Máster en Matemática Computacional

Proyecto Final de Máster

# Análisis de biarquetipos: definición y aplicación al análisis deportivo

*Autor:*
Aleix Alcacer Sales

*Tutora académica:*
Irene Epifanio López

Curso académico 2019/2020

## Resumen

En este trabajo se ha propuesto un nuevo concepto en estadística, el análisis de biarquetipos o biAA, cuyo propósito es obtener arquetipos de filas y columnas de una matriz de datos simultáneamente, es decir, extraer arquetipos de individuos y variables a la vez.

En primer lugar, se ha definido el análisis de biarquetipos matemáticamente y se ha propuesto un método numérico para resolverlo. A continuación, se ha implementado el biAA en el lenguaje R y se ha construido un paquete de R que recoge esta implementación. Además, se ha ilustrado este nuevo concepto con un ejemplo de juguete para que se entendiera y se viera cómo utilizar el paquete creado. Por último, se ha aplicado a un problema de análisis deportivo, donde el biAA ha permitido descubrir patrones escondidos en los datos. En concreto, se ha aplicado a un conjunto de datos que contiene métricas de los jugadores de la NBA.

## Palabras clave

Análisis de biarquetipos, Biclustering difuso, Análisis arquetípico, Análisis deportivo

## Keywords

Biarchetype analysis, Fuzzy biclustering, Archetypal analysis, Sports analytics

# Índice general

**1. Introducción** **7**

**2. Conceptos previos** **9**

2.1. Análisis de arquetipos . . . . . 9

2.2. Análisis clúster . . . . . 12

**3. Análisis de biarquetipos (biAA)** **15**

3.1. Definición . . . . . 16

3.2. Interpretación . . . . . 17

3.3. Procedimiento . . . . . 18

3.4. Elección del mejor modelo . . . . . 19

**4. Paquete de R:** `biaa` **21**

4.1. Descripción del paquete . . . . . 21

4.2. Ejemplo . . . . . 22

**5. Aplicación** **25**

5.1. Contexto y motivación . . . . . 25
5.2. Descripción de los datos . . . . . 26
5.3. Metodología . . . . . 28
5.4. Resultados . . . . . 28

**6. Conclusiones** **43**

**Referencias** **45**

**A. Notación matemática** **47**

# Capítulo 1

# Introducción

La minería de datos es un proceso que permite extraer patrones, previamente desconocidos, que residen de manera oculta en los datos. El objetivo principal consiste en extraer información de un conjunto de datos y convertirla en una estructura fácil de comprender.

La extracción manual de estos patrones de datos se ha llevado a cabo durante siglos. Uno de los primeros métodos desarrollados para identificar patrones es el teorema de Bayes (Bayes, 1763). Sin embargo, a medida que las capacidades técnicas de la informática han ido mejorando, la capacidad de recopilación, almacenamiento y manipulación de datos se ha incrementado drásticamente.

Un punto de inflexión en este crecimiento fue la creación de la *World Wide Web* en el CERN de Ginebra. Esto permitió compartir datos de forma rápida y sencilla con todo el mundo. Otro punto de inflexión ha sido la aparición del internet de las cosas, también llamado $IoT$. Esta tecnología permite la conexión digital de los objetos cotidianos (como pueden ser termostatos, lavadoras, relojes o coches) con internet.

Así, a medida que los conjuntos de datos han crecido en tamaño y complejidad, se han tenido que desarrollar nuevos métodos, como pueden ser las redes neuronales, el análisis clúster, el análisis arquetípico, los algoritmos genéticos, los SVM, los árboles de decisión o las reglas de decisión.

Como consecuencia de estos avances, en la actualidad se puede obtener desde la información del tráfico de las ciudades gracias a los sensores que incorporan los semáforos, a las preferencias musicales de los usuarios dependiendo de las canciones que más escuchas en Spotify.

Dicho esto, el objetivo de este trabajo final del Máster en Matemática Computacional de la

Universidad Jaume I, será proponer una nueva herramienta estadística que va a estar enmarcada dentro de la minería de datos. Este nuevo concepto se llamará análisis de biarquetipos o, simplemente, biAA, y su función principal será la de hallar los patrones extremos dado un conjunto de datos[1]. Para su obtención, usará, a diferencia del análisis de arquetipos, combinaciones convexas tanto de las observaciones como de las variables.

En primer lugar, en el capítulo 2 introduciremos dos conceptos previos que van a ser las bases del biAA: el análisis de arquetipos y el análisis clúster. Después, en el capítulo 3 definiremos, con todo detalle, el análisis de biarquetipos. También veremos cómo interpretar los resultados y detallaremos, paso a paso, el procedimiento que se ha desarrollado para su obtención.

Teniendo en cuenta que el análisis de biarquetipos se va a resolver mediante un método numérico, en el capítulo 4 analizaremos un paquete de R, llamado `biaa`, creado especialmente para poder aplicar el análisis de biarquetipos. Además, programaremos un ejemplo de juguete que muestre, de una forma sencilla, cómo se usa cada función de este paquete.

A continuación, en el capítulo 5 aplicaremos el análisis de biarquetipos a un problema del ámbito del análisis deportivo. En concreto, analizaremos mediante el biAA qué tipo de roles, posición por posición, tienen los jugadores de baloncesto en los partidos de la NBA.

En el capítulo 6 presentaremos las conclusiones que se obtienen de los anteriores capítulos. También analizaremos, en esta parte, los trabajos futuros que puedan surgir a partir de lo visto en los capítulos anteriores, especialmente a partir de la definición de análisis de biarquetipos

Finalmente, en el anexo A estará detallada toda la notación matemática empleada en el trabajo.

[1]En este trabajo, siempre que se haga referencia a un conjunto de datos, se estarán considerando datos cuya estructura sea una matriz con $n$ observaciones (o muestras, las filas de la matriz) y $m$ variables (o atributos, las columnas de la matriz).

# Capítulo 2

# Conceptos previos

En este primer capítulo vamos a definir dos conceptos básicos que son claves para poder entender, de una forma más sencilla, la definición de análisis de biarquetipos propuesta en el capítulo 3.

El primero de ellos, descrito en la sección 2.1, es el análisis de arquetipos. En concreto, nos centraremos en exponer su definición matemática, puesto que ésta será el punto de partida de nuestra definición de análisis de biarquetipos.

El otro concepto necesario, que veremos en la sección 2.2, es el análisis clúster. En particular, será interesante conocer las ideas básicas del biclustering difuso, un tipo de clustering que hace grupos tanto de las observaciones como de las variables y que trabaja con grados de pertenencia a dichos grupos.

## 2.1. Análisis de arquetipos

El análisis de arquetipos, también llamado AA, fue definido por primera vez en Cutler y Breiman (1994). Es un método estadístico no supervisado, cuya finalidad es buscar los casos extremos, llamados arquetipos, en un conjunto de datos. Tal y como se indica en Davis y Love (2010), el disponer de estos patrones extremos facilita, gracias al principio de los opuestos (Thurau y cols., 2012), la interpretación de los resultados.

En el análisis de arquetipos, cada arquetipo se define como una combinación convexa de las observaciones de la muestra y, a su vez, cada observación se aproxima por una combinación convexa de los arquetipos.

### Definición

Sea $X$ una matriz de $n$ observaciones y $m$ variables. El objetivo del análisis de arquetipos es obtener una matriz $Z$ formada por $k$ arquetipos y $m$ variables. Para conseguir estos arquetipos, según se expone en Cabero y Epifanio (2019), el AA calcula dos matrices $\alpha$ y $\beta$ que minimicen la suma residual

$$\begin{aligned} RSS = & \sum_{i=1}^{n}\sum_{j=1}^{m} \|x_{ij} - \hat{x}_{ij}\|^2 = \\ & \sum_{i=1}^{n}\sum_{j=1}^{m} \left\| x_{ij} - \sum_{g=1}^{k} \alpha_{ig} z_{gj} \right\|^2 = \\ & \sum_{i=1}^{n}\sum_{j=1}^{m} \left\| x_{ij} - \sum_{g=1}^{k} \alpha_{ig} \left( \sum_{l=1}^{n} \beta_{gl} x_{lj} \right) \right\|^2 \end{aligned} \tag{2.1}$$

sujeta a las siguientes restricciones

1. $\sum_{j=1}^{k} \alpha_{ij} = 1$ con $\alpha_{ij} \geq 0$ para $i = 1, \ldots, n$
2. $\sum_{l=1}^{n} \beta_{jl} = 1$ con $\beta_{jl} \geq 0$ para $j = 1, \ldots, k$

#### Representación matricial

En vistas a la definición anterior, y una vez conocidos todos los elementos involucrados, los arquetipos se pueden definir de forma matricial como

$$Z_{k\times m} = \alpha_{k\times n} X_{n\times m}$$

y, de manera semejante, una aproximación de la matriz de datos se puede definir como

$$X_{n\times m} \simeq \beta_{n\times k} Z_{k\times m}$$

### Interpretación de las matrices $\alpha$ y $\beta$

Como resultado de la ecuación 2.1 y de las restricciones 1 y 2, se deduce que los pesos que forman la matriz $\alpha$ son los que se utilizan para describir todo el conjunto de datos en función de los arquetipos.

De igual manera, se puede deducir que los pesos de la matriz $\beta$ son los que permiten describir los arquetipos en función de los datos. Además, estos pesos muestran que los arquetipos no tienen por qué ser observaciones del conjunto de datos.

### Elección del mejor modelo

Hay situaciones en las que no se sabe cuántos arquetipos hay que calcular. En estos casos puede ser de utilidad usar el método del codo. Este método sigue la misma filosofía que se emplea para determinar el número de clústers en análisis clúster (Hastie y cols., 2009).

En concreto, se representa en una gráfica la RSS obtenida respecto al número de arquetipos con los que se ha calculado. La idea es que la RSS irá decreciendo conforme vaya aumentando el número de arquetipos, porque el modelo podrá explicar mejor los datos con un mayor número de arquetipos. Pero, cuando se llegue al número óptimo de arquetipos, la mejora (la disminución en RSS) que se obtenga al añadir un nuevo arquetipo no será tan importante como en los pasos anteriores (con un número de arquetipos menor) y, en consecuencia, se podría apreciar un codo en la gráfica, que justo indicaría el número óptimo de arquetipos.

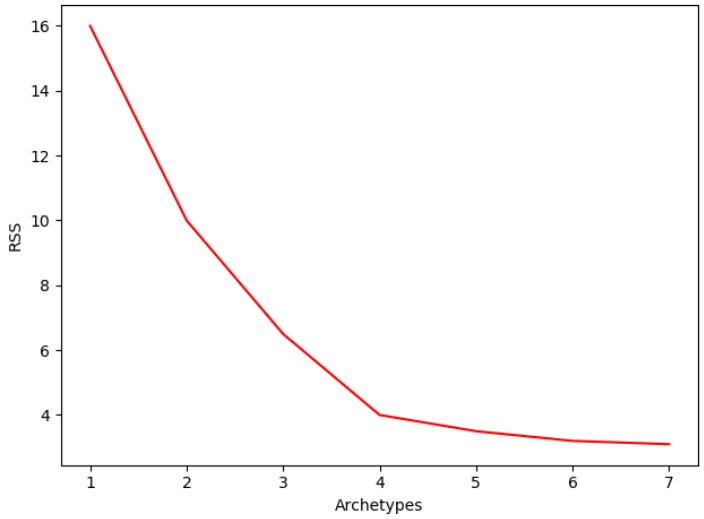



Figura 2.1: RSS para diferentes números de arquetipos.

Por ejemplo, en la figura 2.1 hay representados unas RSS que se han obtenido al aplicar el análisis de arquetipos con diversos números de arquetipos. En ella se observa que hay un codo cuando hay cuatro arquetipos. Por lo tanto, en este ejemplo el número óptimo de arquetipos con el que realizar el análisis de los datos sería cuatro.

## 2.2. Análisis clúster

El análisis clúster es, al igual que el análisis de arquetipos, otro análisis estadístico no supervisado utilizado en la minería de datos.

### Clustering

Los algoritmos utilizados en el clustering tienen como finalidad encontrar grupos de observaciones, llamados clusters, en función de los valores de todas las variables del conjunto de datos. En general, dos observaciones se asignarán al mismo clúster si los valores de éstas son similares entre ellos de acuerdo a una determinada métrica.

Así, por ejemplo, en el algoritmo $k$-means (MacQueen, 1967), para medir el nivel de similitud entre los objetos, se utiliza la distancia Euclídea. Es decir, si dos objetos están muy cerca usando la distancia Euclídea, entonces se asignarán al mismo clúster.

Sea $X_{n\times m}$ una matriz de datos con $n$ observaciones y $m$ variables. El objetivo del $k$-means es encontrar una matriz $C_{k\times m}$ con $k$ centroides y $m$ variables. Para ello, se debe encontrar una matriz $U_{n\times k}$ (donde cada $u_{ig}$ indica si la observación $i$ pertenece al clúster $g$ o no) minimizando la suma residual

$$\text{mín} \left\| X - UC \right\|^2 \tag{2.2}$$

sujeta a la siguiente restricción

1. $\sum_{g=1}^{k} u_{ig} = 1$ con $u_{ig} \in \{0, 1\}$ para $i = 1, \ldots, n$.

### Clustering difuso

Un caso especial de clustering es el clustering difuso, que está basado en la teoría de los conjuntos difusos (Zadeh, 1965).

Como se indica en Bezdek y cols. (1984), el clustering difuso no asigna cada elemento a un único clúster, sino que representa el nivel de similitud entre cada elemento y los clusters con valores entre 0 y 1. De este modo, a cada observación se asignará, por cada clúster, un grado de similitud. Así, si el nivel de similitud es próximo a 1, significa que el elemento y el clúster son muy similares entre sí y si, por el contrario, el valor es próximo a 0, indica que el nivel de similitud entre el clúster y el elemento es muy bajo.

Como consecuencia de esto, es fácil ver que el clustering básico es un caso particular del clustering difuso. En concreto, se obtiene cuando los valores de similitud solo pueden ser 0 o 1. De esta forma, el clúster más similar a la observación tiene un grado de similitud 1 (clúster al que pertenece la observación) y el resto, 0.

Para la obtención de los clusters, en el caso del $k$-means difuso (extensión del $k$-means al clustering difuso) se tiene que minimizar la misma suma residual que en el $k$-means

$$\text{mín} \, \|X - UC\|^2 \tag{2.3}$$

pero sujeta a una restricción más general

1. $\sum_{j=1}^{k} u_{ij} = 1$ con $u_{ij} \geq 0$ para $i = 1, \ldots, n$.

## Biclustering

No obstante, hay situaciones en las que es igual de interesante encontrar grupos tanto de observaciones como grupos de variables. Es en en estos casos donde se aplica el biclustering.

Tal y como se define en Cheng y Church (2000), el biclustering, también llamado *co-clustering* o *two-mode clustering*, consiste en hacer, simultáneamente, grupos de observaciones y grupos de variables en un conjunto de datos. Es decir, cada observación se asignará a un clúster de observaciones (como en el clustering básico) pero cada variable también se asignará a un clúster de variables.

Este tipo de análisis es muy utilizado en el análisis de datos de expresión genética (Kerr y cols., 2008). Sin embargo, también es utilizado en otros campos como en la psicología (Van Mechelen y cols., 2004), en el análisis del tráfico web (Koutsonikola y Vakali, 2009) o en el desarrollo de sistemas de recomendación (Forsati y cols., 2013).

Siguiendo el ejemplo del $k$-means y de acuerdo con Maurizio (2001), el objetivo del $k$-means doble (extensión del $k$-means al biclustering) es encontrar una matriz $C_{k \times c}$ con $k$ centroides y $c$ variables. Para ello, se debe encontrar una matriz $U_{n \times k}$ (donde cada $u_{ig}$ indica si la observación $i$ pertenece al clúster de observaciones $g$ o no) y una matriz $V_{c \times m}$ (donde cada $v_{hj}$ indica si la variable $j$ pertenece al clúster de variables $h$ o no) minimizando la suma residual

$$\text{mín} \, \|X - UCV\|^2 \tag{2.4}$$

sujeta a las restricciones

1. $\sum_{g=1}^{k} u_{ig} = 1$ con $u_{ig} \in \{0, 1\}$ para $i = 1, \ldots, n$.
2. $\sum_{h=1}^{c} v_{hj} = 1$ con $v_{hj} \in \{0, 1\}$ para $j = 1, \ldots, m$.

### Biclustering difuso

Al igual que ocurre con el clustering difuso, si se aplica la teoría de conjuntos difusos (Zadeh, 1965) al biclustering, se consigue el biclustering difuso. Este tipo de biclustering asigna a cada observación un grado de similitud a cada clúster de observaciones y, a su vez, asigna a cada variable del conjunto de datos un grado de pertenencia a cada clúster de variables.

Si continuamos con Maurizio (2001), el $k$-means difuso doble (extensión del $k$-means al biclustering difuso) consiste en minimizar, al igual que en el biclustering, la suma residual

$$\text{mín} \, \|X - UCV\|^2 \tag{2.5}$$

pero sujeta a unas restricciones más flexibles

1. $\sum_{g=1}^{k} u_{ig} = 1$ con $u_{ig} \geq 0$ para $i = 1, \ldots, n$.
2. $\sum_{h=1}^{c} v_{hj} = 1$ con $v_{hj} \geq 0$ para $j = 1, \ldots, m$.

# Capítulo 3

# Análisis de biarquetipos (biAA)

En esta sección vamos a proponer un nuevo concepto en estadística que llamaremos análisis de biarquetipos, en inglés, *biarchetype analysis* (biAA), o también *co-archetype analysis* o *two-mode archetype analysis*. La idea es extender el análisis de arquetipos, que sólo contempla la extracción de arquetipos de filas (individuos), a obtener arquetipos de filas y columnas de la matriz de datos simultáneamente, es decir, a extraer arquetipos de individuos y variables a la vez. Los llamaremos biarquetipos.

El propósito de biAA es el mismo que el contemplado por el análisis bicluster, comentado en el capítulo previo, pero en vez de tener los centroides o puntos centrales de los grupos y los clusters (las asignaciones a cada grupo), lo que tendríamos con biAA son arquetipos, es decir, puntos extremos, tanto de individuos como de variables. Además, conservaremos las ventajas del análisis clásico de arquetipos, pues los arquetipos serán fácilmente interpretables al ser combinaciones convexas (mixturas) de individuos de la muestra y de variables del conjunto de datos. Serán también fácilmente interpretables porque los extremos son más entendibles para las personas, que no los puntos centrales (Davis y Love, 2010). Y por último, tanto los individuos de la muestra como las variables podrán ser explicadas como una combinación convexa, una mezcla, de los biarquetipos. Esto, permitiría establecer grupos con biAA, aunque no fuera su objetivo.

Las posibles aplicaciones de biAA, como técnica estadística no supervisada, serán las mismas que el biclustering, cuando este es aplicado a segmentar un conjunto de datos en los que no hay grupos separados entre sí, cosa que sucede a menudo en la práctica. Las aplicaciones pueden ir desde el campo de la genética, al campo de la minería de texto pasando por el campo del filtrado colaborativo en los sistemas de recomendación (Zhao y cols., 2012).

El concepto de biarquetipo es completamente original, no se ha planteado hasta el momento.

Por ello, en este trabajo, daremos su definición matemática formal y propondremos un método para resolver el análisis de biarquetipos.

## 3.1. Definición

Sea $X$ una matriz de $n$ observaciones y $m$ variables. El objetivo del análisis de biarquetipos es obtener una matriz $Z$ formada por $k$ arquetipos y $c$ variables. Para conseguir estos arquetipos, el biAA calcula cuatro matrices $\alpha$, $\gamma$, $\beta$ y $\theta$ que minimicen la suma residual

$$\begin{aligned} RSS = & \sum_{i=1}^{n}\sum_{j=1}^{m} \|x_{ij} - \hat{x}_{ij}\|^2 = \\ & \sum_{i=1}^{n}\sum_{j=1}^{m} \left\| x_{ij} - \sum_{g=1}^{k}\sum_{h=1}^{c} \alpha_{ig} z_{gh} \gamma_{hj} \right\|^2 = \\ & \sum_{i=1}^{n}\sum_{j=1}^{m} \left\| x_{ij} - \sum_{g=1}^{k}\sum_{h=1}^{c} \alpha_{ig} \left( \sum_{l=1}^{n}\sum_{r=1}^{m} \beta_{gl} x_{lr} \theta_{rh} \right) \gamma_{hj} \right\|^2 \end{aligned} \tag{3.1}$$

sujeta a las siguientes restricciones

1. $\sum_{g=1}^{k} \alpha_{ig} = 1$ con $\alpha_{ig} \geq 0$ para $i = 1, \ldots, n$.
2. $\sum_{h=1}^{c} \gamma_{hj} = 1$ con $\gamma_{hj} \geq 0$ para $j = 1, \ldots, m$.
3. $\sum_{l=1}^{n} \beta_{gl} = 1$ con $\beta_{ig} \geq 0$ para $g = 1, \ldots, k$.
4. $\sum_{r=1}^{m} \theta_{rh} = 1$ con $\theta_{jl} \geq 0$ para $h = 1, \ldots, c$.

### Representación matricial

Considerando la definición anterior, y al igual que sucede en el análisis de arquetipos, la matriz de datos viene aproximada por la siguiente combinación convexa de los arquetipos

$$X_{n\times m} \simeq \alpha_{n\times k} Z_{k\times c} \gamma_{c\times m}$$

y, de igual modo, la matriz de arquetipos viene definida por la siguiente combinación convexa de los datos

$$Z_{k\times c} = \beta_{k\times n} X_{n\times m} \theta_{m\times c}$$

## 3.2. Interpretación de las matrices $\alpha$, $\gamma$, $\beta$ y $\theta$

Teniendo en cuenta la ecuación 3.1 junto con las restricciones 1, 2, 3 y 4, se deducen las siguientes interpretaciones de los resultados:

Por una parte, la matriz $\alpha$ indica el grado de similitud de cada observación de la matriz de datos con cada observación arquetípica. De la misma forma, la matriz $\gamma$ indica el grado de similitud de cada variable de la muestra con cada variable arquetípica.

Por otra parte, la matriz $\beta$ indica el grado de similitud de cada observación arquetípica con cada observación inicial. Finalmente, la matriz $\theta$ indica el grado de similitud de cada variable arquetípica con cada variable del conjunto de datos.

### Ejemplo

Para comprender mejor cómo interpretar estas matrices veamos un ejemplo.

$$\text{Sea } \alpha = \begin{pmatrix} 0.1 & 0.85 & 0.05 \\ 0.35 & 0.35 & 0.3 \\ \vdots & \vdots & \vdots \end{pmatrix}_{5\times 3}, \; Z = \begin{pmatrix} -0.12 & 0 \\ 1 & 0.1 \\ -0.2 & 1.5 \end{pmatrix}_{3\times 2} \text{ y } \gamma = \begin{pmatrix} 0.99 & 0.34 & \cdots \\ 0.01 & 0.66 & \cdots \end{pmatrix}_{2\times 8}.$$

#### Matriz $\alpha$

Respecto a la primera observación de la muestra (primera fila de $\alpha$) podemos decir que tiene una fuerte influencia del segundo arquetipo, ya que el peso asociado a él es 0.85. Sin embargo, si nos centramos en la segunda observación, no podemos decir que ésta tenga una influencia un poco mayor o un poco menor de algún arquetipo, puesto que los tres pesos asociados a éstos, 0.35, 0.35 y 0.3, son muy similares.

#### Matriz $\gamma$

Si analizamos la primera variable de la muestra (la primera columna de la matriz $\gamma$) vemos que está influenciada, prácticamente en su totalidad, por la primera variable arquetípica. Esto es debido a que el peso que acompaña a esta variable arquetípica es 0.99. Por lo que se refiere a la segunda variable, esta tiene un poco más de influencia de la segunda variable (peso asociado 0.66), pero en este caso también está influenciada, en menor medida, por la primera variable

arquetípica.

Como resultado, a partir de los valores de las matrices $\alpha$ y $\gamma$, hemos podido analizar las observaciones de la muestra a partir de los arquetipos obtenidos, es decir, hemos podido describir todos los datos en función de sus patrones extremos. Además, también hemos podido ver las variables en función de las variables extremas. De forma análoga a este procedimiento, si lo que se pretende es analizar los arquetipos, éstos pueden describirse en función de los datos usando las matrices $\beta$ y $\theta$.

## 3.3. Procedimiento

Para resolver el análisis de biarquetipos, se propone un método iterativo que está basado en un algoritmo de minimización alterna. En este caso, los pasos para obtener los arquetipos son los siguientes:

1. Preparación de los datos: Inicializar, de forma aleatoria, las matrices $\alpha$, $\gamma$, $\beta$ y $\theta$ (cumpliendo las restricciones definidas anteriormente).

2. Repetir hasta hacer la $RSS$ lo suficientemente pequeña:

   *a*) Encontrar el mejor $\alpha$ (fijado $\gamma$): Resolver $n$ problemas de mínimos cuadrados convexos usando $X^T = (Z\gamma)^T\alpha^T$.

   *b*) Encontrar el mejor $\gamma$ (fijado $\alpha$): Resolver $m$ problemas de mínimos cuadrados convexos usando $X = (\alpha Z)\gamma$.

   *c*) Recalcular los arquetipos: $Z = \alpha^+ X\gamma^+$.

   *d*) Encontrar el mejor $\beta$ (fijado $\theta$): Resolver $k$ problemas de mínimos cuadrados convexos usando $Z^T = (X\theta)^T\beta^T$.

   *e*) Encontrar el mejor $\theta$ (fijado $\beta$): Resolver $c$ problemas de mínimos cuadrados convexos usando $Z = (\beta X)\theta$.

   *f*) Recalcular los arquetipos: $Z = \beta X\theta$.

   *g*) Calcular el nuevo $RSS$.

Cabe destacar que los problemas de mínimos cuadrados convexos pueden resolverse como se propone en Cutler y Breiman (1994), es decir, usando un problema de mínimos cuadrados penalizado (Lawson y Hanson, 1974). La idea es, dado un problema de mínimos cuadrados $A_{n\times k}X_{k\times m} = B_{n\times m}$, añadir una fila de elementos constantes $c$ a $A$ y una fila de elementos $c$ a

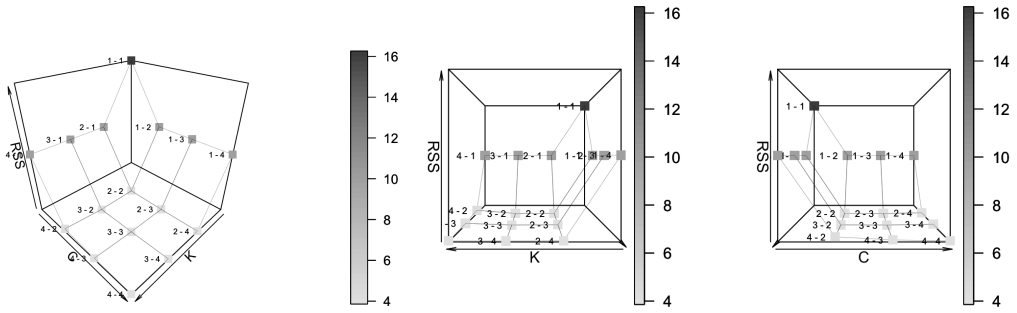
$B$ para obtener un nuevo problema $A_{n+1\times k}X_{k\times m} = B_{n+1\times m}$ cuyo RSS puede expandirse como

$$\begin{aligned} RSS = &\sum_{i=1}^{n+1}\sum_{j=1}^{m}\left\| b'_{ij} - \sum_{h=1}^{k} a'_{ih}x_{hj}\right\|^2 = \\ &\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\left\| b_{ij} - \sum_{h=1}^{k} a_{ih}x_{hj}\right\|^2 + \left\| b'_{n+1,j} - \sum_{h=1}^{k} a'_{n+1,h}x_{hj}\right\|^2\right) = \\ &\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\left\| b_{ij} - \sum_{h=1}^{k} a_{ih}x_{hj}\right\|^2 + \sum_{j=1}^{m}\left\| c - \sum_{h=1}^{k} cx_{hj}\right\|^2\right) = \\ &\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\left\| b_{ij} - \sum_{h=1}^{k} a_{ih}x_{hj}\right\|^2 + \sum_{j=1}^{m} c^2\left\| 1 - \sum_{h=1}^{k} x_{hj}\right\|^2\right) \end{aligned} \tag{3.2}$$

De esta forma, como se observa al final de la ecuación 3.2, si el valor $c$ es muy grande, el término $c^2\left\|1-\sum_{h=1}^{k} x_{hj}\right\|^2$ fuerza que los coeficientes de $X$ sean convexos.

## 3.4. Elección del mejor modelo

En el análisis de biarquetipos, la elección del mejor modelo se realiza de forma similar a cómo se realiza en el análisis de arquetipos.




(a) Vista general. (b) Vista de los perfiles.

Figura 3.1: Representación de las RSS en el análisis de biarquetipos.

En este caso, en lugar de representar la RSS respecto al número de arquetipos, se representa la RSS respecto al número de observaciones arquetípicas y respecto al número de variables arquetípicas. Por ello, la gráfica a analizar será una representación 3D en la que los dos ejes

horizontales estarán formados por el producto cartesiano entre los valores de $k$ y $c$ y en el eje vertical se representará, para cada elemento del producto, la RSS obtenida en el biAA.

La finalidad de este método es encontrar un punto $(k, c)$ tal que la RSS de los puntos siguientes $(k+1, c)$, $(k, c+1)$ y $(k+1, c+1)$ deje de disminuir de forma significativa respecto la RSS del punto $(k, c)$.

Por ejemplo, en la figura 3.1 están representadas las RSS obtenidas en un análisis de biarquetipos con unos valores de $k = (1, 2, 3, 4)$ y $c = (1, 2, 3, 4)$. Si nos centramos en el punto $(2, 2)$, se puede observar que la diferencia entre la RSS de los puntos $(3, 2)$, $(2, 3)$ y $(3, 3)$ y la de este punto es prácticamente idéntica. Por lo tanto, en este ejemplo, el modelo óptimo se consigue cuando se seleccionan dos observaciones arquetípicas y dos variables arquetípicas.

# Capítulo 4

# Paquete de R: `biaa`

En este capítulo vamos a analizar el paquete de R `biaa`, cuya documentación se encuentra en `https://aleix11alcacer.github.io/biaa/index.html`. Este paquete ha sido creado específicamente para poder calcular de forma computacional el análisis de biarquetipos propuesto en el capítulo 3.

Aunque está aún en versión *alpha*, contiene todas las funciones básicas para poder aplicar el análisis de biarquetipos a cualquier conjunto de datos multivariante y poder analizar los resultados de una forma satisfactoria.

## 4.1. Descripción del paquete

Por ahora, el paquete simplemente está compuesto por dos clases y tres funciones:

### Clases

- `biaa`: Representa un modelo concreto del análisis de biarquetipos. Contiene los valores de la $k$ (número de observaciones arquetípicas) y la $c$ (número de variables arquetípicas) escogidos y la RSS obtenida en este modelo. También contiene la matriz de los biarquetipos y las matrices $\alpha$, $\gamma$, $\beta$ y $\theta$.
- `biaaGroup`: Representa a un conjunto de modelos de la clase `biaa`. Tiene dos atributos que indican todos los valores de $k$ y $c$ usados en los modelos almacenados.

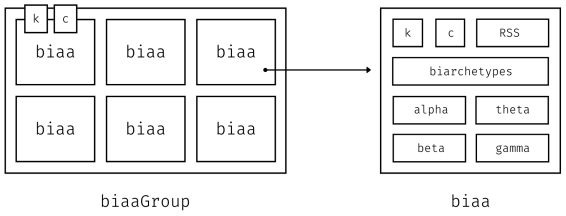



Figura 4.1: Representación esquemática de las clases.

### Funciones

- `biaa`: Aplica el análisis de biarquetipos a un conjunto de datos multivariante con el fin de obtener los biarquetipos. Es necesario pasarle los datos y los valores de $k$ y $c$ que se desean calcular. Se le puede especificar el número de veces que se repite cada modelo, el número máximo de iteraciones en cada modelo y el incremento mínimo entre la RSS de dos iteraciones para que el algoritmo no se pare. Devuelve un objeto de la clase `biaaGroup` que contiene todos los modelos calculados.

- `rssplot`: Dado un objeto de la clase `biaaGroup`, representa en 3 dimensiones las RSS obtenidas en los modelos calculados. Las tres dimensiones son los valores de $k$, los valores de $c$ y los valores de las RSS. Tiene dos parámetros opcionales que permiten rotar el gráfico en cualquier dirección para poder analizar mejor los resultados.

- `getmodel`: Dado un valor de $k$, un valor de $c$ y un objeto de la clase `biaaGropup`, devuelve un objeto de la clase `biaa` que contiene el modelo calculado con esos parámetros.

Más información acerca de estas funciones se puede obtener en `https://aleix11alcacer.github.io/biaa/reference/index.html`.

## 4.2. Ejemplo

En esta sección veremos un ejemplo de juguete que permita entender mejor como se puede aplicar, usando las funciones del paquete `biaa`, el análisis de biarquetipos a un conjunto de datos.

El código en R de este ejemplo se puede encontrar en `https://aleix11alcacer.github.io/biaa/articles/biaa.html`.

### Creación de los datos

En primer lugar, tal y como se observa en la figura 4.2, se genera una matriz de datos que esté formada por dos grupos de observaciones y dos grupos de variables.

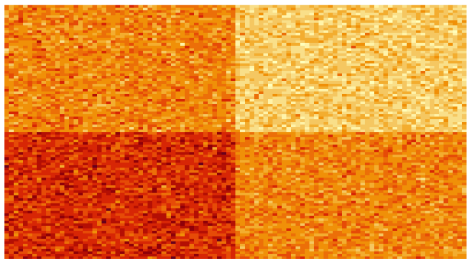


Figura 4.2: Datos originales.

A continuación, se permutan tanto las filas como las columnas de esta matriz para obtener una nueva matriz en la cual los bloques no se puedan distinguir a simple vista. Esta nueva matriz se encuentra representada en la figura 4.3.

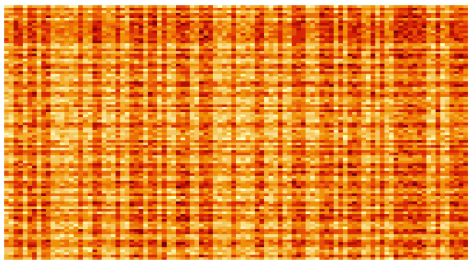


Figura 4.3: Datos permutados aleatoriamente.

### Análisis de biarquetipos

Una vez obtenida la nueva matriz con los datos barajados, se aplica el análisis de biarquetipos a esta matriz para encontrar los biarquetipos. En concreto se calcula el análisis de biarquetipos para los valores $k = (1, 2, 3, 4)$ y los valores $c = (1, 2, 3, 4)$.

Como se puede deducir de la figura 4.4, la mejor solución del biAA se encuentra cuando $k = 2$ y $c = 2$, es decir, cuando se tienen dos grupos de observaciones y dos grupos de variables.

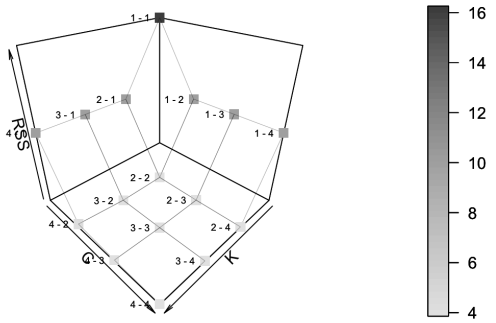



Figura 4.4: RSS obtenido en el análisis de biarquetipos.

## Resultados

Finalmente, se usan los arquetipos para poder agrupar las variables y las observaciones en grupos. De esta forma, se pueden reconstruir los datos originales. Para ello, se agrupan las observaciones de la matriz barajada en función de su similitud con las observaciones arquetípicas y se aplica el mismo método con las variables de la matriz barajada.

La matriz obtenida en este proceso está representada en la figura 4.5 y, como se puede observar, se recuperan bastante bien los grupos de los datos originales.

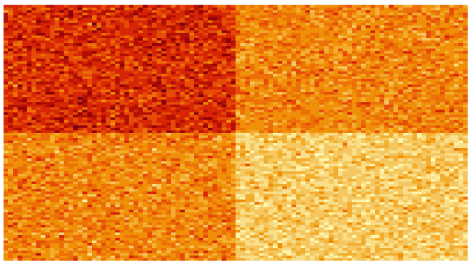


Figura 4.5: Datos recuperados a partir de los biarquetipos.

# Capítulo 5

# Aplicación

Una vez vista la parte teórica y la parte de implementación del análisis de biarquetipos, en esta sección veremos una aplicación del biAA al análisis deportivo.

El campo del análisis deportivo o *Sports Analytics* es un área de aplicación de la estadística de reciente creación, pero con una demanda de métodos y análisis cada vez mayor, pues el mundo deportivo, que en determinadas disciplinas mueve mucho dinero, busca entender la ingente cantidad de datos y estadísticas con las que cuenta hoy en día gracias a la tecnología, para así poder tomar mejores decisiones, y tener ventaja frente a equipos competidores.

Primero, detallaremos cuál es el problema y su motivación. A continuación, describiremos los datos con los que vamos a trabajar y veremos, paso a paso, la metodología usada. Finalmente, analizaremos los resultados obtenidos.

## 5.1. Contexto y motivación

Aunque el baloncesto es un deporte muy conocido, no está de más revisar muy brevemente sus reglas básicas. El baloncesto es un deporte de equipo en el que dos grupos se enfrentan entre sí durante cuatro períodos de tiempo de diez o doce minutos cada uno. El objetivo de ambos equipos es anotar más puntos que el rival encestando un balón en una canasta situada a 3,05 metros del suelo. Por lo que respecta a la puntuación de cada canasta, ésta puede ser de dos o tres puntos (dependiendo de la posición en el campo desde donde se haya realizado el tiro) o de un punto (si la canasta se ha hecho con un tiro libre tras una falta del rival).

En este deporte cada equipo está formado por cinco jugadores repartidos en cinco posiciones

diferentes: base (o *point guard*), escolta (o *shooting guard*), alero (o *small forward*), ala-pívot (o *power forward*) y pívot (o *center*). En la figura 5.1 se puede ver cómo se colocan en el campo los jugadores de un equipo en función de sus posiciones.

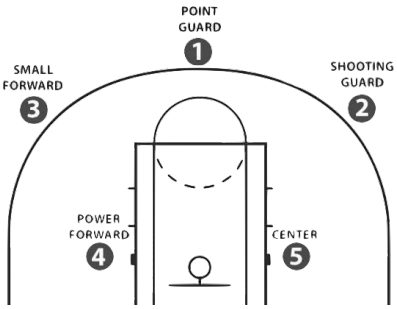



Figura 5.1: Organización espacial de un quinteto de baloncesto.

En este caso, aplicaremos el análisis de biarquetipos a un conjunto de datos que describe el comportamiento de los jugadores de la National Basketball Association, más conocida por sus siglas NBA. De esta forma, por cada posición del quinteto, aplicaremos el análisis de biarquetipos para obtener grupos de jugadores que se comportan de forma similar. Además, con esta técnica, a diferencia del clustering, también obtendremos grupos de variables que nos permitirán interpretar los arquetipos de una manera más sencilla.

En síntesis, vamos a averiguar, por posiciones en el campo, qué tipos (extremos) de jugadores se pueden encontrar en los partidos de la NBA.

## 5.2. Descripción de los datos

Los datos con los que hemos trabajado se han extraído de Sports Reference LLC (2020) y contienen métricas básicas de los últimos cinco años que permiten medir el comportamiento de los jugadores de la NBA. Dichas métricas son las siguientes:

- Player – Nombre del jugador.
- Pos – Posición. Puede ser C (pívot), PF (ala-pívot), SF (alero), SG (escolta) y PG (base).
- Tm – Equipo.
- G – Juegos jugados.
- GS – Juegos jugados de titular.

- MP – Minutos jugados.
- FG – Tiros de campo.
- FGA – Tiros de campo intentados.
- FG % – Porcentaje de acierto en los tiros de campo.
- 3P – Tiros de 3 puntos.
- 3PA – Tiros de 3 puntos intentados.
- 3P % – Porcentaje de acierto en los tiros de 3 puntos.
- 2P – Tiros de 2 puntos.
- 2PA – Tiros de 2 puntos intentados.
- 2P % – Porcentaje de acierto en los tiros de 2 puntos.
- FT – Tiros libres.
- FTA – Tiros libres intentados.
- FT % – Porcentaje de acierto en los tiros libres.
- ORB – Rebotes ofensivos.
- DRB – Rebotes defensivos.
- TRB – Rebotes totales.
- AST – Asistencias.
- STL – Robos.
- BLK – Bloqueos.
- TOV – Pérdidas.
- PF – Faltas personales.
- PTS – Puntos.
- eFG %: Porcentaje de acierto efectivo en los tiros de campo.
- TS %: Verdadero porcentaje de tiro.

Cabe destacar que estas métricas están adaptadas a 100 posesiones. Esto equipara los datos de jugadores que juegan en equipos con un ritmo de juego lento (menos posesiones y, por tanto, menos ocasiones para realizar tiros, rebotes, asistencias...) con los datos de los jugadores que juegan en equipos con un ritmo de juego más alto (más posesiones y, por ende, más tiros, más rebotes, más robos...).

## 5.3. Metodología

En primer lugar hemos filtrado los datos en función de los minutos jugados por partido (variable MP/G) para quitar del análisis aquellos jugadores que hayan jugado muy pocos minutos, ya que sus estadísticas puede que no sean representativas de su juego. Debido a esto, nos hemos quedado con los jugadores que hayan jugado más de 25 minutos por partido durante las últimas cinco temporadas.

Después, hemos seleccionado qué variables de las disponibles en el conjunto de datos analizar. En este caso, hemos utilizado las siguientes variables: FGA, FG %, 3P %, 2P %, FTA, FT %, TRB, AST, STL, BLK, TOV, PF, PTS, eFG % y TS %.

Además, hemos añadido cuatro variables nuevas que permiten describir mejor los datos:

- 2PP – Porcentaje de tiros de dos respecto al total de tiros de campo.
- 3PP – Porcentaje de tiros de tres respecto al total de tiros de campo.
- ORBP – Porcentaje de rebotes ofensivos respecto al total de rebotes.
- DRBP – Porcentaje de rebotes defensivos respecto al total de rebotes.

Conviene destacar que las variables que no hemos utilizado se han descartado porque o son muy similares a algunas de las ya existentes y, de esta forma, evitamos problemas relacionados con la redundancia de variables, o no tienen nada que ver con el comportamiento de los jugadores.

Una vez seleccionadas las variables, hemos separado los jugadores por su posición y hemos obtenido cinco grupos de jugadores. A continuación hemos escalado los datos y, finalmente, hemos aplicado el análisis de biarquetipos a cada uno de los grupos obtenidos utilizando el paquete de R `biaa`. Para obtener el modelo que ofrece la mejor solución al problema, hemos utilizado todas las combinaciones del producto cartesiano entre $k = \{2, 3, 4, 5, 6\}$ y $c = \{2, 3, 4, 5, 6\}$.

## 5.4. Resultados

En esta sección analizaremos por cada posición (bases, escoltas, aleros, ala-pívots y pívots) los biarquetipos obtenidos. En primer lugar veremos las variables más similares a cada variable arquetípica y, a continuación, usando estos datos, analizaremos las características de los jugadores más similares a cada jugador arquetípico obtenido.

## Bases

Tal y como se muestra en la figura 5.2, el codo que indica la mejor solución se encuentra cuando la variable $k$ es igual a 4 y la variable $c$ es igual a 4.

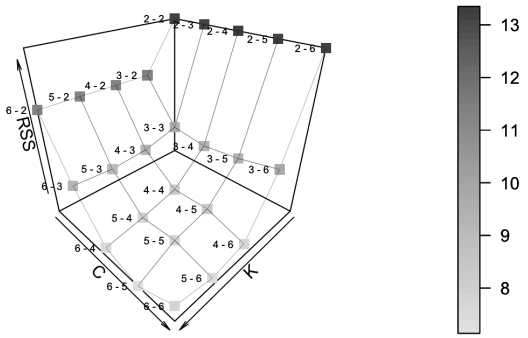



Figura 5.2: RSS obtenido en el análisis de biarquetipos de los bases.

Usando estos valores en el biAA, hemos encontrado los biarquetipos de la tabla 5.1.

| | VA 1 | VA 2 | VA 3 | VA 4 |
|---|---|---|---|---|
| JA 1 | -0.2922365 | 1.5727904 | -1.425607 | -0.1831402 |
| JA 2 | -4.6365145 | 0.9097224 | 2.514952 | 2.1047370 |
| JA 3 | 1.6343527 | 1.0133678 | 2.148200 | -1.2391750 |
| JA 4 | 0.2348413 | -1.6862277 | -1.023031 | 0.1810614 |

Tabla 5.1: Matriz de los biarquetipos obtenidos usando los datos de los bases.

## Variables

La tabla 5.2 contiene las variables de la muestra más similares a cada variable arquetípica. En concreto, en la tabla 5.2a se reflejan variables relacionadas con los triples y los tiros libres. Por lo que se refiere a la tabla 5.2b, ésta contiene variables relacionadas con la defensa, las asistencias y los rebotes. En cuanto a las variables de la tabla 5.2c, éstas están relacionadas con la cantidad y el porcentaje de acierto en los tiros de campo. Finalmente, en la tabla 5.2d están reflejados el porcentaje de rebotes que son ofensivos y el porcentaje de tiros de campo que son de dos.

| Variable | Similitud |
|---|---|
| 3P % | 0.881 |
| 3PP | 0.803 |
| FT % | 0.682 |

(a) VA 1

| Variable | Similitud |
|---|---|
| STL | 1.00 |
| AST | 0.979 |
| TOV | 0.610 |
| TRB | 0.602 |
| DRBP | 0.537 |

(b) VA 2

| Variable | Similitud |
|---|---|
| 2P % | 0.978 |
| eFG % | 0.810 |
| FG % | 0.756 |
| TS % | 0.697 |
| FTA | 0.611 |

(c) VA 3

| Variable | Similitud |
|---|---|
| ORBP | 1.00 |
| 2PP | 0.815 |
| PF | 0.509 |

(d) VA 4

Tabla 5.2: Variables más similares a cada una de las variables arquetípicas en los bases.

**Jugadores**

Basándonos en los resultados de la tabla 5.1, vamos a analizar los jugadores arquetípicos que existen en los bases:

- En la tabla 5.3 se encuentran los bases de la NBA más similares al jugador arquetípico 1. Estos jugadores tienen valores un poco bajos respecto a la primera y a la cuarta variable arquetípica y valores bastante bajos respecto a la segunda variable arquetípica, pero tienen valores altos respecto a la tercera variable arquetípica. En conclusión, si nos basamos en la tabla 5.2 podemos decir que son bases que lanzan poco y mal a canasta pero que consiguen muchas asistencias y robos.

| Jugador | Similitud |
|---|---|
| Ricky Rubio | 0.919 |
| Rajon Rondo | 0.872 |
| Lonzo Ball | 0.728 |

Tabla 5.3: Jugadores más similares al jugador arquetípico 1 (JA 1) en los bases.

- En la tabla 5.4 se encuentran los bases de la NBA más similares al jugador arquetípico 2. Estos jugadores tienen valores extremadamente bajos respecto a la primera variable arquetípica. Sin embargo tienen valores altos respecto a la segunda y tercera variable arquetípica y muy altos respecto a la cuarta variable arquetípica. Es decir, si nos fijamos en la tabla 5.2 podemos decir que son jugadores pésimos tanto desde el triple como desde los tiros libres, aunque tienen una eficiencia en el tiro alta, ya que casi todos sus tiros son de dos puntos (posiblemente mates). Además, son muy buenos robando, asistiendo y cogiendo rebotes.
- En la tabla 5.5 se encuentran los bases de la NBA más similares al jugador arquetípico 3. Estos jugadores tienen valores altos en la primera, en la segunda y en la tercera variable arquetípica pero tienen valores muy bajos respecto a la cuarta variable arquetípica. O sea,

| Jugador | Similitud |
|---|---|
| Ben Simmons | 1.00 |

Tabla 5.4: Jugadores más similares al jugador arquetípico 2 (JA 2) en los bases.

en base a la tabla 5.2 podemos decir que son jugadores que lanzan mucho y muy bien a canasta. También son buenos asistentes y cogen muchos rebotes (la mayoría defensivos).

| Jugador | Similitud |
|---|---|
| Stephen Curry | 0.837 |
| James Harden | 0.815 |
| LeBron James | 0.691 |

Tabla 5.5: Jugadores más similares al jugador arquetípico 3 (JA 3) en los bases.

- En la tabla 5.6 se encuentran los bases de la NBA más similares al jugador arquetípico 4. Estos jugadores tienen valores un poco altos respecto a la primera y a la última variable arquetípica pero tienen valores muy bajos respecto a la segunda y a la tercera variable arquetípica. En definitiva, con los resultados de la tabla 5.2 podemos decir que estos jugadores son bases que tiran pocos tiros y, además, con mal porcentaje de acierto de dos. Tampoco son buenos ni defendiendo ni asistiendo.

| Jugador | Similitud |
|---|---|
| Darius Garland | 1.00 |
| Jamal Murray | 0.746 |
| Derrick Rose | 0.737 |

Tabla 5.6: Jugadores más similares al jugador arquetípico 4 (JA 4) en los bases.

### Escoltas

Tal y como se muestra en la figura 5.3, el codo que indica la mejor solución para los escoltas se encuentra cuando la variable $k$ es igual a 4 y la variable $c$ es igual a 4.

Usando estos valores en el biAA, hemos encontrado los biarquetipos de la tabla 5.7.

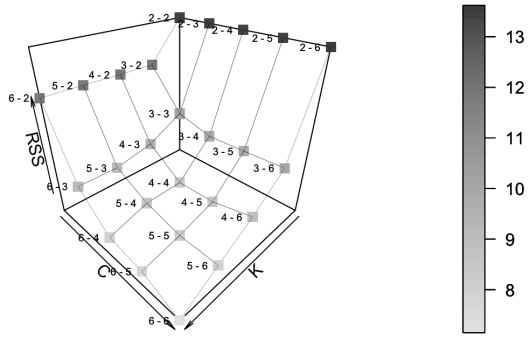



Figura 5.3: RSS obtenido en el análisis de biarquetipos de los escoltas.

| | VA 1 | VA 2 | VA 3 | VA 4 |
|---|---|---|---|---|
| JA 1 | -1.829614 | -0.4713240 | 0.02975573 | 2.3534955 |
| JA 2 | -1.218565 | 2.4112749 | 0.77962293 | -1.6646589 |
| JA 3 | 1.836905 | 0.9810745 | -2.71969830 | -0.9901859 |
| JA 4 | 1.378601 | -1.8338002 | 2.52928309 | -0.8652503 |

Tabla 5.7: Matriz de los biarquetipos obtenidos usando los datos de los escoltas.

**Variables**

La tabla 5.8 contiene las variables de la muestra más similares a cada variable arquetípica. En concreto, en la tabla 5.8a se recogen variables relacionadas con los tiros intentados y los puntos anotados. Por lo que se refiere a la tabla 5.8b, ésta contiene variables relacionadas con la defensa (robos y tapones) y con los rebotes. En cuanto a las variables de la tabla 5.8c, éstas están relacionadas con el porcentaje de acierto en los tiros de campo y con la eficiencia de éstos. Finalmente, en la tabla 5.8d, están reflejados el porcentaje de tiros de campos que son de tres puntos y las faltas personales.

| Variable | Similitud |
|---|---|
| TOV | 0.937 |
| FTA | 0.927 |
| FGA | 0.838 |
| PTS | 0.777 |
| 2PP | 0.658 |

(a) VA 1

| Variable | Similitud |
|---|---|
| STL | 0.922 |
| TRB | 0.858 |
| BLK | 0.811 |
| ORBP | 0.762 |

(b) VA 2

| Variable | Similitud |
|---|---|
| eFG % | 1.00 |
| TS % | 0.975 |
| 3P % | 0.903 |
| FT % | 0.746 |
| FG % | 0.597 |

(c) VA 3

| Variable | Similitud |
|---|---|
| 3PP | 0.758 |
| PF | 0.508 |

(d) VA 4

Tabla 5.8: Variables más similares a cada una de las variables arquetípicas en los escoltas.

**Jugadores**

Basándonos en los resultados de la tabla 5.7, vamos a analizar los jugadores arquetípicos que existen en los escoltas:

- En la tabla 5.9, se encuentran los escoltas de la NBA más similares al jugador arquetípico 1. Estos jugadores tienen valores bastante bajos respecto a la primera variable arquetípica y valores bajos respecto a la segunda variable, pero tienen valores un poco altos respecto a la tercera variable arquetípica y valores muy altos respecto a la última variable arquetípica. En conclusión, si nos basamos en la tabla 5.2, podemos decir que son escoltas que lanzan muy pocos tiros a canasta y, en la mayoría de ocasiones, desde la línea de tres puntos. Además, no son buenos defensores ni cogen muchos rebotes.

| Jugador | Similitud |
|---|---|
| Mychal Mulder | 1.00 |
| J.R. Smith | 0.890 |
| Wesley Matthews | 0.795 |

Tabla 5.9: Jugadores más similares al jugador arquetípico 1 (JA 1) en los escoltas.

- En la tabla 5.10, se encuentran los escoltas de la NBA más similares al jugador arquetípico 2. Estos jugadores tienen valores muy bajos respecto a la primera y a la cuarta variable arquetípica. Sin embargo tienen valores altos respecto a la segunda y a la tercera variable arquetípica. Es decir, si nos fijamos en la tabla 5.8, podemos decir que son jugadores que no intentan muchos tiros ni tienen un porcentaje elevado de tiros desde tres puntos. Sin embargo, tienen unos buenos porcentajes de acierto y son muy buenos cogiendo rebotes y defendiendo.

| Jugador | Similitud |
|---|---|
| Andre Iguodala | 0.769 |
| Danny Green | 0.531 |
| Shai Gilgeous-Alexander | 0.529 |

Tabla 5.10: Jugadores más similares al jugador arquetípico 2 (JA 2) en los escoltas.

- En la tabla 5.11, se encuentran los escoltas de la NBA más similares al jugador arquetípico 3. Estos jugadores tienen valores altos en la segunda variable y valores muy altos en la tercera variable arquetípica, pero tienen valores bajos respecto a la cuarta variable arquetípica y valores extremadamente bajos respecto a la tercera variable arquetípica. O sea, en base a la tabla 5.8, podemos decir que son jugadores que lanzan mucho a canasta pero con muy malos porcentajes y eficiencia. Además, son buenos defensores y

reboteadores y su porcentaje de tiros desde tres puntos es bajo respecto al de todos los escoltas.

| Jugador | Similitud |
|---|---|
| RJ Barrett | 0.802 |
| Andrew Wiggins | 0.607 |
| Marcus Smart | 0.549 |

Tabla 5.11: Jugadores más similares al jugador arquetípico 3 (JA 3) en los escoltas.

- En la tabla 5.12, se encuentran los escoltas de la NBA más similares al jugador arquetípico 4. Estos jugadores tienen valores muy bajos respecto a la segunda y a la cuarta variable arquetípica pero tienen valores muy altos respecto a la primera y tercera variable arquetípica. En definitiva, con los resultados de la tabla 5.8, podemos decir que estos jugadores son escoltas que intentan muchos tiros y, además, con una gran efectividad y acierto. sin embargo, son malos defendiendo, cogen pocos rebotes y, al igual que los anteriores, su porcentaje de tiros desde tres puntos también es bajo.

| Jugador | Similitud |
|---|---|
| J.J. Redick | 0.713 |
| Klay Thompson | 0.643 |
| Devin Booker | 0.593 |

Tabla 5.12: Jugadores más similares al jugador arquetípico 4 (JA 4) en los escoltas.

### Aleros

Tal y como se muestra en la figura 5.4, el codo que indica la mejor solución se encuentra cuando la variable $k$ es igual a 4 y la variable $c$ es igual a 4.

Usando estos valores en el biAA, hemos encontrado los biarquetipos de la tabla 5.13.

| | VA 1 | VA 2 | VA 3 | VA 4 |
|---|---|---|---|---|
| JA 1 | 0.4788568 | 0.2313500 | -1.6477862 | 2.6365053 |
| JA 2 | 3.1910468 | -1.6750555 | 2.9264180 | -2.0003849 |
| JA 3 | -1.7296907 | -0.3108191 | 0.2096416 | -0.1027838 |
| JA 4 | -0.7768359 | 2.2593425 | -2.0406625 | -0.3744917 |

Tabla 5.13: Matriz de los biarquetipos obtenidos usando los datos de los aleros.

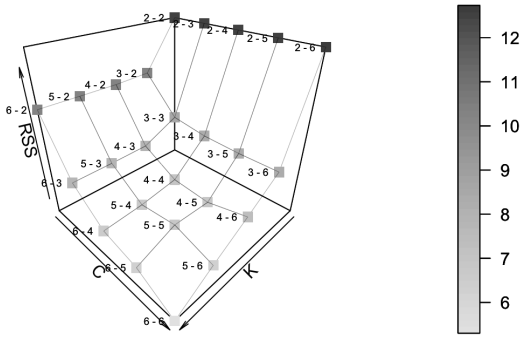



Figura 5.4: RSS obtenido en el análisis de biarquetipos de los aleros.

### Variables

La tabla 5.14, contiene las variables de la muestra más similares a cada variable arquetípica. En concreto, en la tabla 5.14a se recogen variables relacionadas con los porcentajes de acierto en los tiro de campo y con la eficiencia. Por lo que se refiere a la tabla 5.14b, ésta contiene variables relacionadas con los tapones y los rebotes. En cuanto a las variables de la tabla 5.14c, éstas están relacionadas con el porcentaje de tiros desde tres puntos y con el porcentaje de rebotes que son ofensivos. Finalmente, en la tabla 5.14d, están reflejadas variables que indican el uso del balón (tiros intentados, puntos, asistencias, pérdidas...).

| Variable | Similitud |
|---|---|
| TS % | 0.916 |
| 2P % | 0.903 |
| eFG % | 0.886 |
| FG % | 0.567 |
| 3P % | 0.558 |

(a) VA 1

| Variable | Similitud |
|---|---|
| ORBP | 1.00 |
| BLK | 0.906 |
| TRB | 0.633 |

(b) VA 2

| Variable | Similitud |
|---|---|
| 3PP | 0.886 |
| DRBP | 0.699 |
| PF | 0.592 |

(c) VA 3

| Variable | Similitud |
|---|---|
| 1 FTA | 0.924 |
| 2 FGA | 0.896 |
| 3 TOV | 0.839 |
| 4 PTS | 0.799 |
| 5 AST | 0.782 |

(d) VA 4

Tabla 5.14: Variables más similares a cada una de las variables arquetípicas en los aleros.

### Jugadores

Basándonos en los resultados de la tabla 5.13, vamos a analizar los jugadores arquetípicos que existen en los aleros:

- En la tabla 5.15, se encuentran los aleros de la NBA más similares al jugador arquetípico 1. Estos jugadores tienen valores bastante bajos respecto a la tercera variable arquetípica. Sin

embargo, tienen valores un poco altos respecto a la primera y segunda variable arquetípica y valores muy altos respecto a la última variable arquetípica. En conclusión, si nos basamos en la tabla 5.2, podemos decir que son aleros que tienen mucho uso del balón, es decir, hacen muchas jugadas. También tienen buenos porcentajes de tiro y rebotean bien (muchos defensivos). Sin embargo, el porcentaje de tiros que son de tres puntos es bajo.

| Jugador | Similitud |
|---|---|
| Kawhi Leonard | 0.805 |
| Kevin Durant | 0.805 |
| DeMar DeRozan | 0.704 |

Tabla 5.15: Jugadores más similares al jugador arquetípico 1 (JA 1) en los aleros.

- En la tabla 5.16, se encuentran los aleros de la NBA más similares al jugador arquetípico 2. Estos jugadores tienen valores muy bajos respecto a la segunda y a la cuarta variable arquetípica. Sin embargo tienen valores extremadamente altos respecto a la primera y a la tercera variable arquetípica. Es decir, si nos fijamos en la tabla 5.14 podemos decir que son jugadores que no intentan muchos tiros ni cogen muchos rebotes. Sin embargo, tienen unos porcentajes de acierto muy elevados y la mayoría de sus tiros son triples.

| Jugador | Similitud |
|---|---|
| Duncan Robinson | 1.00 |
| Joe Ingles | 0.508 |
| Danuel House | 0.506 |

Tabla 5.16: Jugadores más similares al jugador arquetípico 2 (JA 2) en los aleros.

- En la tabla 5.17, se encuentran los aleros de la NBA más similares al jugador arquetípico 3. Estos jugadores tienen valores intermedios en la segunda, en la tercera y en la cuarta variable y valores muy bajos en la primera variable arquetípica. O sea, en base a la tabla 5.14, podemos decir que son jugadores intermedios que tienen unos porcentajes de acierto en los tiros de campo muy malos.

| Jugador | Similitud |
|---|---|
| Cam Reddish | 0.974 |
| De'Andre Hunter | 0.857 |
| Justise Winslow | 0.738 |

Tabla 5.17: Jugadores más similares al jugador arquetípico 3 (JA 3) en los aleros.

- En la tabla 5.18, se encuentran los aleros de la NBA más similares al jugador arquetípico 4. Estos jugadores tienen valores muy bajos respecto a la primera y a la tercera variable arquetípica pero tienen valores muy altos respecto a la segunda variable arquetípica. En

definitiva, con los resultados de la tabla 5.14, podemos decir que estos jugadores son aleros que intentan pocos triples y, además, con poca efectividad y acierto. Sin embargo, son muy buenos poniendo tapones y reboteando.

| Jugador | Similitud |
|---|---|
| Jonathan Isaac | 0.926 |
| T.J. Warren | 0.683 |
| Miles Bridges | 0.596 |

Tabla 5.18: Jugadores más similares al jugador arquetípico 4 (JA 4) en los aleros.

### Ala-pívots

Tal y como se muestra en la figura 5.5, el codo que indica la mejor solución se encuentra cuando la variable $k$ es igual a 4 y la variable $c$ es igual a 4.

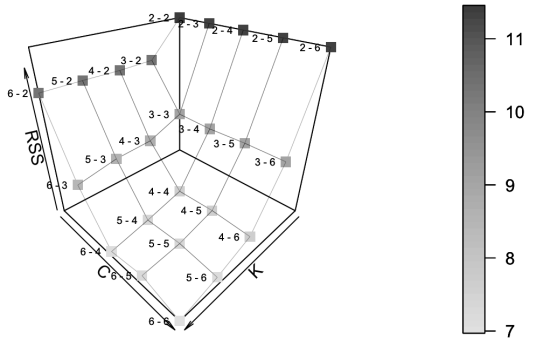



Figura 5.5: RSS obtenido en el análisis de biarquetipos de los ala-pívots.

Usando estos valores en el biAA, hemos encontrado los biarquetipos de la tabla 5.19.

| | VA 1 | VA 2 | VA 3 | VA 4 |
|---|---|---|---|---|
| JA 1 | -0.2981531 | 2.2843019 | -1.9285385 | 1.969913 |
| JA 2 | 2.6158836 | 0.3200589 | 0.9305102 | -1.974919 |
| JA 3 | 0.3262925 | -4.2885159 | -1.0701444 | 1.923676 |
| JA 4 | -1.3388367 | 0.5844577 | 1.2277051 | -1.474463 |

Tabla 5.19: Matriz de los biarquetipos obtenidos usando los datos de los ala-pívots.

### Variables

La tabla 5.20, contiene las variables de la muestra más similares a cada variable arquetípica. En concreto, en la tabla 5.20a, se recogen variables relacionadas con la defensa y las asistencias. Por lo que se refiere a la tabla 5.20b, ésta contiene variables relacionadas con los puntos anotados y los tiros de campo intentados. En cuanto a las variables de la tabla 5.20c, éstas están relacionadas con los triples, con los tiros libres y con el porcentaje de rebotes que son ofensivos. Finalmente, en la tabla 5.20d, están recogidas variables que reflejan el porcentaje de acierto en los tiros de campo, el porcentaje de rebotes ofensivos capturados y la efectividad.

| Variable | Similitud |
|---|---|
| STL | 0.985 |
| PF | 0.886 |
| BLK | 0.650 |
| AST | 0.645 |

(a) VA 1

| Variable | Similitud |
|---|---|
| PTS | 1.00 |
| FTA | 0.978 |
| FGA | 0.852 |
| TS % | 0.652 |
| TOV | 0.573 |

(b) VA 2

| Variable | Similitud |
|---|---|
| 1 3PP | 0.899 |
| 2 DRBP | 0.845 |
| 3 FT % | 0.812 |
| 4 3P % | 0.646 |

(c) VA 3

| Variable | Similitud |
|---|---|
| ORBP | 0.913 |
| 2P % | 0.744 |
| FG % | 0.661 |
| eFG % | 0.619 |
| 2PP | 0.596 |

(d) VA 4

Tabla 5.20: Variables más similares a cada una de las variables arquetípicas en los ala-pívots.

### Jugadores

Basándonos en los resultados de la tabla 5.19, vamos a analizar los jugadores arquetípicos que existen en los ala-pívots:

- En la tabla 5.21, se encuentran los ala-pívots de la NBA más similares al jugador arquetípico 1. Estos jugadores tienen valores bastante bajos respecto a la tercera variable arquetípica. Sin embargo, tienen valores muy altos respecto a la segunda y a la cuarta variable arquetípica. En conclusión, si nos basamos en la tabla 5.20, podemos decir que son ala-pívots que intentan muchos tiros de campo y consiguen muchos puntos. Además, tienen muy buenos porcentajes en los tiros de dos y muchos de sus rebotes son ofensivos. Sin embargo, son malos respecto a los triples y a los tiros libres.

| Jugador | Similitud |
|---|---|
| Zion Williamson | 1.00 |
| John Collins | 0.804 |
| Bam Adebayo | 0.659 |

Tabla 5.21: Jugadores más similares al jugador arquetípico 1 (JA 1) en los ala-pívots.

- En la tabla 5.22, se encuentran los ala-pívots de la NBA más similares al jugador arquetípico 2. Estos jugadores tienen valores muy bajos respecto a la cuarta variable arquetípica. Sin embargo tienen valores altos respecto a la tercera variable y extremadamente altos respecto a la primera variable arquetípica. Es decir, si nos fijamos en la tabla 5.20, podemos decir que son jugadores muy buenos en la parte defensiva y, dentro de los ala-pívots, tienen un porcentaje elevado de triples intentados respecto a los tiros totales. Por el contrario, sus porcentajes de acierto en el tiro son bajos y casi no capturan rebotes ofensivos.

| Jugador | Similitud |
|---|---|
| Draymond Green | 0.679 |
| Robert Covington | 0.611 |

Tabla 5.22: Jugadores más similares al jugador arquetípico 2 (JA 2) en los ala-pívots.

- En la tabla 5.23, se encuentran los ala-pívots de la NBA más similares al jugador arquetípico 3. Estos jugadores tienen valores bajos en la tercera variable y valores extremadamente bajos en la segunda variable arquetípica. Por otro lado, tienen valores bastante elevados en la cuarta variable arquetípica. O sea, en base a la tabla 5.20, podemos decir que son jugadores que casi no intentan ningún tiro por partido y casi no intentan triples. Además, capturan bastantes rebotes ofensivos y su porcentaje de acierto en el tiro, especialmente de dos, es alto.

| Jugador | Similitud |
|---|---|
| P.J. Tucker | 0.539 |

Tabla 5.23: Jugadores más similares al jugador arquetípico 3 (JA 3) en los ala-pívots.

- En la tabla 5.24, se encuentran los ala-pívots de la NBA más similares al jugador arquetípico 4. Estos jugadores tienen valores muy bajos respecto a la primera y a la cuarta variable arquetípica pero tienen valores altos respecto a la segunda y, sobretodo, a la tercera variable arquetípica. En definitiva, con los resultados de la tabla 5.20, podemos decir que estos ala-pívots intentan bastantes tiros y muchos de éstos son triples. Sin embargo, son malos defendiendo y en el porcentaje de acierto en los tiros de campo.

| Jugador | Similitud |
|---|---|
| Danilo Gallinari | 0.999 |
| Lauri Markkanen | 0.815 |
| Harrison Barnes | 0.755 |

Tabla 5.24: Jugadores más similares al jugador arquetípico 4 (JA 4) en los ala-pívots.

### Pívots

Tal y como se muestra en la figura 5.6, el codo que indica la mejor solución para los pívots se encuentra cuando la variable $k$ es igual a 4 y la variable $c$ es igual a 4.

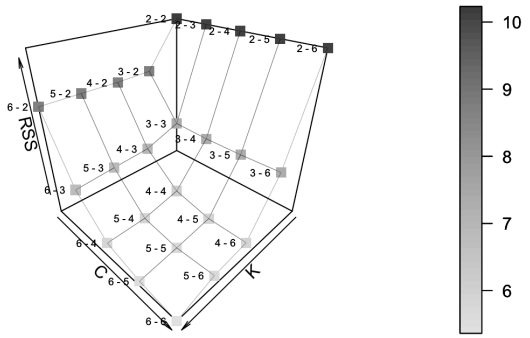



Figura 5.6: RSS obtenido en el análisis de biarquetipos de los pívots.

Usando estos valores en el biAA, hemos encontrado los biarquetipos de la tabla 5.25.

| | VA 1 | VA 2 | VA 3 | VA 4 |
|---|---|---|---|---|
| JA 1 | 0.8752719 | 2.7985107 | -0.9973723 | -1.1141375 |
| JA 2 | -1.8042945 | -0.6155108 | -0.1135502 | 2.5112030 |
| JA 3 | -1.3058777 | -0.5088128 | 2.6186335 | 0.5825802 |
| JA 4 | 0.8792024 | -0.5924802 | -0.4925579 | -0.8534247 |

Tabla 5.25: Matriz de los biarquetipos obtenidos usando los datos de los pívots.

### Variables

La tabla 5.26, contiene las variables de la muestra más similares a cada variable arquetípica. En concreto, en la tabla 5.26a, se recogen variables relacionadas con el triple, los tiros libres y los rebotes defensivos. Por lo que se refiere a la tabla 5.26b, ésta contiene variables relacionadas con la cantidad de tiros libres, con los puntos, con los robos y con la cantidad de pérdidas. En cuanto a las variables de la tabla 5.26c, éstas están relacionadas con el porcentaje de tiros de campo y la eficiencia. Finalmente, en la tabla 5.26d, están recogidas variables relacionadas con los rebotes ofensivos y la cantidad de tiros que son de dos.

| Variable | Similitud |
|---|---|
| 3PP | 0.975 |
| FT % | 0.957 |
| 3P % | 0.950 |
| DRBP | 0.814 |
| FGA | 0.597 |

(a) VA 1

| Variable | Similitud |
|---|---|
| TOV | 1.00 |
| FTA | 0.785 |
| STL | 0.694 |
| PTS | 0.650 |
| PF | 0.548 |

(b) VA 2

| Variable | Similitud |
|---|---|
| eFG % | 0.896 |
| TS % | 0.868 |
| 2P % | 0.842 |
| FG % | 0.738 |
| BLK | 0.601 |

(c) VA 3

| Variable | Similitud |
|---|---|
| ORBP | 0.963 |
| 2PP | 0.689 |

(d) VA 4

Tabla 5.26: Variables más similares a cada una de las variables arquetípicas en los pívots.

**Jugadores**

Basándonos en los resultados de la tabla 5.25, vamos a analizar los jugadores arquetípicos que existen en los pívots:

- En la tabla 5.27, se encuentran los pívots de la NBA más similares al jugador arquetípico 1. Estos jugadores tienen valores bastante bajos respecto a la tercera y a la cuarta variable arquetípica. Sin embargo, tienen valores muy altos respecto a la primera y a la segunda variable arquetípica. En conclusión, si nos basamos en la tabla 5.26 podemos decir que son pívots que lanzan muchos tiros, con una gran cantidad de triples (para ser pívots). También tienen buenos porcentajes en triples y tiros libres pero, para la posición en la que juegan, no tienen buenos porcentajes de dos ni cogen muchos rebotes ofensivos.

| Jugador | Similitud |
|---|---|
| DeMarcus Cousins | 0.982 |
| Joel Embiid | 0.878 |
| Nikola Jokić | 0.574 |

Tabla 5.27: Jugadores más similares al jugador arquetípico 1 (JA 1) en los pívots.

- En la tabla 5.28, se encuentran los pívots de la NBA más similares al jugador arquetípico 2. Estos jugadores tienen valores muy bajos respecto a la primera y a la segunda variable arquetípica. Por el contrario, tienen valores extremadamente altos respecto a la cuarta variable arquetípica. Es decir, si nos fijamos en la tabla 5.26, podemos decir que son jugadores que no tiran bien desde el triple ni desde los tiros libres. Sin embargo, muchos de sus rebotes son ofensivos y la mayoría de sus tiros son de dos puntos.
- En la tabla 5.29, se encuentran los pívots de la NBA más similares al jugador arquetípico 3. Estos jugadores tienen valores un poco altos en la cuarta variable y valores muy altos en la tercera variable arquetípica. Sin embargo, tienen valores bajos en la segunda variable y valores bastante bajos en la primera variable arquetípica. O sea, en base a la tabla 5.26,

| Jugador | Similitud |
|---|---|
| Steven Adams | 0.846 |
| Tristan Thompson | 0.703 |
| Andre Drummond | 0.635 |

Tabla 5.28: Jugadores más similares al jugador arquetípico 2 (JA 2) en los pívots.

podemos decir que son pívots que tiran poco y mal desde el triple y desde los tiros libres, pero que tienen unos porcentaje de acierto y una efectividad muy elevados.

| Jugador | Similitud |
|---|---|
| DeAndre Jordan | 1.00 |
| Rudy Gobert | 0.946 |
| Clint Capela | 0.685 |

Tabla 5.29: Jugadores más similares al jugador arquetípico 3 (JA 3) en los pívots.

- En la tabla 5.30, se encuentran los pívots de la NBA más similares al jugador arquetípico 4. Estos jugadores tienen valores bajos respecto a la segunda, tercera y cuarta variable arquetípica, pero tienen valores muy altos respecto a la primera variable arquetípica. En definitiva, con los resultados de la tabla 5.26, podemos decir que estos jugadores son pívots que son buenos respecto a los triples y los tiros libres y tienen un porcentaje de rebotes defensivos alto.

| Jugador | Similitud |
|---|---|
| Al Horford | 1.00 |
| Serge Ibaka | 1.00 |
| Brook Lopez | 0.980 |

Tabla 5.30: Jugadores más similares al jugador arquetípico 4 (JA 4) en los pívots.

# Capítulo 6

# Conclusiones

En este trabajo se ha propuesto un nuevo concepto en estadística, el análisis de biarquetipos. En concreto, se definido matemáticamente y se ha propuesto un método para resolverlo. También se ha implementado el biAA en el lenguaje R y hasta se ha construido un paquete de R que recoge esta implementación. Además, se ha ilustrado el análisis de biarquetipos con datos de juguete para que se entendiera este nuevo concepto. Por último, se ha aplicado a un problema del análisis deportivo, donde el biAA ha permitido descubrir patrones escondidos en los datos, dejando que fueran los propios datos los que hablaran. Concretamente, hemos analizado métricas de los jugadores de la NBA, obteniendo, por cada posición existente en un quinteto de baloncesto, qué tipo de jugadores (roles) existen en esta competición. Estos resultados pueden ayudar a los ejecutivos a entender, de forma más sencilla, qué perfil tiene cada jugador (ya que los jugadores están expresados como una mixtura de biarquetipos) a la hora de buscar nuevos fichajes.

Como trabajo futuro, hay muchas preguntas abiertas por explorar debido a que los biarquetipos son un concepto inédito y original en estadística. Un primer paso podría ser realizar una comparativa entre los resultados obtenidos por biclustering y biAA en un trabajo de simulación y con bases de datos reales de distintos campos. Por otro lado, al ser una técnica nueva y con muchos campos de aplicación al igual que el biclustering, a nivel práctico los posibles trabajos futuros son enormes.

Pero es que a nivel teórico, los posibles trabajos futuros son también numerosos. El análisis de biarquetipos se podría extender al análisis de biarquetipoides, es decir, si restringimos los posibles arquetipos a que sean elementos de la muestra, tanto en individuos como en variables (o solo en individuos y no en variables). Pero además, el biAA se ha definido únicamente para datos numéricos de variables continuas, igual que originalmente fue definido el análisis de arquetipos. Sin embargo, podría extender a otro tipo de datos, como datos funcionales, binarios, ordinales, mezclas, datos censurados, datos circulares, etc.

# Referencias


Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*(53), 370–418.

Bezdek, J. C., Ehrlich, R., y Full, W. (1984). FCM: the fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), 191–203.

Cabero, I., y Epifanio, I. (2019). Archetypal analysis: An alternative to clustering for unsupervised texture segmentation. *Image Analysis & Stereology*, *38*(2), 151–160. doi: 10.5566/ias.2052

Cheng, Y., y Church, G. M. (2000). Biclustering of expression data. En *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* (Vol. 8, pp. 93–103).

Cutler, A., y Breiman, L. (1994). Archetypal analysis. *Technometrics*, *36*(4), 338–347.

Davis, T., y Love, B. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, *21*, 234-42. doi: 10.1177/0956797609357712

Forsati, R., Doustdar, H. M., Shamsfard, M., Keikha, A., y Meybodi, M. R. (2013). A fuzzy co-clustering approach for hybrid recommender systems. *International Journal of Hybrid Intelligent Systems*, *10*(2), 71–81.

Hastie, T., Tibshirani, R., y Friedman, J. (2009). The Elements of Statistical Learning. New York, NY, USA: Springer.

Kerr, G., Ruskin, H. J., Crane, M., y Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in Biology and Medicine*, *38*(3), 283–293.

Koutsonikola, V. A., y Vakali, A. (2009). A fuzzy bi-clustering approach to correlate web users and pages. *IJ Knowledge and Web Intelligence*, *1*(1/2), 3–23.

Lawson, C. L., y Hanson, R. J. (1974). *Solving least squares problems.* Prentice Hall.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. En *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297).

Maurizio, V. (2001). Double k-means clustering for simultaneous classification of objects and variables. En S. Borra, R. Rocci, M. Vichi, y M. Schader (Eds.), *Advances in Classification and Data Analysis* (pp. 43–52). Berlin, Heidelberg: Springer.

Sports Reference LLC. (2020). *2018-19 NBA Player Stats: Per 100 Possessions.* `https://www.basketball-reference.com/play-index/psl_finder.cgi` (29/05/2020).

Thurau, C., Kersting, K., Wahabzada, M., y Bauckhage, C. (2012). Descriptive matrix factorization for sustainability adopting the principle of opposites. *Data Mining & Knowledge Discovery*, *24*, 325-354. doi: 10.1007/s10618-011-0216-z

Van Mechelen, I., Bock, H.-H., y De Boeck, P. (2004). Two-mode clustering methods: a structured overview. *Statistical Methods in Medical Research*, *13*(5), 363–394.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353.

Zhao, H., Wee-Chung Liew, A., Z Wang, D., y Yan, H. (2012). Biclustering analysis for pattern discovery: current techniques, comparative studies and applications. *Current Bioinformatics*, *7*(1), 43–55.

# Anexo A

# Notación matemática

| Símbolo | Descripción |
|---|---|
| $X_{n\times m}$ | Matriz con $n$ filas y $m$ columnas. |
| $x_{ij}$ | Elemento de la matriz $X$ situado en la $i$-ésima fila y en la $j$-ésima columna. |
| $\hat{x}$ | Aproximación del elemento $x$. |
| $X^T$ | Transpuesta de la matriz $X$. |
| $X^+$ | Pseudo-inversa de Moore-Penrose de la matriz $X$. |