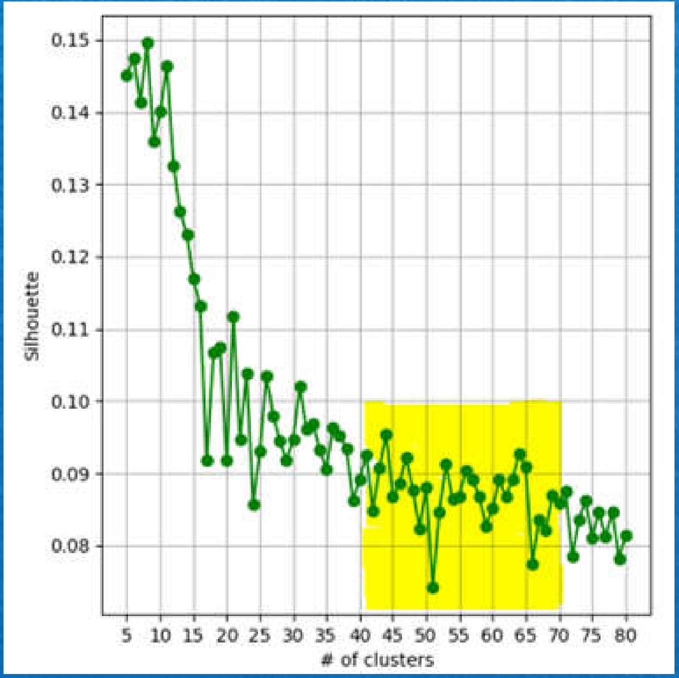
$$S = \frac{b - a}{max(a, b)}, \quad -1 \leq S \leq 1$$

  - **a**: mean distance b/t a sample and all other points in the same cluster.
  - **b**: mean distance between a sample and all other points in the next nearest cluster.

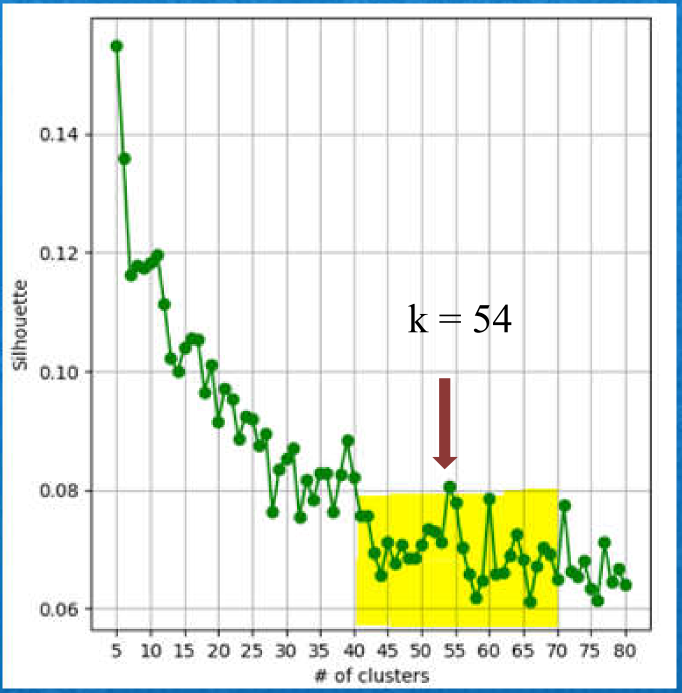
Clustering Algorithm

K-means

- Deciding value of k, finally



Original Data

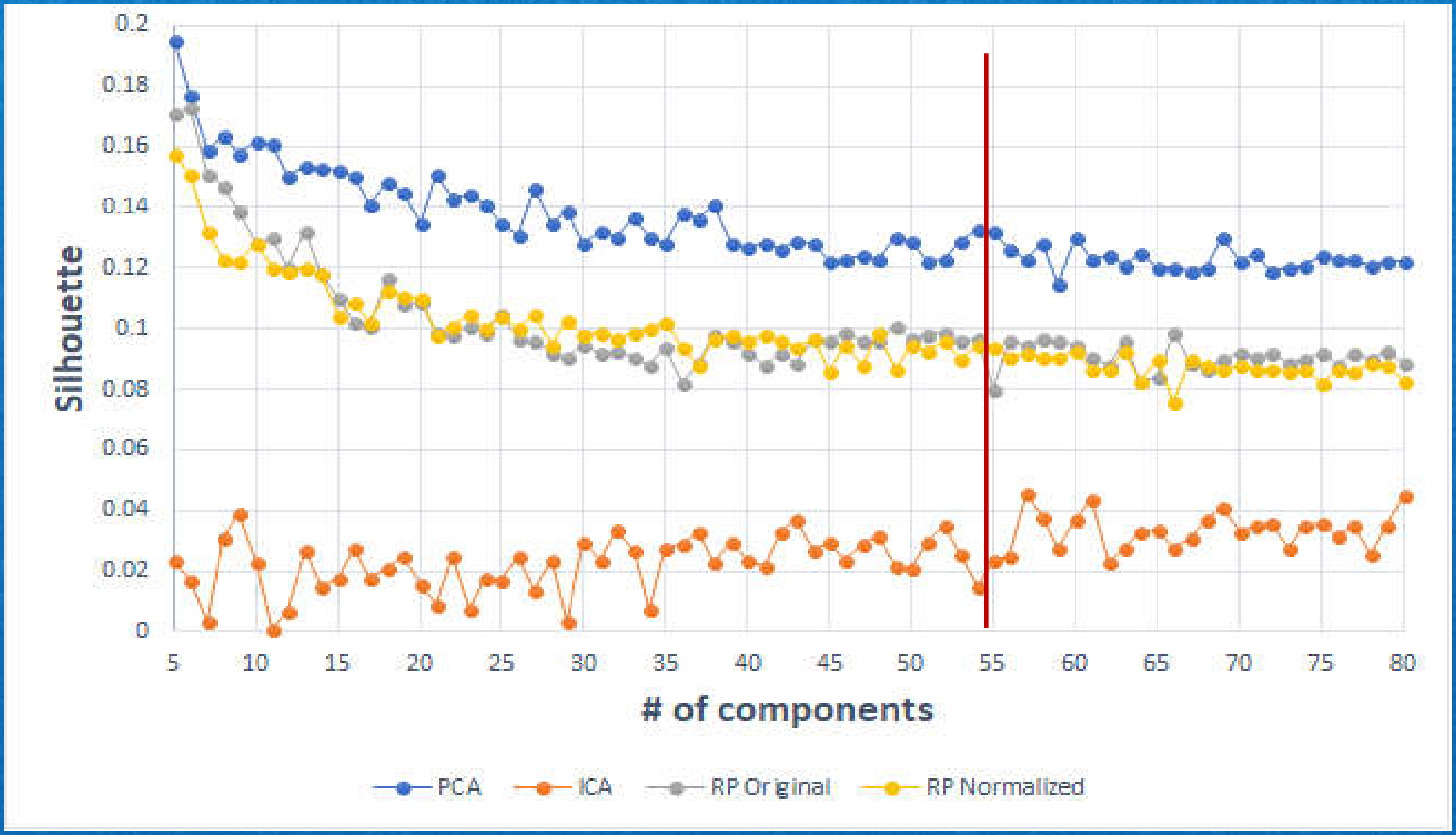


Normalized Data

Clustering Algorithm

K-means

- Dimensionality Reduction Data



Clustering Algorithm

Gaussian Mixture EM

- Assume data is a mixture of multivariate Gaussian distributions
- EM can find most likelihood k Gaussian distributions
- Hidden variables $< z_1, z_2, \cdots z_k >$, where z_j is the probability that x generated by distribution j.

Clustering Algorithm

Gaussian Mixture EM

- E-M steps
  - E step: re-calculate each z from previous-assigned gaussian distribution
  - M step: update centers by taking weighted average

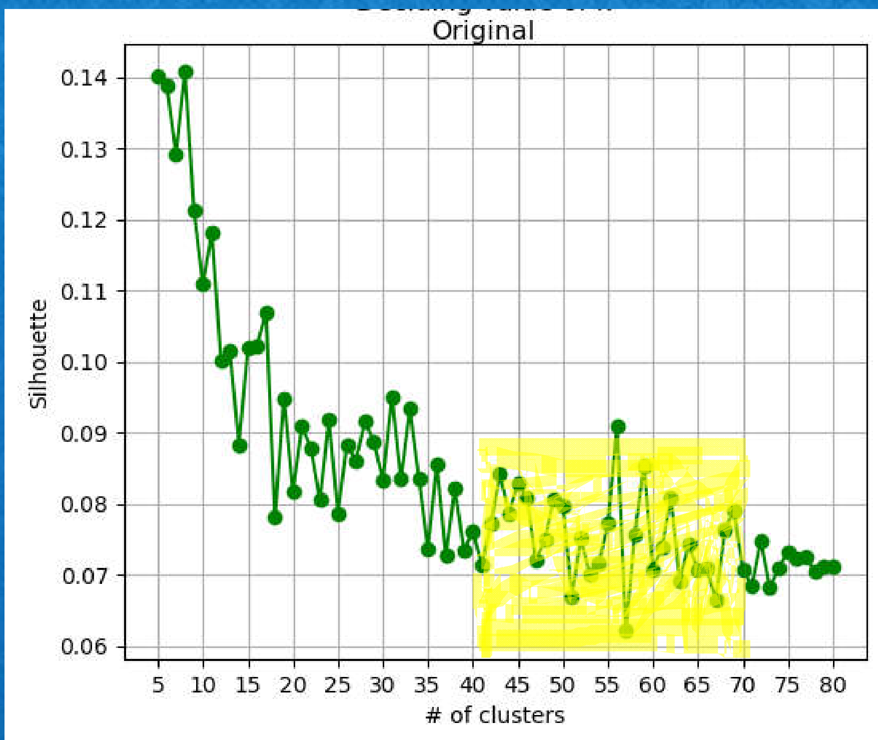
$$\text{“E” step:} \quad \mathbb{E}[z_{ij}] = \frac{p(x = x_i | \mu = \mu_j)}{\sum_{l=1}^{k} p(x = x_i | \mu = \mu_l)}$$

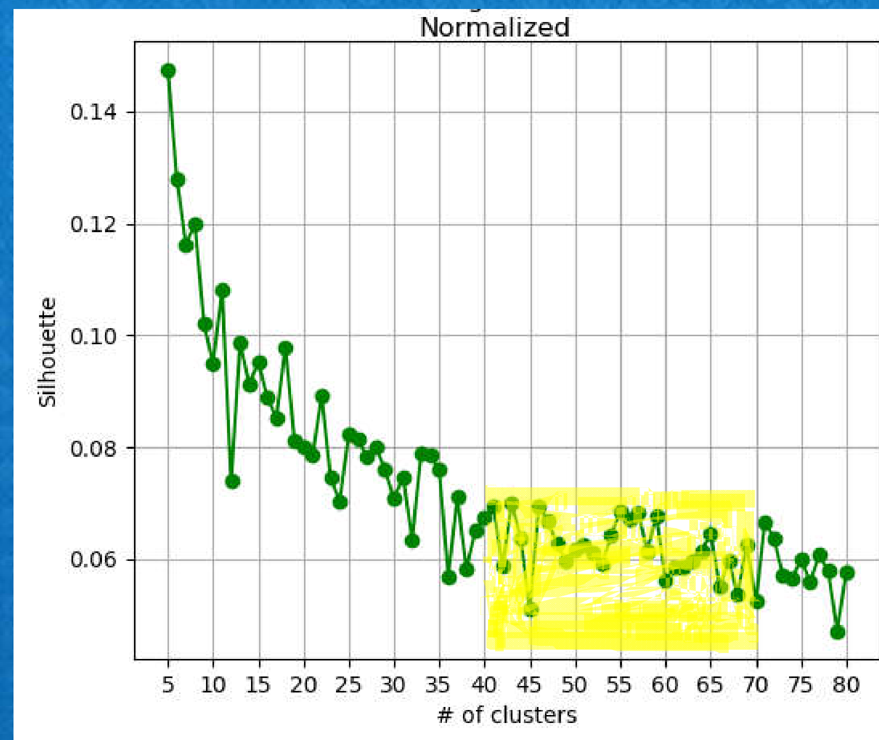
$$\text{“M” step:} \quad \mu_j = \frac{\sum_i \mathbb{E}[z_{ij}] x_i}{\sum_i \mathbb{E}[z_{ij}]}$$

- Linkage to K-means

Clustering Algorithm

Gaussian Mixture EM

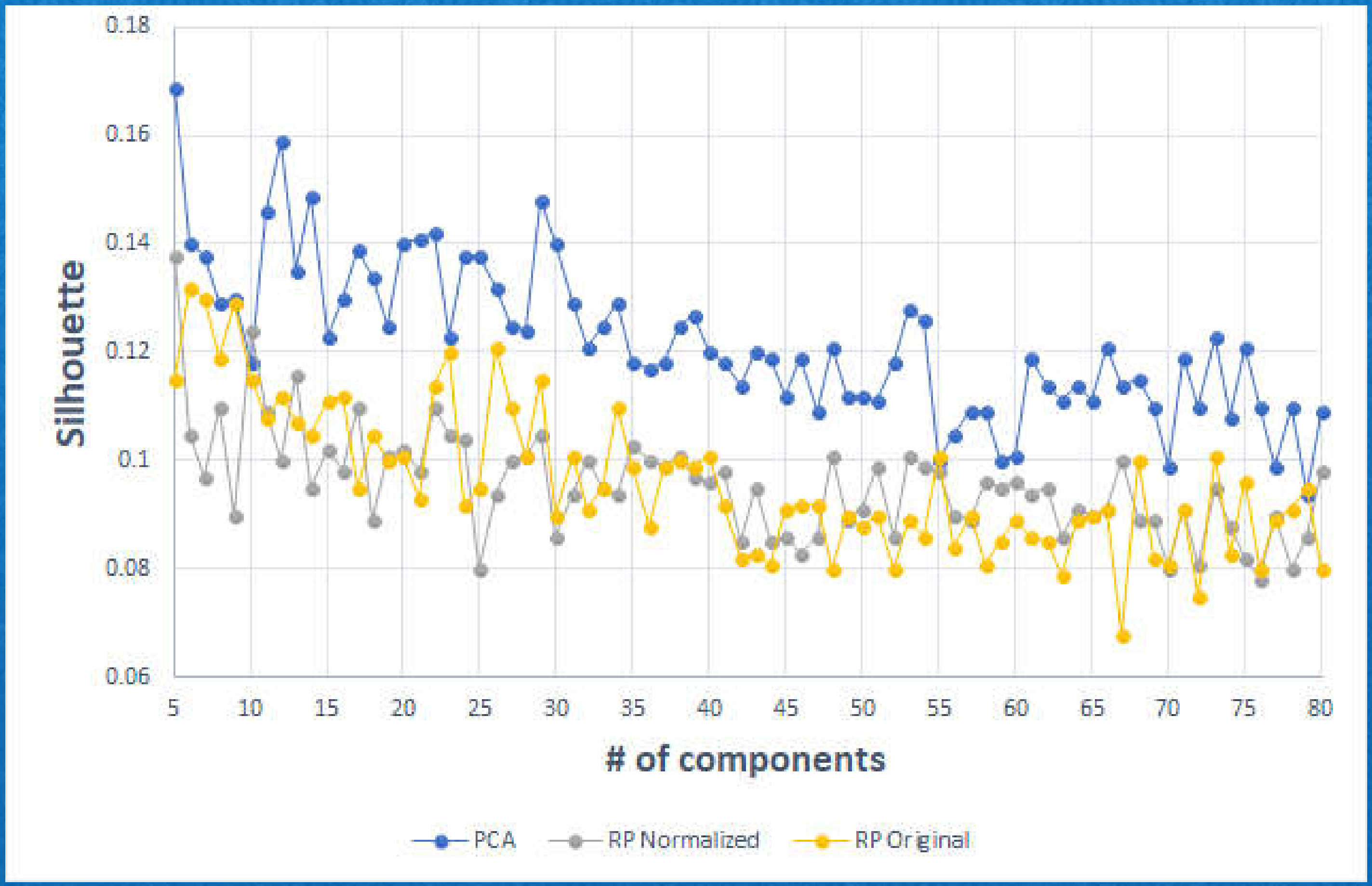




Clustering Algorithm

Gaussian Mixture EM

- Dimensionality Reduction Data



Results

- Manually examine each clustering results
- K = 54
- EM on PCA has best result

Data	K-means		Gaussian Mixture EM	
	Silhouette	Clustering Examination	Silhouette	Clustering Examination
Original	0.09	OK	0.07	Poor
Normalized	0.08	Good	0.06	Good
PCA_17d	0.13	Better	0.13	Best*
ICA_25d	0.02	Very poor	0.03	Very poor
RP_original_18d	0.11	Poor	0.09	Poor
RP_normalized_17d	0.10	Poor	0.1	Poor

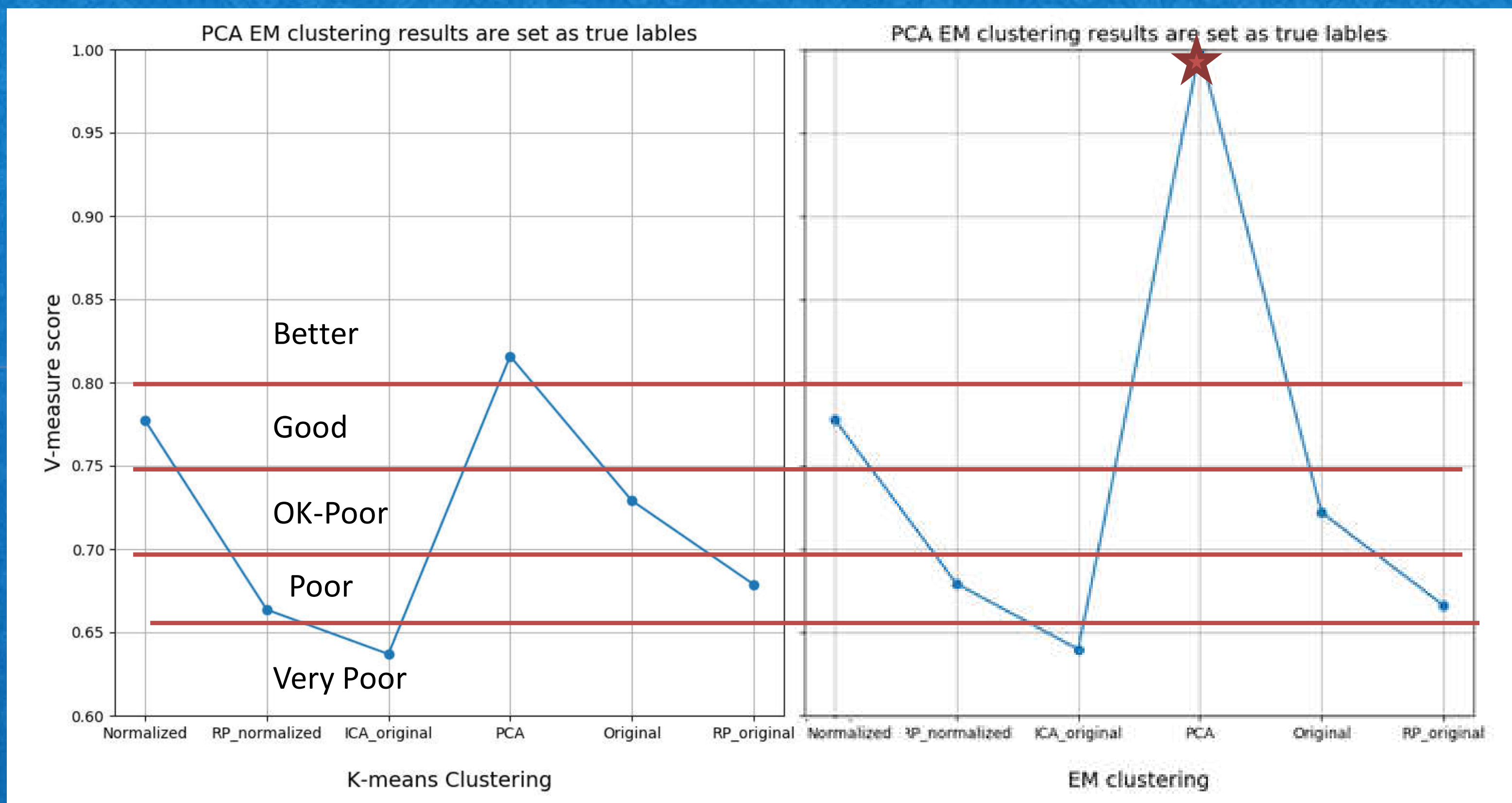
Results

- Measure similarity of two clustering results
  - Assume EM on PCA data is true label
  - V-measure
    - **h**: homogeneity -- how well each cluster contains only members from a single true class
    - **c**: completeness -- if all members from a true class are assigned to the same cluster
    - range [0, 1]: 0-bad; 1-perfect

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

Results

- Link v-measure to clustering examination



Results

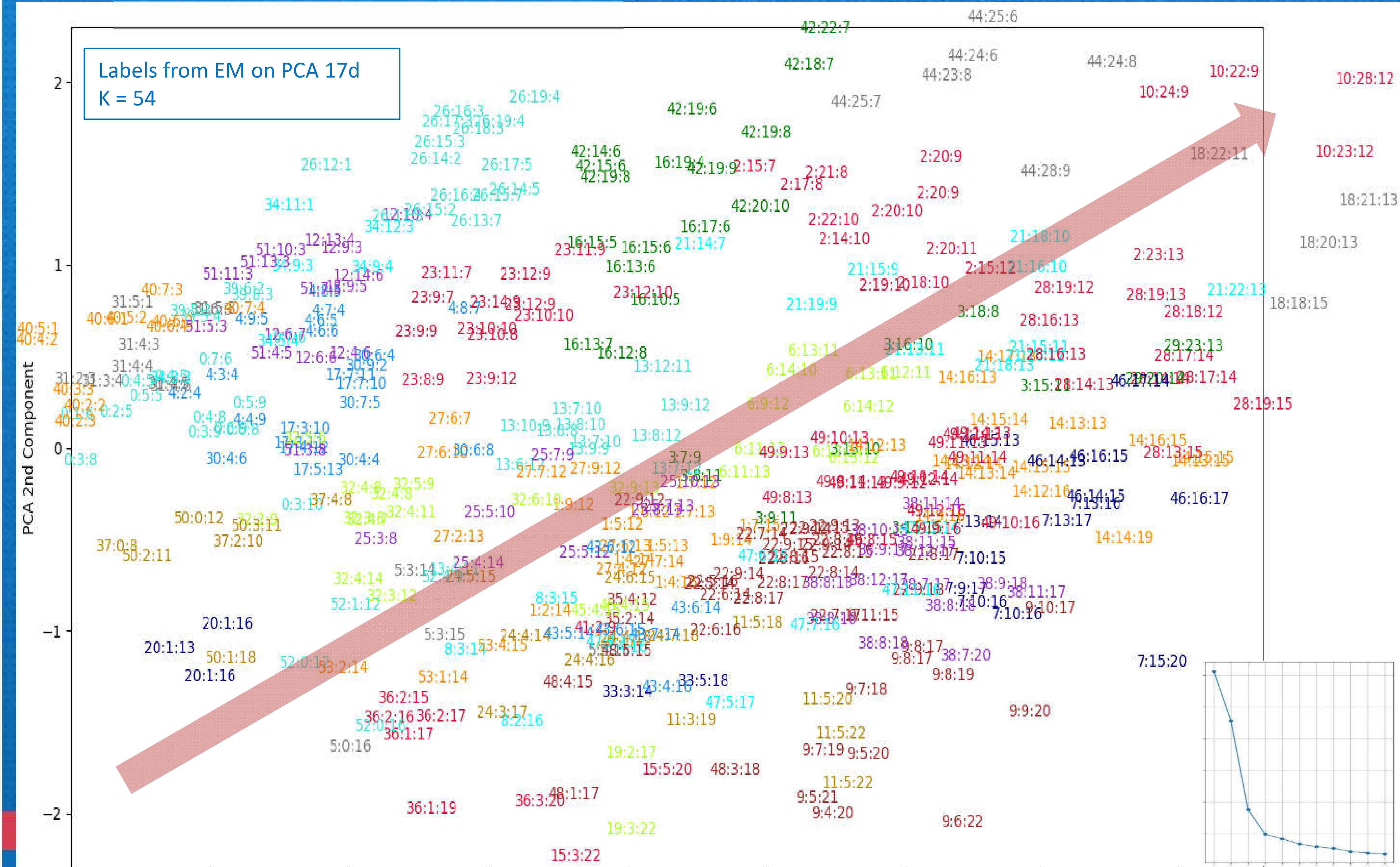
- Link v-measure to Quality of clustering
- Saving human work to check results

V-measure	Quality of Clusteirng Result
> 0.80	Excellent
0.75-0.80	Good
0.70-0.75	OK
0.65-0.70	Poor
< 0.65	Very Poor

Results

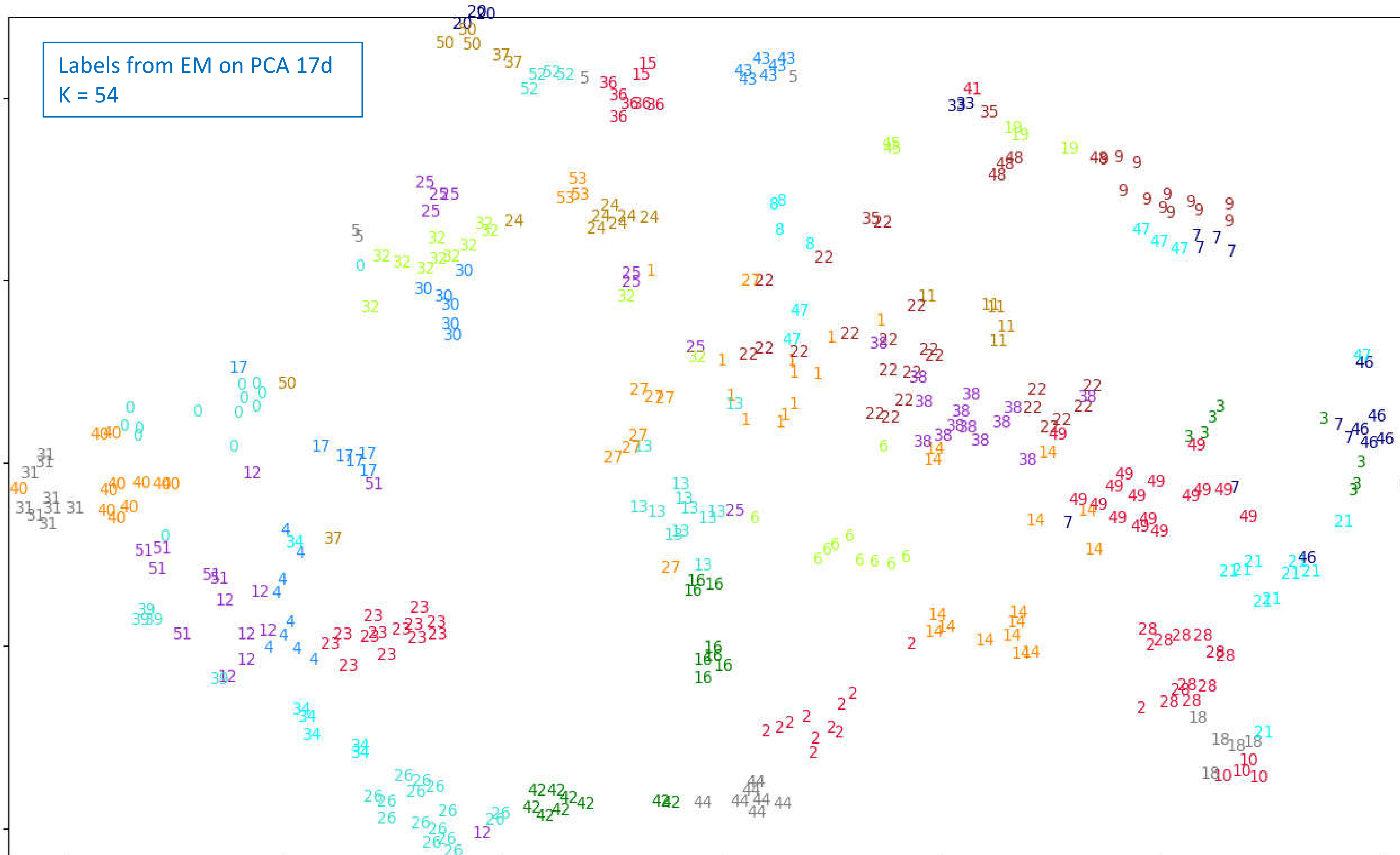
Labels	Efficiency	Avg_Shot_Dist
10		
Derrick Rose	23.11	12.51
Dwyane Wade	24.78	9.90
LeBron James	28.58	12.25
Russell Westbrook	22.38	9.11
18		
Carmelo Anthony	22.66	11.95
Kevin Durant	24.95	14.98
Kevin Martin	18.18	15.25
Kobe Bryant	21.39	13.96
Monta Ellis	20.29	13.69
29		
Chris Paul	23.10	13.71
Steve Nash	20.99	14.27
44		
Amar'e Stoudemire	24.59	8.36
Blake Griffin	25.63	6.72
Dwight Howard	28.31	4.44
Kevin Love	28.37	9.94
LaMarcus Aldridge	23.09	8.77
Pau Gasol	25.40	7.36
Zach Randolph	24.43	6.86

PCA 2D Plot (Label : Efficiency : Avg. Shot Dist.)



t-SNE

$(k \ll n^{1/5} \approx 3)$



Improvement on overlapping

- Build 3-NN graph
  - 54 centroids → 54 vertices
  - for each vertices:
    - find 3 nearest neighbors and add edges to them
- Each cluster will have 3 neighbor clusters
- Players from neighbors can be considered as secondary option