Identifying Player Prototypes in Basketball using Archetypoid Analysis

Daniel Fernández

Universitat Politècnica de Catalunya - BarcelonaTech, Spain

XI RETREAT GRBIO 2026
Vilanova i la Geltrú, July 1st, 2026

Team work

Marica Manisera **Paola Zuccolotto**

Big&Open Data Innovation Laboratory (BODaI-Lab)
University of Brescia, Italy

Daniel Fernández **Jordi Cortés**

Institute for Research and Innovation in Health (IRIS)
Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Spain

Outline

Motivation

Motivation example

NBA Player Statistics Season 2024-25

$n = 501$ **players**

$p = 17$ **variables**

Var	Description
MIN	**Total minutes played**
PTS	**Points per game**
REB	**Rebounds per game**
AST	Assists per game
FGA	Field goal attempts
FG%	Field goal efficiency
P3A	3-point attempts
P3%	3-point efficiency
FTA	Free throw attempts
FT%	Free throw efficiency
STL	Steals per game
BLK	Blocks per game
TOV	Turnovers per game
PF	Personal fouls

Head of the NBA dataset (some variables)

Player	Team	PTS	FG%	P3%	REB	AST	BLK	MIN
A.J. Lawson	TOR	22.5	42.1	32.7	8.2	3.0	0.6	486
Aaron Gordon	DEN	24.4	53.1	43.6	8.1	5.3	0.5	1447
Aaron Holiday	HOU	20.1	43.7	39.8	4.6	4.9	0.7	792
Aaron Nesmith	IND	22.8	50.7	43.1	7.5	2.3	0.7	1122
Aaron Wiggins	OKC	24.7	48.8	38.3	8.0	3.6	0.5	1744
Adam Flagler	OKC	15.4	26.0	19.4	6.4	2.8	0.7	203

Motivation example

NBA Player Statistics Season 2024-25

n = 501 players

p = 17 variables

Var	Description
MIN	**Total minutes played**
PTS	**Points per game**
REB	**Rebounds per game**
AST	Assists per game
FGA	Field goal attempts
FG%	Field goal efficiency
P3A	3-point attempts
P3%	3-point efficiency
FTA	Free throw attempts
FT%	Free throw efficiency
STL	Steals per game
BLK	Blocks per game
TOV	Turnovers per game
PF	Personal fouls

Head of the NBA dataset (some variables)

Player	Team	PTS	FG%	P3%	REB	AST	BLK	MIN
A.J. Lawson	TOR	22.5	42.1	32.7	8.2	3.0	0.6	486
Aaron Gordon	DEN	24.4	53.1	43.6	8.1	5.3	0.5	1447
Aaron Holiday	HOU	20.1	43.7	39.8	4.6	4.9	0.7	792
Aaron Nesmith	IND	22.8	50.7	43.1	7.5	2.3	0.7	1122
Aaron Wiggins	OKC	24.7	48.8	38.3	8.0	3.6	0.5	1744
Adam Flagler	OKC	15.4	26.0	19.4	6.4	2.8	0.7	203

- Can we found players who are **extreme** profiles based on their performance? Archetypoid analysis (**Standard ADA**).

Motivation example

NBA Player Statistics Season 2024-25

$n = 501$ players

$p = 17$ variables

Var	Description
MIN	**Total minutes played**
PTS	**Points per game**
REB	**Rebounds per game**
AST	Assists per game
FGA	Field goal attempts
FG%	Field goal efficiency
P3A	3-point attempts
P3%	3-point efficiency
FTA	Free throw attempts
FT%	Free throw efficiency
STL	Steals per game
BLK	Blocks per game
TOV	Turnovers per game
PF	Personal fouls

Head of the NBA dataset (some variables)

Player	Team	PTS	FG%	P3%	REB	AST	BLK	MIN
A.J. Lawson	TOR	22.5	42.1	32.7	8.2	3.0	0.6	486
Aaron Gordon	DEN	24.4	53.1	43.6	8.1	5.3	0.5	1447
Aaron Holiday	HOU	20.1	43.7	39.8	4.6	4.9	0.7	792
Aaron Nesmith	IND	22.8	50.7	43.1	7.5	2.3	0.7	1122
Aaron Wiggins	OKC	24.7	48.8	38.3	8.0	3.6	0.5	1744
Adam Flagler	OKC	15.4	26.0	19.4	6.4	2.8	0.7	203

- Can we found players who are **extreme** profiles based on their performance? Archetypoid analysis (**Standard ADA**).
- Are they players determined by ADA **representative**? e.g. do they have play enough minutes? Our proposal: **Penalized ADA**.

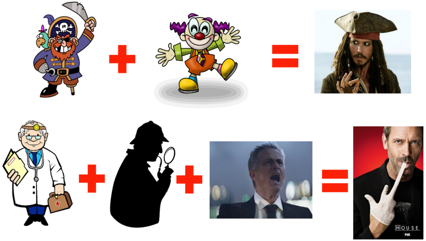
Background: AA & ADA

What are the Archetypes?

- Archetype (Wikipedia): from Greek:
  - *archē*: “beginning”, “origin”.
  - *tupos*: “pattern”, “model”, or “type”.
  - **Original pattern from which copies are made.**

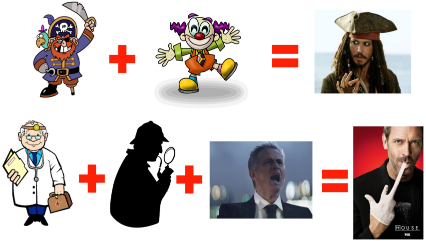
What are the Archetypes?

- Archetype (Wikipedia): from Greek:
  - *archē*: "beginning", "origin".
  - *tupos*: "pattern", "model", or "type".
  - **Original pattern from which copies are made.**
- Archetypes in everyday language:
  - Jack Sparrow: 40% pirate and 60% clown.
  - Dr. House: 20% doctor, 30% detective, and 50% bad temper.



What are the Archetypes?

- Archetype (Wikipedia): from Greek:
  - *archē*: “beginning”, “origin”.
  - *tupos*: “pattern”, “model”, or “type”.
  - **Original pattern from which copies are made.**
- Archetypes in everyday language:
  - Jack Sparrow: 40% pirate and 60% clown.
  - Dr. House: 20% doctor, 30% detective, and 50% bad temper.



- In Statistics, the **concept of archetypes** is the same as in common life.

Archetype Analysis (AA). Definition

- **Objective** (Cutler and Breiman, 1994): to find a few, not necessarily observed, extremal cases or pure types (the archetypes) such that:
  - all the observations are **obtained by convex combinations of the archetypes**, and
  - all the archetypes are convex combinations of the observations.

Archetype Analysis (AA). Definition

- **Objective** (Cutler and Breiman, 1994): to find a few, not necessarily observed, extremal cases or pure types (the archetypes) such that:
  - all the observations are **obtained by convex combinations of the archetypes**, and
  - all the archetypes are convex combinations of the observations.

- **Solution** (Cutler and Breiman, 1994): Alternating minimizing algorithm and implemented in R by Eugster and Leisch (2009).

Archetype Analysis (AA). Example

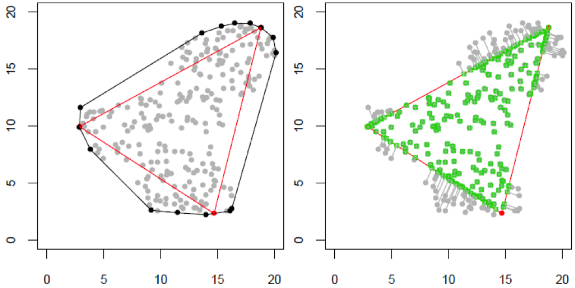


Figure 2: Visualization of three archetypes.

Suppose we want to find $k = 3$ archetypes:

1. Start with 3 random points.

Archetype Analysis (AA). Example

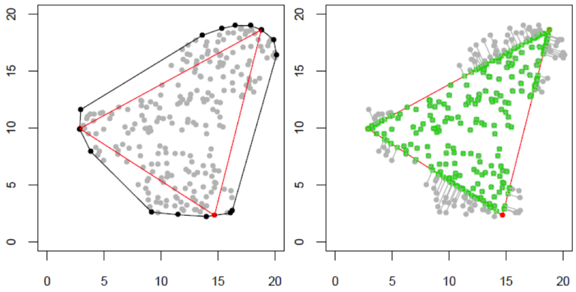


Figure 2: Visualization of three archetypes.

Suppose we want to find $k = 3$ archetypes:

1. Start with 3 random points.
2. The algorithm iteratively updates these points to identify archetypes.

Archetype Analysis (AA). Example

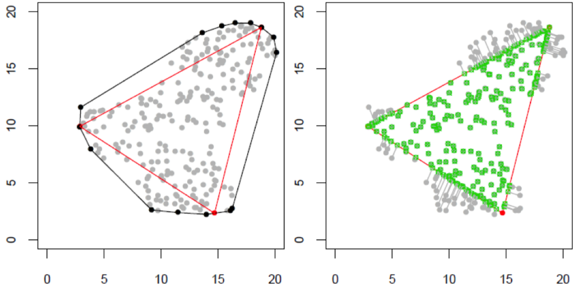


Figure 2: Visualization of three archetypes.

Suppose we want to find $k = 3$ archetypes:

1. Start with 3 random points.
2. The algorithm iteratively updates these points to identify archetypes.
3. In the end, they are the 3 corners of the data.

Archetype Analysis (AA). Example

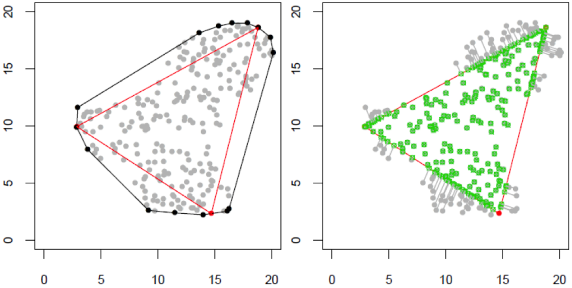


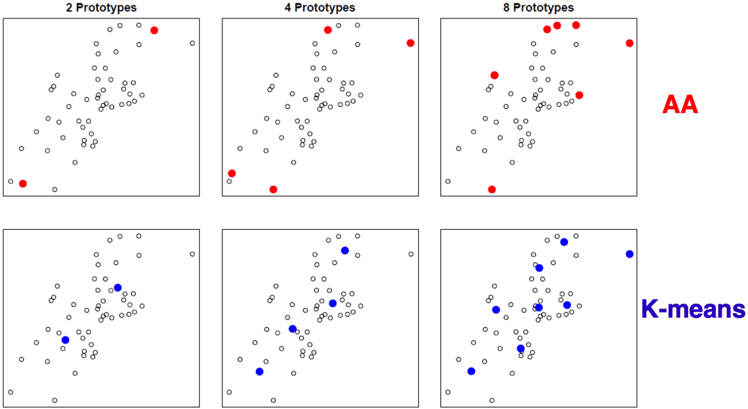
Figure 2: Visualization of three archetypes.

Suppose we want to find $k = 3$ archetypes:

1. Start with 3 random points.
2. The algorithm iteratively updates these points to identify archetypes.
3. In the end, they are the 3 corners of the data.
4. The **observations** are approximated by convex combinations of the archetypes.

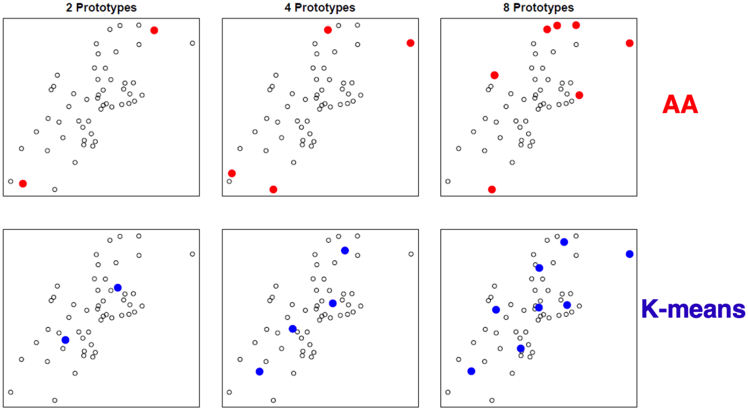
Archetype Analysis (AA). Example

Archetypes **are more interpretable** than the centroids of k-means.



Archetype Analysis (AA). Example

Archetypes **are more interpretable** than the centroids of k-means.



For humans, it is **easier to understand opposites than centroids**: they provide **an intuitive understanding** of the results (Vinúe *et al.*, 2015; Thurau *et al.*, 2012).

AA for multivariate data

- Let X be an $n \times m$ matrix with n observations and m variables.
- The objective of AA is to find the matrix Z of k m-dimensional archetypes.
- AA computes two matrices α and β which **minimize the residual sum of squares (RSS)**:

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2,$$

under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ for $j = 1, \ldots, k$.

Archetypoid Analysis (ADA)

- **A**rchetypoid **A**nalysis (ADA) is a **specific case** of Archetype Analysis (AA).

Archetypoid Analysis (ADA)

- **A**rchetypoid **A**nalysis (ADA) is a **specific case** of Archetype Analysis (AA).
- **Objective** (Vinúe, Epifanio, & Alemany, 2015): To find a few observed, **extremal cases or pure types (the archetypoids)** such that:
  - All observations are approximated by convex combinations of the archetypoids (**as before**), and
  - All archetypoids are **real observations**.

ADA for multivariate data

- The objective of ADA is to find the matrix Z of k m-dimensional archetypes.
- AA computes two matrices α and β which **minimize the residual sum of squares (RSS)**:

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2,$$

under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \in \{0, 1\}$ and $j = 1, \ldots, k.$

- **ADA solution** (Vinué, Epifanio, & Alemany, 2015): algorithm based on PAM.

Penalized ADA

Penalized ADA

- **Main idea**: Modify the **objective function**, explicitly controlling the selection via regularization (weights).

Penalized ADA

- **Main idea**: Modify the **objective function**, explicitly controlling the selection via regularization (weights).

$$RSS = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \mathbf{z}_j \right\|^2 = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{\ell=1}^{n} \beta_{j\ell} \mathbf{x}_\ell \right\|^2 + \lambda \sum_{\mathbf{j=1}}^{\mathbf{k}} \sum_{\ell=\mathbf{1}}^{\mathbf{n}} \beta_{\mathbf{j}\ell} \mathbf{w}_\ell,$$

Penalized ADA

- **Main idea**: Modify the **objective function**, explicitly controlling the selection via regularization (weights).

$$RSS = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \mathbf{z}_j \right\|^2 = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{\ell=1}^{n} \beta_{j\ell} \mathbf{x}_\ell \right\|^2 + \lambda \sum_{\mathbf{j=1}}^{\mathbf{k}} \sum_{\ell=\mathbf{1}}^{\mathbf{n}} \boldsymbol{\beta}_{\mathbf{j}\ell} \mathbf{w}_\ell,$$

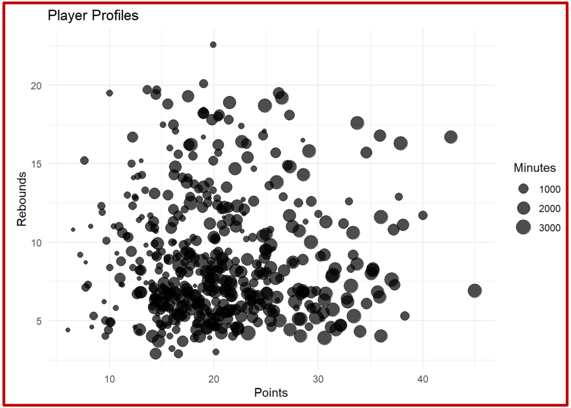
under the constraints

1. $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and $\sum_{\ell=1}^{n} \beta_{j\ell} = 1$ with $\beta_{j\ell} \in \{0, 1\}$ and $j = 1, \ldots, k$.
2. The **penalty weights** are defined as:
   - $w_\ell^0 = 1/(\mathrm{minutes}_\ell + \epsilon)$, where ϵ is a small constant to avoid division by zero,
   - $w_\ell = w_\ell^0 / \bar{w}^0$, where $\bar{w}^0$ is the mean of all the w_ℓ^0,
   - $\lambda \geq 0$ controls the strength of the penalty; $\lambda = 0$ represents no penalization.

- **Interpretation**: High-minute players $\rightarrow$ low contribution to the penalty

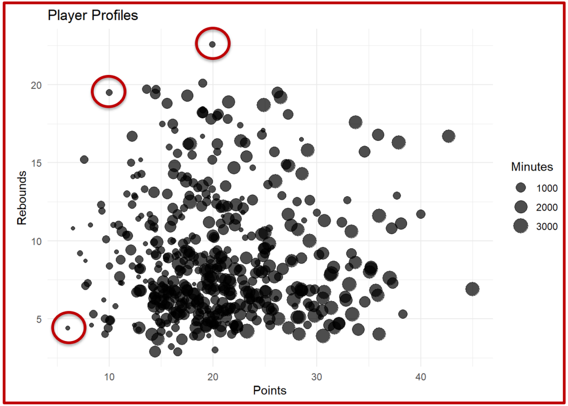
Application: NBA data set

ADA results (players)



- The variables **Points** & **Rebounds** are our features of interest.
- The variable **Minutes** serves to compute weights.

Standard ADA



- The variables **Points** & **Rebounds** are our features of interest.
- The variable **Minutes** serves to compute weights.
- Note that some players with very **low minutes** are located at the **extremes** of the data cloud $\Rightarrow$ Archetypoids in the **standard ADA**.

ADA results: Standard vs. Penalized (I)

- Example: $k = 3$ archetypoids, 20 starting points, & $\lambda = 1$ as the penalty strength.

	Standard	**Penalized**
RSS	0.027	0.159
Avg. minutes	988.97	2421.67
Computation time	27.95	41.11

- **RSS** is higher in the penalized version $\Rightarrow$ adding the penalty selects **less extreme** but more **representative** archetypoids.
- **Average minutes** nearly triples in the penalized version (989 $\rightarrow$ 2422 min) $\Rightarrow$ confirms the penalty favors **high-minute** players.

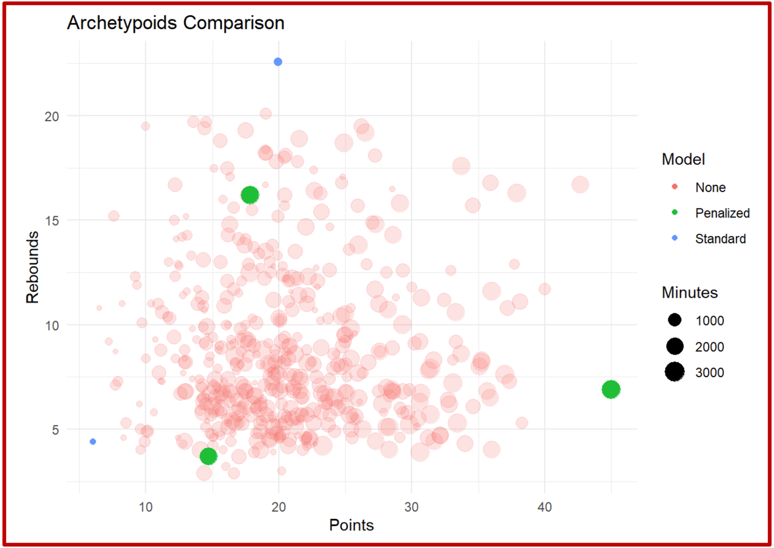
ADA results: Standard vs. Penalized (II)

Selected archetypoids

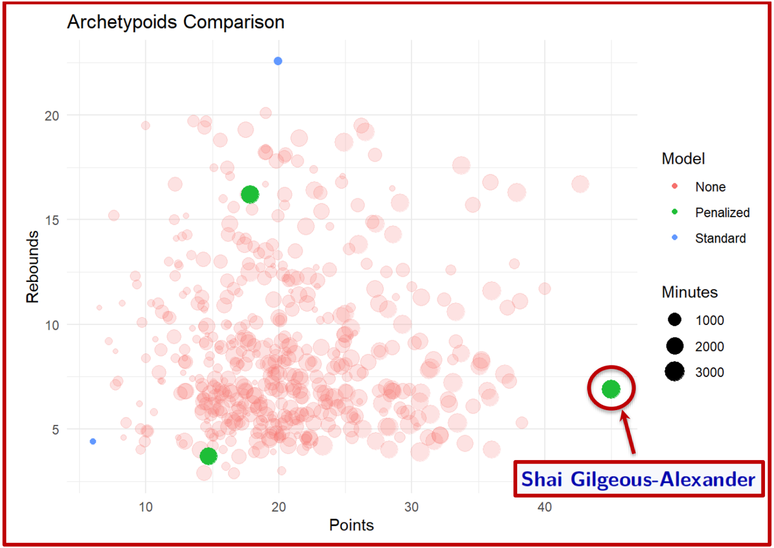
Player	PTS	REB	MIN	Model
Joe Ingles	6.0	4.4	114.1	Standard
Oscar Tshiebwe	19.9	22.6	255.2	Standard
Shai Gilgeous-Alexander	45.0	6.9	2597.6	**Both**
Rudy Gobert	17.8	16.2	2388.4	Penalized
Kentavious Caldwell-Pope	14.7	3.7	2279.0	Penalized

- Standard ADA picks players with **very low minutes** (Joe Ingles: 114 min, Oscar Tshiebwe: 255 min), which may be **misleading** in practice.
- **Shai Gilgeous-Alexander** is selected in **both** versions: he is simultaneously **extreme and representative**.
- $\lambda \geq 0$ controls the trade-off: higher λ favors **representativeness**, lower λ favors **extremeness**.

ADA results: Standard vs. Penalized (III)



ADA results: Standard vs. Penalized (III)



Penalized ADA results (players)

- Each NBA player is represented as a mixture of the 3 archetypoids.
  Example Cluster 3:

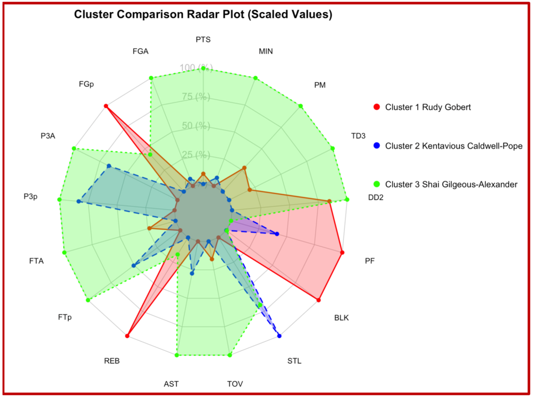
Cluster 3 (N=58; 11.6%)

Representative: Shai Gilgeous-Alexander

- Point guard - Shooting guard.
- 4-time NBA All-Star.
- 3-time All-NBA First Team member.
- 2-time NBA MVP.

Penalized ADA results (players)

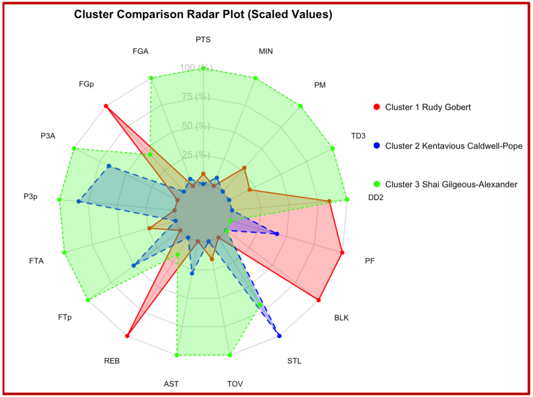
Cluster 3 (N=58; 11.6%). Representative: Shai Gilgeous-Alexander



- **Cluster 3** players are **elite, high-usage franchise stars who dominate offensively**.

Penalized ADA results (players)

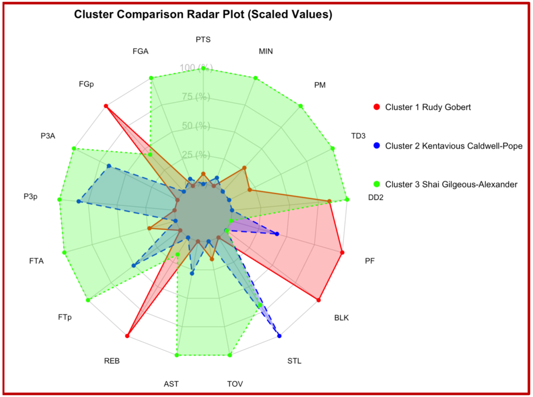
Cluster 3 (N=58; 11.6%). Representative: Shai Gilgeous-Alexander



- **Cluster 3** players are **elite, high-usage franchise stars who dominate offensively**.
- **Leading all clusters** in scoring (32.9 pts), assists (7.22), minutes (1,857), and free throw attempts (7.57).

Penalized ADA results (players)

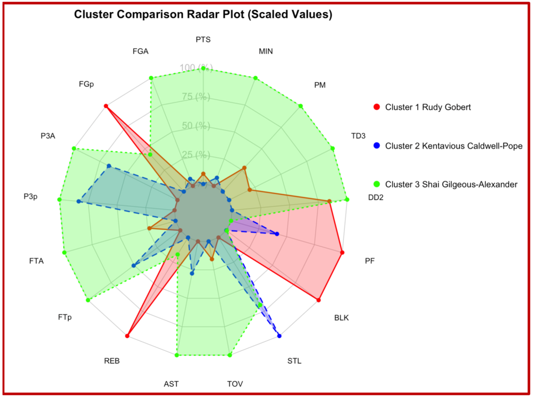
Cluster 3 (N=58; 11.6%). Representative: Shai Gilgeous-Alexander



- **Cluster 3** players are **elite, high-usage franchise stars who dominate offensively**.
- **Leading all clusters** in scoring (32.9 pts), assists (7.22), minutes (1,857), and free throw attempts (7.57).
- They are the only cluster with a positive plus/minus (+1.37), confirming their **winning impact**.

Penalized ADA results (players)

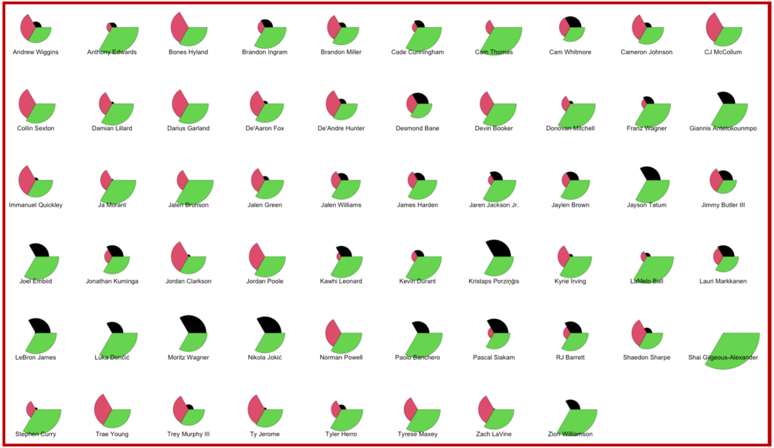
Cluster 3 (N=58; 11.6%). Representative: Shai Gilgeous-Alexander



- **Cluster 3** players are **elite, high-usage franchise stars who dominate offensively**.
- **Leading all clusters** in scoring (32.9 pts), assists (7.22), minutes (1,857), and free throw attempts (7.57).
- They are the only cluster with a positive plus/minus (+1.37), confirming their **winning impact**.
- Though their heavy playmaking responsibility comes at the cost of the **highest turnover rate** (3.77).

Penalized ADA results (players)

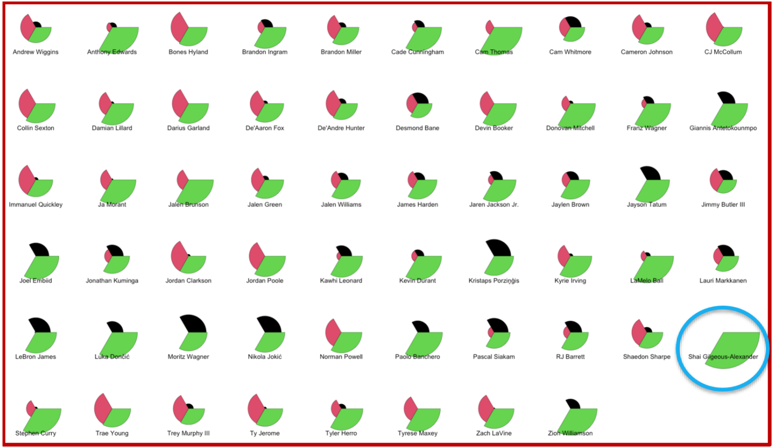
Shai Gilgeous-Alexander (green archetypoid)



Rudy Gobert, and Kentavious Caldwell-Pope

Penalized ADA results (players)

Shai Gilgeous-Alexander (green archetypoid)



Rudy Gobert, and Kentavious Caldwell-Pope

Penalized ADA results (players)

Shai Gilgeous-Alexander (green archetypoid)

Rudy Gobert, and Kentavious Caldwell-Pope

Summary

- We introduced a **penalized ADA** framework that incorporates player **representativeness** into the selection process.

Summary

- We introduced a **penalized ADA** framework that incorporates player **representativeness** into the selection process.
- Standard ADA may select **outliers** with few minutes played, leading to **misleading** prototypes.
- The penalized version tends to select players with **higher minutes**, improving **interpretability** and practical relevance.

Summary

- We introduced a **penalized ADA** framework that incorporates player **representativeness** into the selection process.
- Standard ADA may select **outliers** with few minutes played, leading to **misleading** prototypes.
- The penalized version tends to select players with **higher minutes**, improving **interpretability** and practical relevance.
- In **sports analytics**, where real-world relevance is essential, penalized ADA provides a more **actionable** solution than standard ADA.

Main References of this Talk

- Cutler, A., and Breiman, L. (1994). Archetypal analysis. *Technometrics*, 36(4), 338-347.
- Eugster, M., and Leisch, F. (2009). From spider-man to hero-archetypal analysis in R. *Journal of Statistical Software*, 30(8), 1–23.
- Fernández, D., Epifanio, I., and McMillan, L. F. (2021). Archetypal analysis for ordinal data. *Information Sciences*, 579, 281-292.
- Vinúe, G., Epifanio, I., and Alemany, S. (2015). Archetypoids: A new approach to define representative archetypal data. *Computational Statistics & Data Analysis*, 87, 102-115.
- Vinúe, G., and Epifanio, I. (2021). Robust archetypoids for anomaly detection in big functional data. *Advances in Data Analysis and Classification*, 15, 437-462.

The End

Acknowledgments:

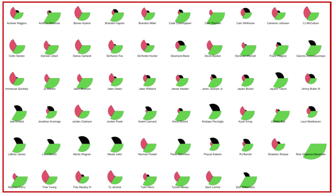
- MICIU & FEDER [PID2023-148033OB-C21] (Spain)
- Departament de Recerca i Universitats de la Generalitat de Catalunya [2021 SGR 01421] (Spain)
- NZ Government funding the Marsden Fund [E11772-5342] (New Zealand)

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

UNIÓN EUROPEA
FONDO EUROPEO DE DESARROLLO REGIONAL
"Una manera de hacer Europa"

AGENCIA ESTATAL DE INVESTIGACIÓN

Thanks for listening!!!



Influence of the Penalty (λ)

- For fixed $k = 3$, we compute the **average minutes** of the selected archetypoids across a range of λ values.

- As λ increases:
  1. **Avg. minutes** $\uparrow \rightarrow$ more representative players.
  2. **RSS** $\uparrow \rightarrow$ less extreme archetypoids.
  3. **Ratio** (min/RSS) $\uparrow \rightarrow$ better balance.

- Large λ favors **high-minute** players at the cost of a higher RSS.

Results for varying λ ($k = 3$)

λ	Avg. Min	RSS	Ratio
0.00	1238	8.69	142
0.10	1589	8.72	182
0.25	1589	8.72	182
0.50	1589	8.72	182
0.75	1589	8.72	182
1.00	1589	8.72	182
2.50	2063	8.92	231
5.00	2241	9.06	247
10.00	2603	9.92	262