UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

ÚS D'ANÀLISI MULTIVARIANT PER A LA DEFINICIÓ DE NOVES POSICIONS DE JOC AL BASQUETBOL PROFESSIONAL

Autor: Marc Maset López

Directors: Dr. Josep Vives Santa Eulàlia
Dra. M. Carme Florit Selma
Realitzat a: Departament d'estadística

Barcelona, 19 de juny de 2021

Abstract

Data and sports have been side to side for years. Anyone who watches sports during the weekend, and not only in professional teams, will most likely find people gathering data from those games: points or goals scored, fouls, the time a player has spent on court... Nowadays, mainly powerful teams are following a trend consisting on deeply analyzing these data with advanced methods. This project is an introductory example to these kind of studies.
Using data from Liga EBA, the fourth tier in the Spanish basketball competition, and well-known multivariate analysis methods such as CPA or *k-means* clustering, this project has as its main objective the classification of players depending on their statistical performance, escaping the classical guard - forward - center division. What this thesis looks for is a totally objective classification based in no more than statistics, following the footsteps of what M. Alagappan and his coworkers did in the NBA.

Resum

Les dades i l'esport han anat de la mà des de fa anys. Qualsevol que faci un recorregut per pavellons, pistes o camps del territori durant un cap de setmana de competició no necessàriament profesional veurà molta gent acumulant dades de partits: punts, gols, faltes, temps de joc... Recentment s'ha desenvolupat una tendència, especialment en clubs potents, d'analitzar aquestes dades amb mètodes avançats. En aquest treball es fa una introducció a aquests tipus d'estudi.
A partir de dades de la Liga EBA, la quarta categoria del bàsquet estatal, i de mètodes d'anàlisi multivariant coneguts com l'ACP o el clústering *k-means*, per citar-ne alguns, aquest treball persegueix l'objectiu de classificar els jugadors en funció dels seus valors estadístics, fugint de la classificació clàssica dels jugadors en base, aler o pivot a partir de la seva estatura i qualitats físiques. Es busca una classificació totalment objectiva, seguint les passes d'estudis similars realitzats per M. Alagappan i els seus col·laboradors a l'NBA, la lliga de bàsquet dels Estats Units.

2020 Mathematics Subject Classification. 15A18, 91C99, 62H25, 62H30

Agraïments

Aquest treball no hagués estat possible sense molta gent que ha aportat el seu granet de sorra perquè el projecte tirés endavant. Aquesta pàgina és per a tots ells.

- Marta, Josep, Maria, Oriol i tota la família. Gràcies no només per haver-me aguantat durant els mesos que ha durat el treball sinó també per haver suportat quatre anys de carrera parlant només de "mates i bàsquet". Tenir una família que em recolzi com vosaltres ho feu no té preu.
- Josep Vives i Carme Florit. Gràcies per haver-vos ofert a tutoritzar-me el treball tot i no ser del vostre àmbit d'estudi després d'haver estat buscant tutors durant tres mesos.
- Àlex Rodríguez, Sergi Muñoz, Miquel Villaró i Marta Maset, gràcies per la lectura i revisió del treball.
- Els cossos tècnics del Sènior EBA i els Cadets A i 1 de la UE Mataró, en especial a Victor Solanes, Joan Rubio i Sergi Teodoro. Gràcies per obrir-me les portes d'un lloc on es creu en l'analítica, i a ajudar-me a desenvolupar una passió i una vocació de futur que espero que comenci amb aquest treball.
- A tothom que m'ha contactat preguntant pel treball, gràcies per mostrar interès en un tema en expansió com aquest.

Índex

1 Introducció

El projecte

Aquest projecte té com a objectiu la classificació estadística dels jugadors de Liga EBA. La part principal del treball presenta la base matemàtica sobre la qual es fonamenten els mètodes usats a la pràctica, des de l'ACP fins als algoritmes de classificació, de forma similar a un manual bàsic d'anàlisi multivariant. Un cop acabat l'estudi teòric comença una segona secció enfocada a l'anàlisi de les dades recollides, on es busca aplicar els mètodes vistos durant la primera part del treball a un cas real: la classificació dels jugadors de la Liga EBA, la quarta divisió del basquetbol estatal, en funció de diferents característiques estadístiques.

Tota la programació del projecte està feta a partir de Python i R. Python s'ha usat per fer el web-scrapping i extreure totes les dades d'internet des de la web de la Federación Española de Baloncesto, mentre que tot el tractament d'aquestes dades i l'aplicació dels mètodes estadístics es fa en R, el llenguatge per excel·lència de l'anàlisi de dades.

Estructura de la Memòria

La part principal de la memòria està estructurada en dues grans seccions. En primer lloc apareix la vessant matemàtica del treball, amb un estudi profund de quatre àmbits de l'anàlisi multivariant: l'anàlisi de components principals, el biplot, el clústering i els algoritmes de classificació. En el primer capítol es realitza una anàlisi de l'ACP, anant des de la seva construcció com a combinació lineal de les variables originals fins arribar a la projecció dels individus en aquests nous eixos, estudiant també l'ACP normalitzada i la quantitat de components principals que cal retenir segons el problema. Al final del capítol s'estudia el biplot com una forma de representar en uns mateixos eixos els individus estudiats amb l'ACP i les variables de la matriu de dades. A aquest efecte es presenta la descomposició en valors singulars d'una matriu, que generalitza la diagonalització de matrius quadrades i permet construir aquesta visualització conjunta d'individus i variables.

Al capítol 3 hi ha un canvi important de temàtica, passant dels algoritmes de visualització als algoritmes de clústering. Després d'un estudi general de la idea que hi ha darrere d'aquest tipus de mètodes s'estudien per separat els algoritmes jeràrquics i els no jeràrquics. Pels primers és bàsic el concepte de geometria ultramètrica, i adaptar després alguns conceptes a espais mètrics -que és on es treballa la majoria de vegades-, mentre que l'estudi dels segons està altament relacionat amb l'algoritme *k-means*, molt conegut. Per acabar, es tracta el *silhouette plot*, una forma d'avaluar la classificació obtinguda. D'aquí és d'on es pretenen obtenir els grups de jugadors que són l'objectiu principal del treball.

Per acabar la secció teòrica, al quart capítol del treball es tracten de manera breu els algoritmes de classificació, que reparteixen els individus segons una variable resposta categòrica. En particular, s'estudien els algoritmes de K-Nearest Neighbors, basat en el classificador de Bayes, i l'anàlisi discriminant tant lineal com quadràtic.

En quant a la vessant pràctica del treball, es dedica l'últim capítol de la memòria a explicar el procés seguit en l'aplicació de tota la teoria a un cas pràctic com és la classificació estadística dels jugadors de Liga EBA. Aquest capítol està orientat molt especialment als resultats obtinguts, fent ús de tots els gràfics que retorna R per tal de treure conclusions, alhora que també es fa algun comentari sobre la recollida de dades des de la web de la Federación Española de Baloncesto fent ús de Python. Aquest aspecte no es tracta profundament perquè no és l'objectiu del treball però sí cal fer-ne esment per saber amb quines dades s'està treballant.

Aquest projecte disposa també d'un annex on es completa la informació donada al treball. En aquestes pàgines s'hi troben diversos resultats previs a l'estudi realitzat, necessaris per poder provar alguns dels teoremes que apareixen, així com proposicions i enunciats relacionats amb

l'estudi del treball però amb una temàtica més propera a l'àlgebra lineal que a l'anàlisi multivariant.

En una segona secció de l'annex es presenta informació complementària a l'anàlisi de les dades dels jugadors, des de taules explicant el significat de les variables fins algunes imatges que completen la informació obtinguda. Per últim, els codis de Python i R utilitzats durant el treball estan disponibles en arxius adjunts.

2 Anàlisi de components principals per l'estudi dels individus

L'anàlisi de components principals o ACP és un mètode que permet reduir la dimensió de les dades d'una matriu per buscar una representació gràfica més significativa dels individus. Donada una matriu de dades $X \in \mathcal{M}_{n\times p}$ es poden representar els n individus com punts a $\mathbb{R}^p$, on cadascuna de les coordenades ve donada pel valor que pren la variable j per l'individu i. L'ACP busca projectar aquest núvol de punts sobre un subespai $\mathbb{R}^k \subseteq \mathbb{R}^p$ de tal manera que es maximitzi la variància projectada[2]. És una eina prèvia al clústering de dades, un dels objectius finals d'aquest treball, i permet fer un primer estudi dels individus i les variables[3].
Les referències principals per aquest capítol són [4], [9] i [13]. [24] ofereix informació extra sobre el biplot, i de [3] s'ha extret tot allò relacionat amb la descomposició en valors singulars.

2.1 Centrat de les dades

Degut a les avantatges tècniques que reporta, el primer pas per a qualsevol ACP és centrar la matriu de dades X. Per a fer-ho cal definir la matriu de centrat:

Definició 2.1 (Matriu de centrat)**.** *Siguin $I \in \mathcal{M}_{n\times n}(\mathbb{R})$ la matriu identitat de dimensió n i $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ la matriu definida segons $A = (a_{ij})_{1\leq i,j\leq n}$, $a_{ij} = 1\ \forall i, j$. Es defineix aleshores la matriu de centrat com la matriu*

$$H = I - \frac{1}{n}A.$$

Posant $H = (h_{ij})_{1\leq i,j\leq n}$, es té que

$$h_{ij} = \begin{cases} \frac{n-1}{n} & si\ i = j, \\ -\frac{1}{n} & si\ i \neq j. \end{cases}$$

Definició 2.2 (Matriu de dades centrada)**.** *Sigui $X \in \mathcal{M}_{n\times p}$ una matriu de dades i H la matriu de centrat segons s'ha construït en la definició anterior. Aleshores, es defineix la matriu centrada com HX.*

La Proposició A.1 de l'annex A mostra algunes propietats importants de la matriu de centrat i de la matriu de dades centrada. El procés de centrat de les dades provoca que el centre de gravetat del núvol de punts dels individus -això és, el punt que té per coordenada j-èsima el valor $\overline{X_j}$- es trobi en l'origen de coordenades. El fet de centrar les dades permet també evitar alguns dels problemes que sorgeixen en aplicar els mètodes de classificació.

2.2 Construcció de les Components Principals

En aquesta secció es tracta la construcció matemàtica de l'ACP. Per a fer-ho, l'objectiu és construir un conjunt de combinacions lineals de les variables originals,

$$\varphi_i = \sum_{j=1}^{p} \alpha_{ij} X_j, \tag{2.1}$$

de manera que es minimitzi la pèrdua d'informació[4].

[2] Hi ha una altra definició de l'ACP buscant optimalitat respecte de mínims quadrats, però aquest treball s'enfocarà pel costat de la maximització de la variància. En [9], pp. 199-203, es troba explicada aquesta altra versió.

[3] Durant tot el capítol es treballa sota la hipòtesi que totes les columnes de X tenen variància positiva. Si això no és així aquella columna no és rellevant per l'estudi perquè és una variable constant, de manera que es pot obviar.

[4] Es tracta més profundament la idea d'informació donada per la matriu de dades a l'apartat 2.4, quan apareix el concepte d'inèrcia

Definició 2.3 (Components Principals)**.** *Sigui $X = (X_1, ..., X_p)$ una matriu de dades. Sigui $t \leq p$ i $\alpha_1, ..., \alpha_t \in \mathbb{R}^p$. Es defineixen les t primeres components principals com les combinacions lineals (2.1) per $i \leq t$ tals que compleixen:*

(A) $\mathbb{V}(\varphi_1) \geq \mathbb{V}(\varphi_2) \geq ... \geq \mathbb{V}(\varphi_t)$.

(B) $\forall j \in \{1, ..., t\}$, $Cov(\varphi_j, \varphi_k) = 0 \ \forall k < j$.

És clar que la clau perquè es satisfacin (A) i (B) rau en la tria correcta dels vectors de coeficients $\{\alpha_i\}_{i=1}^t$. El següent teorema determina quina és l'elecció que cal fer:

Teorema 2.4 (Construcció de les components principals I)**.** *Sigui X una matriu de dades de dimensions $n \times p$ amb matriu de covariàncies σ tal que els seus VAPs són tots diferents. Aleshores,*

(I) $\exists! \alpha_1 \in \mathbb{R}^p$ *tal que* $\alpha_1 \alpha_1^T = 1$ *i que maximitza* $\mathbb{V}(\varphi_1)$.

(II) $\forall 2 \leq k \leq t \leq p \ \exists! \alpha_k \in \mathbb{R}^p$ *tal que* $\alpha_k \alpha_k^T = 1$, $\alpha_k \alpha_j^T = 0 \ \forall j < k$ *i maximitza* $\mathbb{V}(\varphi_k)$*, prenent el màxim entre totes les possibles φ_k incorrelacionades amb $\varphi_j \ \forall j < k$.*

Demostració. El primer objectiu és maximitzar $\mathbb{V}(\varphi_1)$ sota la restricció $\|\alpha_1\| = \alpha_1 \alpha_1^T = 1$. A partir del Lema A.2 i posant $\alpha_1 = (\alpha_{11}, ..., \alpha_{1p})$ és fàcil veure que

$$\mathbb{V}(\varphi_1) = \mathbb{V}\left(\sum_{j=1}^{p} \alpha_{1j} X_j\right) = \alpha_1 \sigma \alpha_1^T. \tag{2.2}$$

Aquesta és la funció que es vol optimitzar. Cal tenir en compte la restricció $\alpha_1 \alpha_1^T = 1$, i per això es defineix

$$f(v) = v \sigma v^T + \lambda_1 (1 - v v^T), \tag{2.3}$$

on λ_1 és el multiplicador de Lagrange corresponent a aquesta condició. Per trobar un extrem respecte v d'aquesta funció cal derivar respecte v matricialment i igualar a zero:

$$\frac{\partial f}{\partial v}(v) = 2(\sigma - \lambda_1 I_p) v^T = 0. \tag{2.4}$$

Si $v \neq 0$ es té que

$$\sigma - \lambda_1 I_p = 0, \tag{2.5}$$

és a dir, que λ_1 és un VAP de σ, i per (2.4) v és el VEP de VAP λ_1.

Ara cal veure que aquesta parella VAP-VEP maximitza $\mathbb{V}(\varphi_1)$ i que $v = v_1$ és únic. Per a fer-ho s'ha de determinar exactament quin VAP és λ_1. En (2.2) s'ha vist que

$$\mathbb{V}(\varphi_1) = \alpha_1 \sigma \alpha_1^T,$$

però també que s'ha de complir $\alpha_1 = v_1$, de manera que

$$\mathbb{V}(\varphi_1) = \alpha_1 \sigma \alpha_1^T = v_1 \sigma v_1^T = \lambda_1.$$

Per tal que $\mathbb{V}(\varphi_1)$ sigui màxim, doncs, λ_1 ha de ser el VAP de valor màxim, i sota la restricció $v_1 v_1^T = 1$ efectivament hi ha un únic vector de coeficients que satisfà (I), donat que tots els VEPs de VAP λ_1 són múltiples de v_1.

Pel que fa a (II), la demostració es fa per $k = 2$ però és un raonament molt fàcil d'estendre al cas general. En aquest cas es busca maximitzar $\mathbb{V}(\varphi_2)$ sota dues condicions: $\alpha_2 \alpha_2^T = 1$ -el vector és unitari- i $\alpha_1 \alpha_2^T = 0$ -el nou vector és ortogonal a v_1-. Per tant, en la mateixa línia que s'ha seguit a (2.3), es defineix

$$f(v) = v \sigma v^T + \lambda_2 (1 - v v^T) - \mu v_1 v^T,$$

on v_1 és el vector que compleix (I) i λ_2, μ són els multiplicadors de Lagrange corresponents a cadascuna de les restriccions extres. Com que v_1 compleix (I), a més, es poden recuperar (2.4) i (2.5) per més endavant.
De moment, igual que abans, com que es cerca un màxim de f cal derivar respecte v i igualar a zero:

$$\frac{\partial f}{\partial v}(v) = 2(\sigma - \lambda_2 I_p)v^T - \mu v_1^T = 0. \tag{2.6}$$

Ja que $v_1 \neq 0$ es pot multiplicar a banda i banda per l'esquerra per v_1, arribant a

$$v_1\left[2(\sigma - \lambda_2 I_p)v^T - \mu v_1^T\right] = 2v_1\sigma v^T - 2\lambda_2 v_1 v^T - \mu v_1 v_1^T = 0,$$

i donat que $v_1 v^T = 0$ i $v_1 v_1^T = 1$, (2.6) és equivalent a

$$2v_1\sigma v^T - \mu = 0. \tag{2.7}$$

Recuperant (2.4),

$$(\sigma - \lambda_1 I_p)v_1^T = 0,$$

i prenent $v \neq 0$ es pot multiplicar per l'esquerra als dos costats de la igualtat per v i s'obté

$$v(\sigma - \lambda_1 I_p)v_1^T = v\sigma v_1^T - \lambda_1 v v_1^T = 0.$$

De nou, $vv_1^T = 0$, i per tant

$$v\sigma v_1^T = 0 \Leftrightarrow (v\sigma v_1^T)^T = v_1\sigma^T v^T \overset{(*)}{=} v_1\sigma v^T = 0,$$

on en $(*)$ s'aplica que σ és simètrica. Per (2.7), llavors, $\mu = 0$, i tornant a (2.6) es té

$$2(\sigma - \lambda_2 I_p)v^T = 0 \Leftrightarrow (\sigma - \lambda_2 I_p)v^T = 0,$$

arribant per tant a la mateixa situació que abans: λ_2 ha de ser un VAP de σ i $v = v_2$ el VEP de VAP λ_2. De nou, $\mathbb{V}(\varphi_2) = v_2\sigma v_2^T = \lambda_2$, de manera que per maximitzar $\mathbb{V}(\varphi_2)$ es necessita que λ_2 sigui el primer VAP que faci que v_2 compleixi les propietats que demana (II). Com que per hipòtesi tots els VAPs són diferents, λ_2 ha de ser el segon VAP més gran de σ, i llavors prenent v_2 el VEP unitari de VAP λ_2 es té que $v_1 v_2^T = 0$ perquè els VEPs de diferents VAPs són ortogonals. A més, $\mathbb{V}(\varphi_2) = \lambda_2 = \max_{j\in\mathcal{S}} \lambda_j$ on $\mathcal{S} := \{j : v_j v_j^T = 1, v_i v_j^T = 0\ \forall i < j\}$.

Repetint aquest procés per tot $k \leq t$ es té la construcció de les components principals per a variables que tenen matriu de covariàncies amb tots els VAPs diferents. □

Observació 2.5. Com ja s'ha comentat, per $k > 2$ es pot seguir amb un procés totalment anàleg, però cada cop amb més restriccions. Per exemple, per $k = 3$ es vol maximitzar $v\sigma v^T$ sota les condicions $vv^T = 1$, $v_1 v^T = v_2 v^T = 0$, de manera que es defineix la funció

$$f_3(v) = v\sigma v^T + \lambda_3(1 - vv^T) - \mu_1 v_1 v^T - \mu_2 v_2 v^T$$

i es poden fer servir les condicions trobades per v_1 i per v_2:

$$(\sigma - \lambda_1 I_p)v_1^T = 0, \qquad (\sigma - \lambda_2 I_p)v_2^T = 0,$$

per arribar a que $\mathbb{V}(\varphi_3)$ ha de ser el major VAP de σ diferent de λ_1 i λ_2 (λ_3, per tant) i que $v = v_3$.

Per matrius de covariàncies amb VAPs repetits es pot fer servir aquesta modificació del teorema. La seva prova es troba a l'annex A:

Teorema 2.6 (Construcció de les components principals II)**.** *Sigui X una matriu de dades de dimensions $n \times p$ amb matriu de covariàncies σ amb algun(s) VAP(s) repetit(s). Aleshores,*

(I) Si $\lambda_1 = ... = \lambda_r > \lambda_{r+1} > ... > \lambda_p$, $\exists! U \subset \mathbb{R}^p$ subespai vectorial de dimensió r tal que $U = Span[\alpha_1, ..., \alpha_r]$, $\forall i\ \alpha_i\alpha_i^T = 1$ i que maximitza $\mathbb{V}(\varphi_1) + ... + \mathbb{V}(\varphi_r)$.

(II) Si $\lambda_1 > ... > \lambda_s = ... = \lambda_{r+s-1} > \lambda_{r+s} > ... > \lambda_p$, $\exists! U \subset \mathbb{R}^p$ *subespai vectorial de dimensió* r *tal que* $U = Span[\alpha_s, ..., \alpha_{s+r-1}]$, $\forall i \in \{s, ..., s+r-1\}$ $\alpha_i \alpha_i^T = 1$, $\forall i \in \{s, ..., s+r-1\}$ $\forall j < s$ $\alpha_i \alpha_j^T = 0$ *i que maximitza* $\mathbb{V}(\varphi_s) + ... + \mathbb{V}(\varphi_{r+s})$, *prenent el màxim entre totes les possibles* $\varphi_s, ..., \varphi_{r+s}$ *incorrelacionades amb* φ_j $\forall j < s$.

Observació 2.7. El fet que hi hagi VAPs repetits a σ provoca que la base de $\mathbb{R}^p$ formada per les components principals sigui una mica pitjor, ja que com que els VEPs que la formen no sempre són de VAPs diferents no tenen perquè ser ortogonals entre ells, i per tant la base no té perquè ser ortogonal. Sempre és possible escollir una base ortonormal de U, però llavors les components principals no maximitzen necessàriament la variància, que és el principal objectiu d'aquest mètode.

Observació 2.8. A la pràctica és molt més usual fer servir el Teorema 2.4, degut a la poca probabilitat que la matriu σ tingui dos VAPs exactament iguals si es treballa amb dades reals.

2.3 Projecció dels individus a les noves coordenades

Un cop trobades les components principals cal projectar els individus de la matriu de dades sobre aquests nous eixos. Per a fer-ho convé fer un breu repàs d'àlgebra lineal, considerant el procés de l'ACP com un canvi de base de $\mathbb{R}^p$ amb la base canònica $\mathcal{B}_e$ a $\mathbb{R}^p$ amb la base $\mathcal{B}_u$ formada pels VEPs de σ. Sigui

$$\Psi : (\mathbb{R}^p, \mathcal{B}_e) \longrightarrow (\mathbb{R}^p, \mathcal{B}_u)$$

l'aplicació de canvi de base, amb matriu

$$U = \begin{pmatrix} u_{11} & \cdots & u_{p1} \\ \vdots & \ddots & \vdots \\ u_{1p} & \cdots & u_{pp} \end{pmatrix},$$

on $u_k = (u_{k1}, ..., u_{kp})$ és el VEP de VAP $\lambda_{(k)}$ (el VAP amb el k-èsim major valor). Amb això, la projecció d'un individu j que prengui valors $x_j = (x_{j1}, ..., x_{jp})$ és

$$\begin{aligned} \Psi(x_{j1}, ..., x_{jp}) = (x_{j1}, ..., x_{jp})U = (x_{j1}, ..., x_{jp}) \begin{pmatrix} u_{11} & \cdots & u_{p1} \\ \vdots & \ddots & \vdots \\ u_{1p} & \cdots & u_{pp} \end{pmatrix} \\ = \left(\sum_{k=1}^{p} x_{jk} u_{1k}, ..., \sum_{k=1}^{p} x_{jk} u_{pk} \right) = \left(\langle x_j, u_1 \rangle, ..., \langle x_j, u_p \rangle \right), \end{aligned}$$

on $\langle \cdot, \cdot \rangle$ representa el producte escalar usual de $\mathbb{R}^p$.

Amb aquest procés es poden calcular les coordenades de cadascun dels individus del núvol de punts original en la nova base de $\mathbb{R}^p$, que no és una base qualsevol. Als Teoremes 2.4 i 2.6 s'ha provat que $\forall 1 \leq t \leq p$, $u_1, ..., u_t$ són els vectors que maximitzen $\sum_{i=1}^{t} \mathbb{V}(\varphi_i)$, de manera que la projecció dels punts sobre les seves primeres t coordenades dona la representació en $\mathbb{R}^t$ que maximitza la variabilitat per aquella t, per a tot $1 \leq t \leq p$. Per tant, un cop aplicada Ψ, cal projectar sobre les primeres coordenades tots els punts (el nombre de coordenades, que de moment es denotarà per q, es concretarà amb els criteris de l'apartat 2.6) fent ús de l'aplicació

$$\pi_q : \mathbb{R}^p \longrightarrow \mathbb{R}^q, \quad \pi_q(y_1, ..., y_p) = (y_1, ..., y_q, 0, ..., 0).$$

En definitiva, doncs, donat un individu x_i representat per la fila i-èsima de la matriu de dades, la seva projecció en el subespai generat per $u_1, ..., u_q$ ve donada per

$$\pi_q(\Psi(x_i)) = \pi_q(\langle x_i, u_1 \rangle, ..., \langle x_i, u_p \rangle) = (\langle x_i, u_1 \rangle, ..., \langle x_i, u_q \rangle, 0, ..., 0).$$

2.4 Informació donada per cadascuna de les CP

Amb aquest apartat es busca justificar la importància de l'ACP normalitzada que es tracta més endavant. Perseguint aquest objectiu, cal conèixer la importància de cadascuna de les components principals, entenent el concepte "importància" com la quantitat de variabilitat que expliquen. A aquest fi es necessita definir el concepte d'inèrcia, una generalització de la variància a més dimensions.

Definició 2.9 (Inèrcia)**.** *Sigui $X = (X_1, ..., X_p)$ una matriu de dades, amb $X_1, ..., X_p$ variables. Es defineix la inèrcia de la matriu de dades com*

$$I_X = \sum_{i=1}^{p} \mathbb{V}(X_j).$$

Aquesta definició permet calcular el percentatge d'informació que proporciona cada component principal. Abans és necessària aquesta proposició, la demostració de la qual es troba a l'annex A:

Proposició 2.10. *Siguin $X \in \mathcal{M}_{n\times p}(\mathbb{R})$ una matriu de dades i σ la seva matriu de covariàncies. Sigui també $Spec(\sigma) = \{\lambda_1, ..., \lambda_p\}$. Aleshores es satisfà*

$$I_X = \sum_{j=1}^{p} \lambda_j.$$

A partir d'això es pot veure la proporció d'inèrcia explicada per cada component principal, que és important per tal de determinar quantes components principals s'agafen per a la visualització. La inèrcia explicada per la component principal associada al s-èsim VEP es denota per τ_s, i es troba segons

$$\tau_s = \frac{\mathbb{V}(\varphi_s)}{I_X} = \frac{\lambda_s}{I_X}.$$

També és útil la inercia explicada per les t primeres components principals: es pot definir simplement com

$$\tilde{\tau}_t = \sum_{i=1}^{t} \tau_i = \sum_{i=1}^{t} \frac{\lambda_i}{I_X} = \frac{\lambda_1 + ... + \lambda_t}{\lambda_1 + ... + \lambda_p},$$

i de manera similar, tornant a la definició original d'inèrcia, es pot quantificar la proporció d'inèrcia que representa cadascuna de les variables de la matriu de dades segons

$$\chi_s = \frac{\mathbb{V}(X_s)}{\sum_{i=1}^{p} \mathbb{V}(X_i)} = \frac{\mathbb{V}(X_s)}{I_X}. \tag{2.8}$$

Aquest valor es pot entendre com el "pes" que té cada variable en l'anàlisi de components principals, i juga un paper important ja que justifica la necessitat de treballar amb l'ACP normalitzada en determinades situacions.

2.5 ACP estandarditzada

L'expressió (2.8) té una conseqüència important. Com es mostra a l'expressió de χ_s el pes que té cada variable depèn de la seva variància, i per tant no totes tenen la mateixa importància en l'ACP. Aquest biaix pot afectar l'estudi, i per això és recomanable fer una normalització prèvia a l'anàlisi, especialment en situacions on les variables no estan mesurades en les mateixes unitats. Es pot definir el procés de la següent manera:

Definició 2.11 (Normalització de la matriu de dades)**.** *Siguin $X = (X_1, ..., X_p) \in \mathcal{M}_{n\times p}$ una matriu de dades, $\sigma \in \mathcal{M}_{p\times p}$ la seva matriu de covariàncies i $\mu_X \in \mathcal{M}_{n\times p}$ la matriu definida segons*

$$\mu_X = (\mu_{ij})_{1\leq i\leq n, 1\leq j\leq p}, \quad \mu_{ij} = \overline{X_j}.$$

Es defineix la matriu de dades normalitzada com la matriu $Z \in \mathcal{M}_{n\times p}$ resultant de l'aplicació matricial

$$\begin{array}{rccc} \Phi: & \mathcal{M}_{n\times p} & \longrightarrow & \mathcal{M}_{n\times p} \\ & X & \longmapsto & (diag(\sigma))^{-\frac{1}{2}}(X-\mu_X) \end{array} \tag{2.9}$$

és a dir, que $\forall (i,j) \in \{1,...,n\}\times\{1,...,p\}$ es té

$$z_{ij} = \frac{x_{ij} - \overline{X_j}}{\sqrt{\mathbb{V}(X_j)}}.$$

Observació 2.12. A la Proposició A.1(IV) s'han calculat totes les entrades de la matriu HX, concloent que a la posició i,j hi ha el valor

$$x_{ij} - \frac{1}{n}\sum_{k=1}^{n} x_{kj} = x_{ij} - \overline{X_j},$$

de manera que $\forall j \in \{1,...,p\}$ es té

$$Z_j = \frac{(HX)_j}{\sqrt{\mathbb{V}(X_j)}}.$$

La següent proposició justifica la utilitat de definir aquesta matriu normalitzada. La prova es troba a l'annex A:

Proposició 2.13 (Normalitat de les columnes de la matriu normalitzada)**.** *Siguin X una matriu de dades i $Z = \Phi(X) = (Z_1,...,Z_p)$ la corresponent matriu normalitzada. Aleshores es compleix:*

(I) Les columnes de Z són centrades.

(II) Les columnes de Z tenen variància 1.

Observació 2.14. D'aquesta proposició es poden treure dues conclusions importants respecte la normalització de l'ACP. Per una banda, com que totes les variables tenen variància igual a 1 no n'hi haurà cap que domini sobre les altres a l'hora de crear les visualitzacions, de manera que es veuran igualment representades totes les variables. No obstant això, el fet de normalitzar les columnes de X provoca que el núvol de punts canviï de forma, modificant lleugerament els resultats sobretot a nivell de valors anòmals. En qualsevol cas aquesta normalització és una pràctica molt utilitzada, perquè els beneficis en termes de qualitat de la representació són molt més importants que aquest canvi en la forma del núvol de punts, i normalitzant les dades s'eviten també alguns problemes que donen els algoritmes de classificació que es tractaran més endavant.

L'objectiu ara és replicar el procés fet al Teorema 2.4 però per a la matriu normalitzada. En (2.1) es veu que la component principal φ_1 sobre X, denotada d'ara endavant per φ_1^X, es defineix escollint $\alpha_1 := (\alpha_{11},...,\alpha_{1p})$ de manera que

$$\varphi_1^X = \sum_{j=1}^{p} \alpha_{1j} X_j$$

compleixi unes determinades propietats, entre elles maximitzar

$$\mathbb{V}(\varphi_1^X) = \alpha_1 \sigma \alpha_1^T, \tag{2.10}$$

on

$$\sigma_{ij} = \begin{cases} \mathbb{V}(X_i) & \text{si } i = j, \\ Cov(X_i, X_j) & \text{altrament.} \end{cases}$$

El que es busca ara és repetir aquesta construcció sobre Z. Sigui $\tilde{\sigma}$ la matriu de covariàncies de Z, i suposant que té tots els VAPs diferents es pot aplicar el Teorema 2.4(I): $\exists! \beta_1 \in \mathbb{R}^p$ tal que $\beta_1\beta_1^T = 1$ i que maximitza $\mathbb{V}(\varphi_1^Z)$. Aplicant la mateixa igualtat que a (2.10) es té

$$\mathbb{V}(\varphi_1^Z) = \beta_1 \tilde{\sigma} \beta_1^T$$

i per tant, seguint el procediment del teorema, cal diagonalitzar $\tilde{\sigma}$. Amb els següents resultats s'obté una relació entre $\tilde{\sigma}$ i X, però per a fer-ho es requereixen primer dues definicions:

Definició 2.15 (Coeficient de correlació)**.** *Siguin X i Y dues variables. Es calcula el seu coeficient de correlació com*

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{\mathbb{V}(X)}\sqrt{\mathbb{V}(Y)}} \in [-1,1].$$

Definició 2.16 (Matriu de correlacions)**.** *Siguin $X_1, ..., X_n$ variables. Es defineix la matriu de correlacions entre aquestes variables com la matriu*

$$R = (r_{ij}), \quad r_{ij} = \begin{cases} 1 & \text{si } i = j, \\ \rho(X_i, X_j) & \text{altrament.} \end{cases}$$

En cas que les variables $X_1, ..., X_n$ siguin columnes d'una matriu X, es denota aquesta matriu de correlacions com R_X.

Teorema 2.17 (Diagonalització de la matriu de correlacions en l'ACP normalitzada)**.** *Sigui $X = (X_1, ..., X_p) \in \mathcal{M}_{n\times p}$ una matriu de dades i $Z = \Phi(X)$ la corresponent matriu normalitzada. Seguint les notacions anteriors, $\tilde{\sigma} = R_X$.*

Demostració. Sense perdre generalitat es pot suposar que X és centrada, ja que $Cov(X+a, Y+b) = Cov(X,Y)$ per $a, b \in \mathbb{R}$. Amb això, $\forall j \in \{1, ..., p\}$ es té

$$Z_j = \frac{X_j}{\sqrt{\mathbb{V}(X_j)}},$$

i per tant:

- Si $i = j$, $\tilde{\sigma}_{ij} = \mathbb{V}(Z_i) = 1$ perquè les columnes de Z són centrades.
- Si $i \neq j$,

$$\tilde{\sigma}_{ij} = Cov(Z_i, Z_j) = Cov\left(\frac{X_i}{\sqrt{\mathbb{V}(X_i)}}, \frac{X_j}{\sqrt{\mathbb{V}(X_j)}}\right) = \frac{Cov\,(X_i, X_j)}{\sqrt{\mathbb{V}(X_i)}\sqrt{\mathbb{V}(X_j)}} = \rho(X_i, X_j).$$

□

Aquest teorema, junt amb els Teoremes 2.4 i 2.6, són les bases pràctiques de l'ACP. Tant si s'han normalitzat les dades com si no es pot fer servir el procés del Teorema 2.4 (o el Teorema 2.6 si és necessari) per a construir les components principals. Cal, però, tenir en compte que si s'ha estandarditzat X s'han buscar els VEPs i els VAPs de la matriu de correlacions mentre que si no s'ha fet es necessita diagonalitzar la matriu de covariàncies.

2.6 Selecció de components principals

Com s'ha comentat anteriorment, l'ACP és una tècnica de reducció de la dimensió d'una matriu de dades. Amb el procés vist als Teoremes 2.4 i 2.6 es poden definir p components principals, cadascuna d'elles associada a un VEP de la matriu de covariàncies, però construir totes les components principals no genera cap benefici a nivell de visualització de les dades perquè la dimensió és la mateixa. Per això, cal decidir quantes components principals fer servir per a representar els individus, buscant mantenir un equilibri entre la complexitat de la representació -relacionada amb el nombre de components principals preses- i la informació que aporta aquesta representació, és a dir, el valor $\tilde{\tau}_t$. Per a fer-ho es pot fer ús de diversos criteris:

a. *Last elbow rule* o criteri del bastó trencat. Aquest criteri fa ús d'un *screeplot*, un gràfic que representa a l'eix d'abscisses els valors propis, ordenats en funció del seu valor, i a l'eix d'ordenades els valors d'aquests VAPs. Aleshores, fixat un valor $M > 0$ prou gran es conserven

$$\max\{t \in \{2, ..., p\} : \lambda_t - \lambda_{t-1} > M\}$$

components principals.

b. Criteri de variabilitat acumulada. Aquest criteri usa únicament el valor $\tilde{\tau}_t$, fixant el nombre de components principals a retenir com

$$\min\{t \in \{1, ..., p\} : \tilde{\tau}_t \geq \gamma\},$$

on γ és un paràmetre fixat per l'investigador (sovint $\gamma = 0.8$).

c. Criteri de Kaiser. Sigui

$$\overline{\lambda} = \frac{I_X}{p}$$

la mitjana dels valors dels VAPs. Aleshores, segons el criteri de Kaiser, el nombre de components principals que cal fer servir és

$$t = \#\{\lambda \in \{\lambda_1, ..., \lambda_p\} : \lambda > \overline{\lambda}\},$$

és a dir, es prenen tantes components principals com VAPs hi hagi amb valors superiors a la mitjana[5].

2.7 El biplot: estudi simultani d'individus i variables

Fins ara s'ha tractat l'ACP exclusivament des del punt de vista de l'estudi dels individus de la matriu de dades, però es pot fer un anàlisi similar per les seves variables. Aquest doble estudi es fonamenta en la tècnica del biplot, que permet representar en uns mateixos eixos els individus i les variables (això és, les files i les columnes de la matriu de dades). Al seu torn, aquesta tècnica troba el seu principi en la descomposició en valors singulars de les matrius, que es tracta a continuació.

2.7.1 Descomposició en valors singulars

Des dels cursos bàsics d'àlgebra se sap que els conceptes de valor i vector propi estan associats únicament a matrius quadrades. La descomposició en valors singulars busca una generalització d'aquests conceptes, i per extensió de la diagonalització de matrius, a matrius $A \in \mathcal{M}_{m\times n}$, amb $m > n$, a partir d'una descomposició de la forma

$$A = U\Sigma V^T$$

on $U \in \mathcal{M}_{m\times m}$ i $V \in \mathcal{M}_{n\times n}$ són matrius ortogonals i $\Sigma \in \mathcal{M}_{m\times n}$ és una matriu diagonal.
Per tal d'arribar a aquesta descomposició, la següent definició és bàsica:

Definició 2.18 (Valors singulars)**.** *Sigui $A \in \mathcal{M}_{n\times p}$ una matriu. Siguin $\lambda_1, ..., \lambda_p$ els valors propis de la matriu $A^TA \in \mathcal{M}_{p\times p}$. Es defineixen els valors singulars de la matriu A com*

$$\sigma_j = \sqrt{\lambda_j}, \quad \forall j \in \{1, ..., p\}.$$

És comú ordenar els valors singulars de manera decreixent. Suposant que es tenen $r < p$ VAPs no nuls, $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > \sigma_{r+1} = ... = \sigma_p = 0$.

Observació 2.19. $(A^TA)^T = A^T(A^T)^T = A^TA$, de manera que A^TA és una matriu simètrica que com a tal diagonalitza amb VAPs reals no negatius. Per tant, σ_j està ben definit $\forall j$.

A partir d'aquesta definició ja es pot enunciar el teorema de descomposició en valors singulars, necessari per construir el biplot:

Teorema 2.20 (Descomposició en valors s†ingulars)**.** *Sigui $X \in \mathcal{M}_{n\times p}$, $n \geq p$, una matriu qualsevol. Aleshores, existeixen matrius $U \in \mathcal{M}_{n\times n}$ ortogonal, $\Sigma \in \mathcal{M}_{n\times p}$ diagonal i $V \in \mathcal{M}_{p\times p}$ ortogonal tals que*

$$X = U\Sigma V^T.$$

[5]Hi ha estudis (veure [6], pàg. 7), que consideren $c\overline{\lambda}$ com a llindar, sovint c=0.7. Aquestes variacions funcionen millor amb ACP no normalitzades, però no són d'ús generalitzat.

Demostració. La prova segueix dos passos: en primer lloc la construcció d'aquestes tres matrius, i després la comprovació de les propietats demanades.
Fent servir la Definició 3.1 es pot construir Σ segons

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_p \\ 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

V també es pot trobar de manera bastant natural. Com que $X^TX \in \mathcal{M}_{p\times p}$ és una matriu simètrica ha de diagonalitzar amb valors propis reals i no negatius, és a dir, que $\exists V, D \in \mathcal{M}_{p\times p}$ tals que

$$X^TX = VDV^T.$$

En concret, D és la matriu diagonal que té a la posició d_{ii} l'i-èsim valor propi en ordre decreixent i V és la matriu que té a la columna j-èsima el vector propi de VAP λ_j. Posant $V = (v_1, ..., v_n)$ es troba la segona matriu de la descomposició.

Per últim cal construir U. Si es consideren $k \leq n$ VAPs no nuls $\sigma_1, ..., \sigma_k$ de la matriu X^TX, es pot definir

$$u_i = \frac{Xv_i}{\sigma_i} \quad \forall i \in \{1, ..., k\}. \tag{2.11}$$

Aquests vectors són ortonormals si els $\{v_i\}$ ho són. En efecte, per una banda

$$\|u_i\|^2 = \left\| \frac{Xv_i}{\sigma_i} \right\|^2 = \frac{1}{\sigma_i^2}\|Xv_i\|^2 \overset{(*)}{=} \frac{1}{\lambda_i} v_i^T X^T X v_i \overset{(**)}{=} \frac{1}{\lambda_i} v_i^T \lambda_i v_i = \frac{\lambda_i}{\lambda_i}\|v_i\|^2 = 1,$$

on en $(*)$ s'usa que $\|x\|^2 = x^Tx$ i en $(**)$ que v_i és VEP de VAP λ_i de X^TX, i per altra banda si $i \neq j$ es té que

$$u_i^T u_j = \left(\frac{Xv_i}{\sigma_i}\right)^T \frac{Xv_j}{\sigma_j} = \frac{1}{\sigma_i\sigma_j}(Xv_i)^T Xv_j = \frac{1}{\sigma_i\sigma_j} v_i^T X^T X v_j = \frac{1}{\sigma_i\sigma_j} v_i^T \lambda_j v_j = 0,$$

donat que $v_i^T v_j = 0$. Falta, però, completar la matriu U, ja que els $\{u_i\}$ només en formen les primeres k columnes. Per a fer-ho calen $n-k$ vectors tals que $\{u_1, ..., u_k, u_{k+1}, ..., u_n\}$ formin una base ortogonal de $\mathbb{R}^n$, que es poden trobar a partir de vectors $w_{k+1}, ..., w_n$ qualssevol tals que $\{u_1, ..., u_k, w_{k+1}, ..., w_n\}$ siguin linealment independents, i per tant una base de $\mathbb{R}^n$. Per acabar, seguint el procés d'ortonormalització de Gram-Schmidt, es converteix aquest conjunt de vectors en un conjunt ortonormal.

Arribats a aquest punt només falta comprovar que es compleixen les condicions demanades. En primer lloc, per la pròpia construcció, és clar que tant U com V són matrius ortogonals, i Σ és diagonal. Per tant, només cal comprovar que $X = U\Sigma V^T$, o equivalentment que $XV = U\Sigma$, perquè com que V és una matriu ortogonal es té que $V^T = V^{-1}$. Més amunt s'ha dit que hi ha $k \leq n$ VAPs no nuls $\lambda_1, ..., \lambda_k$ de X^TX, i per tant pels VEPs des del $k+1$ fins al n es satisfà

$$Xv_i = \sigma_i v_i = 0v_i = 0.$$

Amb això,

$$XV = X\begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} = \begin{pmatrix} Xv_1 & \cdots & Xv_n \end{pmatrix} \overset{(2.11)}{=} \begin{pmatrix} \sigma_1 u_1 & \cdots & \sigma_k u_k & 0 & \cdots & 0 \end{pmatrix}$$

$$= \begin{pmatrix} u_1 & \cdots & u_k & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \sigma_k & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} = U\Sigma, \qquad (2.12)$$

de manera que efectivament U, V i Σ compleixen les propietats requerides. □

Abans de veure l'aplicació d'aquest teorema a les representacions gràfiques cal fer dues observacions:

Observació 2.21 (No unicitat de la DVS)**.** La matriu V no té perquè ser única. Per exemple, en situacions on X^TX tingui VAPs repetits, els VEPs d'aquest VAP es poden posar en diferent ordre, de manera que V perd la condició d'unicitat. Imposant que V sigui ortogonal però no ortonormal, cada columna de V pot ser substituïda per un múltiple d'aquesta, fent de nou que no sigui única.
Per altra banda, U tampoc té perquè ser única: de fet, només ho és si el nombre de VAPs no nuls de X^TX és n -si no depèn dels vectors que s'afegeixen per crear la base ortonormal- i, a més, que tots els VAPs de la matriu són diferents -si no, l'ordre dels VEPs d'aquell VAP en les columnes de la matriu també trenca aquesta unicitat.

Observació 2.22. Aquesta descomposició en valors singulars és una extensió del procés de diagonalització de les matrius quadrades. Si $n = p$, aleshores V és la matriu que té per columnes els VEPs ortonormals de X, $U = V^T$ (i totes dues són matrius ortonormals), i Σ és la matriu que té a la diagonal els VAPs de X, en l'ordre corresponent.

2.7.2 Aplicació de la SVD a la representació dels biplots

A partir de (2.12) s'obtenen les coordenades per representar les files i les columnes de X: les primeres venen donades per les n files de $U\Sigma$ i les segones són les files de V. En general, per tal de trobar una solució biplot, es busca representar alhora les matrius $A := U\Sigma^\alpha$ i $B := V\Sigma^{1-\alpha}$, $0 \leq \alpha \leq 1$, de tal manera que $AB^T = U\Sigma^\alpha(V\Sigma^{1-\alpha})^T = U\Sigma^\alpha(\Sigma^T)^{1-\alpha}V^T = X$. El valor d'$\alpha$ dona la qualitat de la representació de files i columnes: quan $\alpha = 1$ es té que $A = U\Sigma$ és la representació més fidel dels individus, si $\alpha = 0$ passa el mateix per les columnes (les variables). És una pràctica habitual fer servir el valor $\alpha = \frac{1}{2}$.

J. C. Gower va proposar un mètode alternatiu basat únicament en la construcció de components principals. Com s'ha tractat al capítol anterior, $Y = XT$ és una manera de representar els individus en uns eixos que maximitzen la variància projectada. A partir d'aquesta descomposició es pot representar una variable X_j com els punts

$$x_j(\alpha_j) = (0, ..., \alpha_j, ..., 0), \quad m_j \leq \alpha_j \leq M_j,$$

posant

$$m_j := \min_{i\in\{1,..,n\}} X_{ij}, \quad M_j := \max_{i\in\{1,..,n\}} X_{ij}.$$

Aleshores, la variable X_j es representa a partir del vector $x_j(\alpha)T$.

3 Clústering

L'inici d'aquest capítol marca el canvi d'objectiu en l'estudi de la matriu de dades. El capítol 2 s'ha centrat en la reducció de la dimensió per a poder-les visualitzar de manera senzilla, mentre que d'ara endavant el principal focus és la classificació dels individus en diferents grups, usant mètodes de classificació o clústering.
Els mètodes de clústering es diferencien en dos tipus, jeràrquics i no jeràrquics, sent la principal diferència la forma de construir els clústers: els primers els formen a partir de la fusió o la divisió de dos grups ja construits, mentre que els segons ho fan a partir de particions del conjunt sencer dels individus. Tots dos tenen avantatges i inconvenients, que es poden aprofitar en el que s'anomena mètode seqüencial. Aquest és l'enfoc que pren la part pràctica del treball, construint una classificació en dos passos: començant amb algoritmes jeràrquics per consolidar la partició amb algoritmes no jeràrquics. Això es tracta amb molta més profunditat al capítol cinquè, on s'explica el procés seguit amb la base de dades dels jugadors de la Liga EBA. Aquest capítol només es focalitza sobre la vessant matemàtica dels dos tipus de mètodes.
Les referències principals per aquest capítol són [4], [9], [11] i [14], i [21] té informació sobre el *silhouette plot*, el mètode usat per determinar la qualitat de la classificació.

Abans de començar amb els mètodes de clústering cal definir què són les variàncies intra- i entre-clústers per entendre millor l'objectiu d'aquests algoritmes:

Definició 3.1 (Variàncies intra i entre clúster)**.** *Sigui una població dividida en H grups diferents i X la variable que es vol estudiar. Siguin $\overline{X}$ la mitjana global, $\overline{X}_h$ la mitjana de la variable en cadascun dels grups $h \in \{1, ..., H\}$ i V_h la variància mostral en cada grup. Sigui també n la mida de la població, i n_h la mida de cada grup. Aleshores, es defineixen:*

1. *La variància intra-classes com*
$$V_{intra} = \sum_{h=1}^{H} \frac{n_h}{n} V_h.$$
2. *La variància entre classes com*
$$V_{entre} = \sum_{h=1}^{H} \frac{n_h}{n} \left(\overline{X}_h - \overline{X}\right)^2.$$

A partir d'aquestes definicions es pot trencar la variància global de tots els individus en la suma d'aquestes variàncies intra-grups i entre-grups, com mostra la Proposició A.8 de l'annex A. Portant-ho al terreny del clústering, es pot trencar la variància de cada variable de la matriu de dades com la suma de les variàncies dins de cada clúster i la variància entre els clústers. L'objectiu de tots els mètodes de clústering, tant jeràrquics com no jeràrquics, és aconseguir crear grups que redueixin al mínim possible V_{intra}, és a dir, construir grups el màxim d'homogenis que es pugui. Com que la variància total és constant, aquesta minimització implica que V_{entre} creixi, és a dir, que els grups seran més heterogenis entre ells.

3.1 Clústering jeràrquic

El clústering jeràrquic[6] és la primera fase de la classificació seqüencial. Aquests mètodes es basen en una successió de particions $\{C_i\}_{i\geq 1}$ del conjunt Ω del total dels elements, que van ajuntant grups d'individus que estiguin a prop. La definició d'aquesta proximitat és el que dona varietat a aquests mètodes, ja que diferents formes de calcular distàncies produeixen diferents agrupacions. Per a poder treballar-hi, el primer que cal fer és definir el concepte de jerarquia indexada:

[6]Els mètodes jeràrquics es poden dividir en dos subgrups, en funció de la monotonia de la successió de particions. Si cada partició és formada per unions d'elements de la partició anterior es diu que el mètode és aglomeratiu, mentre que si els conjunts de la partició j són formats per trossos de conjunts de particions anteriors el mètode és divisiu. Els dos tipus de mètodes són interessants i tenen propietats a estudiar, però per coherència amb el codi de la part pràctica, desenvolupat seguint algoritmes aglomeratius, només s'estudien aquests.

Definició 3.2 (Jerarquia indexada)**.** *Sigui Ω un conjunt. Es defineix una jerarquia indexada sobre Ω com una parella (C, α) tals que $C \subseteq \mathcal{P}(\Omega)$ i compleixen:*

(A) $\forall c, c' \in C$, $c \cap c' \in \{c, c', \emptyset\}$.

(B) $\forall c \in C$, $c = \bigcup_{c' \in \mathcal{C}} c'$, *on* $\mathcal{C} = \{c' \in C : c' \subseteq c\}$

(C) $\Omega = \bigcup_{c \in C} c$.

(D) $\alpha : C \longrightarrow \mathbb{R}_+$ *és una aplicació sobre els reals positius complint:*

(a) $\alpha(i) = 0$ $\forall i \in C$.

(b) $\forall c \subset c'$, $\alpha(c) \leq \alpha(c')$.

Aleshores, es diu que els elements de C són els clústerings i α s'anomena índex.

Observació 3.3. α es pot veure com una mesura de la homogeneïtat dels clústers. Si c és un clúster, valors grans d'$\alpha(c)$ indiquen una heterogeneïtat alta en c, mentre que valors petits de la funció representen grups molt homogenis.

Observació 3.4. D'ara endavant, per distingir les unions d'elements dels clústerings, es denotaran aquestes últimes amb el símbol $\vee$. Així, $c_1 \cup c_2$ denotarà l'unió de dos elements (no exclou que sigui la unió de dos clústers), i $c_1 \vee c_2$ indicarà que c_1 i c_2 són clústers i que $c_1 \vee c_2$ és un clústering sobre Ω. Sempre s'entendrà, a més, que $\vee$ denota unions disjuntes.

3.2 Geometries ultramètriques i dendrogrames

En aquest apartat s'introdueix el concepte de dendrograma, una representació gràfica molt útil per visualitzar els processos de clústering jeràrquic. Per a fer-ho cal definir primer què són els espais ultramètrics:

Definició 3.5 (Espai ultramètric)**.** *Un espai ultramètric és una parella (Ω, u) on Ω és un conjunt finit i u és una aplicació*

$$u : \quad \Omega \times \Omega \quad \longrightarrow \quad \mathbb{R}_+$$

que compleix les següents tres propietats: $\forall i, j, k \in \Omega$,

(A) $u(i, j) \geq 0$.

(B) $u(i, j) = u(j, i)$.

(C) $u(i, k) \leq \sup\{u(i, j), u(j, k)\}$.

En aquest cas es diu que u és una distància ultramètrica.

Observació 3.6. Tota distància ultramètrica és mètrica. En efecte,

$$\sup\{u(i,j), u(j,k)\} = \frac{u(i,j) + u(j,k) + |u(i,j) - u(j,k)|}{2} \leq$$

$$\leq \frac{u(i,j) + u(j,k) + |u(i,j)| + |u(j,k)|}{2} = u(i,j) + u(j,k)$$

i per tant, si u és ultramètrica,

$$u(i,k) \leq \sup\{u(i,j), u(j,k)\} \leq u(i,j) + u(j,k).$$

Es pot definir també la noció de dendrograma:

Definició 3.7 (Dendrograma)**.** *Sigui Ω un conjunt finit amb cardinal n. Un dendrograma o arbre ultramètric és un graf connex i sense cicles amb un punt anomenat arrel i n punts extrems equidistants de l'arrel.*

I la de triangle ultramètric:

Definició 3.8 (Triangle ultramètric)**.** *Sigui* (Ω, u) *un espai ultramètric i siguin* i, j, $k \in \Omega$ *tres punts. Es diu que el triangle format per aquests és ultramètric si es compleix*

$$u(i,j) \leq u(i,k) = u(j,k),$$

suposant que el costat més curt és el que uneix i *i* j*.*

Aquests dos conceptes són els que permeten representar els processos de clústering, gràcies al següent teorema:

Teorema 3.9 (Representació dels dendrogrames)**.** *Sigui* (Ω, u) *un espai ultramètric. Aleshores, es pot representar com un arbre ultramètric que tingui per punts extrems els elements de* Ω*.*

Demostració. Sigui un triangle $\{i,j,k\}$ d'elements de Ω, amb ij com a costat més curt. Es denota per γ_{ab} el node on s'ajunten els punts $a, b \in \Omega$ i $u(a,b)$ l'altura del dendrograma on s'ajunten, que es pot provar que és una distància ultramètrica de manera senzilla. Es pot definir una relació d'ordre total $\widetilde{\leq}$ en el conjunt de nodes: $\gamma_{ab} \widetilde{\leq} \gamma_{cd} \Leftrightarrow a$ i b s'uneixen abans que c i d en el dendrograma. Al triangle que s'ha escollit, $\gamma_{ij} \widetilde{\leq} \gamma_{ik} \widetilde{=} \gamma_{jk}$, perquè primer s'uneixen i i j i després el conjunt $\{i,j\}$ s'uneix amb k, motiu pel qual es dona la darrera igualtat. Per tant, $u(i,k) = u(j,k) = u(i,j) + h$, on $h > 0$ és la distància vertical entre el node on s'uneixen i i j i on s'uneixen el conjunt $\{i,j\}$ amb l'element k. Això prova que $\{i,j,k\}$ és un triangle ultramètric, i estenent això al conjunt de tots els punts de Ω es prova que es pot representar aquest espai com un dendrograma. □

Una altra propietat important queda palesa en la següent proposició, que demostra que donat un clústering $\Omega = \bigcup_{i=1}^{m} c_i$, el fet d'unir dos clústers per moure's al següent pas de l'algoritme no trenca el fet que u segueixi sent una distància ultramètrica. Abans, però, cal un lema previ que assegura que la distància que la nova distància conserva les propietats d'ultramètrica:

Lema 3.10. *En un espai ultramètric, tot triangle és ultramètric.*

Demostració. Siguin (Ω, u) un espai ultramètric i i, j, $k \in \Omega$. Prenent de nou ij com el costat més curt, és evident que la primera desigualtat de la Definició 3.8 es compleix.
Per provar la igualtat, s'usa la tercera propietat de les distàncies ultramètriques:

$$u(i,k) \leq \sup\{u(i,j), u(j,k)\}$$

i

$$u(j,k) \leq \sup\{u(i,k), u(i,j)\}.$$

Ara, com que el costat curt és ij, $\sup\{u(i,j), u(j,k)\} = u(j,k)$ i $\sup\{u(i,k), u(i,j)\} = u(i,k)$, i per tant

$$u(i,k) \leq u(j,k)$$

i

$$u(j,k) \leq u(i,k),$$

provant la igualtat i per tant que $\{i,j,k\}$ és un triangle ultramètric. □

Proposició 3.11 (Distància ultramètrica després d'ajuntar grups)**.** *Sigui*

$$\Omega = c_1 \vee ... \vee c_m$$

un clústering sobre un espai ultramètric (Ω, u)*. Siguin* c_i *i* c_j *els dos clústers més propers, és a dir, tals que* $u(c_i, c_j)$ *és mínim. Llavors, es pot definir una distància ultramètrica* u' *sobre el clústering*

$$\Omega = c_1 \vee ... \vee (c_i \cup c_j) \vee ... \vee c_m.$$

Demostració. Es defineix primer u' segons

$$u'(c_k, c_i \cup c_j) = u(c_k, c_i) = u(c_k, c_j) \quad \text{per } k \neq i, j,$$

$$u'(c_l, c_m) = u(c_l, c_m) \quad \text{per } \{i, j\} \cap \{l, m\} = \emptyset.$$

Pel lema anterior, $u'(c_k, c_i \cup c_j)$ està ben definida: si $u(c_i, c_j)$ és mínima, aleshores el triangle $\{c_i, c_j, c_k\}$ és ultramètric, és a dir, $u(c_i, c_j) \leq u(c_i, c_k) = u(c_j, c_k)$.
Ara només queda comprovar les propietats de distància ultramètrica, que són:

- $u'(c_a, c_b) \geq 0 \ \forall c_a, c_b \subseteq \Omega$.
- $u'(c_a, c_b) = u'(c_b, c_a) \ \forall c_a, c_b \subseteq \Omega$.
- $u(c_a, c_d) \leq \sup\{u(c_a, c_b), u(c_b, c_d)\} \ \forall c_a, c_b, c_d \subseteq \Omega$.

És evident que escollint c_a, c_b i c_d tals que $\{c_a, c_b, c_d\} \cap \{c_i, c_j\} = \emptyset$, u' compleix totes les propietats perquè pren els mateixos valors que prenia u, que és ultramètrica. Es pren doncs un triangle $\{c_a, c_b, c_i \cup c_j\}$, i

- $u'(c_a, c_i \cup c_j) = u(c_a, c_i) \geq 0$, i igual per b.
- $u'(c_a, c_i \cup c_j) = u(c_a, c_i) = u(c_i, c_a) = u'(c_i \cup c_j, c_a)$.
- La tercera propietat s'ha de comprovar de dues maneres: amb $c_i \cup c_j$ jugant el paper de c_a (o de c_d, és simètric) i fent la funció de c_b.
  - $u'(c_a, c_d) = u(c_a, c_d) \leq \sup\{u(c_a, c_i), u(c_i, c_d)\} = \sup\{u'(c_a, c_i \cup c_j), u'(c_i \cup c_j, c_d)\}$.
  - $u'(c_a, c_i \cup c_j) = u(c_a, c_i) \leq \sup\{u(c_a, c_b), u(c_b, c_i)\} = \sup\{u'(c_a, c_b), u'(c_b, c_i \cup c_j)\}$.

Amb això queda demostrat que u' és també distància ultramètrica. □

Per acabar la secció es presenta el teorema que permet relacionar les jerarquies indexades que defineixen els clústerings amb les seves representacions en forma de dendrogrames.

Teorema 3.12 (Relació entre jerarquies indexades i geometries ultramètriques)**.** *Sigui Ω un conjunt finit i (C, α) una jerarquia indexada sobre Ω. Aleshores, es pot definir una distància ultramètrica u sobre Ω. Recíprocament, tot espai ultramètric (Ω, u) defineix una jerarquia indexada (C, α).*

Demostració. Sigui (C, α) una jerarquia indexada. Es defineix

$$u(i, j) = \alpha(c_{ij}),$$

on

$$c_{ij} = \bigcap_{c \in C : i, j \in c} c \tag{3.1}$$

és el menor clúster que conté i i j. Cal veure que u és una distància ultramètrica:

- $u(i, j) \geq 0$. Com $u(i, j) = \alpha(c_{ij})$ i α és una funció índex, $\alpha(c_i) = 0 \ \forall i \in C$ i $\forall c \subset c', \alpha(c) \leq \alpha(c')$. Si hi hagués alguna parella d'elements i, j tals que $u(i, j) = \alpha(c_{ij}) < 0$, com que $\{i\} \subseteq c_{ij}$, aleshores $\alpha(c_i) \leq \alpha(c_{ij}) < 0$, incomplint la primera condició de la definició d'α. Per tant, $u(i, j) \geq 0 \ \forall i, j \in C$.
- $u(i, j) = u(j, i)$. Evidentment, $c_{ij} = c_{ji}$, i per tant $u(i, j) = \alpha(c_{ij}) = \alpha(c_{ji}) = u(j, i)$.
- $u(i, j) \leq \sup\{u(i, k), u(k, j)\}$. Sigui un triangle $\{i, j, k\}$, i els clústers c_{ik} i c_{jk} definits com en (3.1). Com que (C, α) és una jerarquia indexada, en particular ha de complir que $\forall c, c' \in C$, $c \cap c' \in \{c, c', \emptyset\}$. Com que $c_{ik} \cap c_{jk} \neq \emptyset$ ja que $k \in c_{ik} \cap c_{jk}$, es té bé que $c_{ik} \subset c_{jk}$ o bé que $c_{jk} \subset c_{ik}$:

- $c_{ik} \subset c_{jk} \Rightarrow \{i,k\} \in c_{jk} \Rightarrow c_{ij} \subset c_{jk} \Rightarrow u(i,j) = \alpha(c_{ij}) \leq \alpha(c_{jk}) = u(j,k)$ $\leq \sup\{u(i,k), u(j,k)\}$.
- $c_{jk} \subset c_{ik} \Rightarrow \{j,k\} \in c_{ik} \Rightarrow c_{ij} \subset c_{ik} \Rightarrow u(i,j) = \alpha(c_{ij}) \leq \alpha(c_{ik}) = u(i,k)$ $\leq \sup\{u(i,k), u(j,k)\}$.

Per tant, en qualsevol cas, es compleix la propietat ultramètrica.

Amb això queda demostrat que donada una jerarquia indexada sobre Ω es pot definir una distància ultramètrica u sobre el mateix conjunt.

Per provar la implicació contrària cal construir un algoritme. Sigui ara (Ω, u) un espai ultramètric, i es vol generar una jerarquia indexada fent ús de la distància ultramètrica u. Aquest procediment s'anomena *algoritme fonamental de classificació*:

1. Es considera la partició $\Omega = c_1 \vee ... \vee c_n$ i es calcula $u(c_i, c_j)\ \forall i, j$.
2. S'ajunten els clústers c_i i c_j tals que $u(c_i, c_j) = \min_{A,B} u(c_A, c_B)$, i es defineix una nova distància u' seguint la Proposició 3.11, que garanteix la conservació de la propietat ultramètrica.
3. A partir de la nova partició $\Omega = c_1 \vee ... \vee (c_i \cup c_j) \vee ... \vee c_n$ es repeteix el pas 2, ajuntant cada cop els dos clústers que tinguin $u(c_i, c_j)$ mínima, fins arribar a un únic clúster Ω. A més, cada cop que s'ajunten dos clústers, es defineix la funció índex com

$$\alpha(c_i \cup c_j) = u(c_i, c_j).$$

D'aquesta manera s'ha pogut, a partir d'un espai ultramètric, definir una jerarquia indexada (C, α). □

Observació 3.13. Aquest teorema mostra el procediment estàndard per construir jerarquies indexades en espais ultramètrics. En la majoria dels casos, i en particular després d'haver realitzat un anàlisi de components principals, es treballa en $(\mathbb{R}^n, d_{Eucl})$, un espai no ultramètric. En efecte, considerant $k_n = (k, ..., k) \in \mathbb{R}^n\ \forall k \in \mathbb{R}$,

$$\sqrt{n} = d_{Eucl}((0_n, 1_n) \not\leq \frac{1}{2}\sqrt{n} = \sup\left\{ d_{Eucl}\left(0_n, \left(\frac{1}{2}\right)_n\right), d_{Eucl}\left(\left(\frac{1}{2}\right)_n, 1_n\right)\right\}.$$

Això justifica la necessitat d'adaptar l'algoritme a una distància no ultramètrica.

3.3 Algoritmes de classificació jeràrquica en espais mètrics

Sigui (Ω, d) un espai mètric. Si d és ultramètrica es pot aplicar l'algoritme vist al Teorema 3.12, però en els casos on d no ho sigui cal modificar-lo. A continuació es mostra aquesta variant:

Algoritme genèric per espais mètrics

1. Sigui la partició $\Omega = c_1 \vee ... \vee c_n$. Es calcula $d(c_i, c_j)\ \forall i, j$.
2. S'ajunten els i, j tals que $d(c_i, c_j)$ sigui mínim i es recalcula la distància d'un individu a aquest clúster segons

$$d'(c_k, (c_i \cup c_j)) = f(d(c_i, c_k), d(c_j, c_k)).$$

   Aquesta funció f determina quin dels mètodes s'usa. L'únic requisit que ha de complir és transformar triangles respecte d en triangles ultramètrics respecte d'.
3. Es considera la nova partició, i es repeteix el pas 2 fins arribar a un sol clúster.

Observació 3.14. La gran diferència entre els dos mètodes es troba al pas 2. Als espais ultramètrics es pot definir una distància ultramètrica sobre la nova partició de Ω, mentre que en aquest cas cal una d' ultramètrica a partir d'una funció de distàncies d, en general no ultramètrica. Aquesta f es pot agafar de moltes maneres diferents, però el treball en mostra només tres: el mètode del mínim o *single linkage*, el mètode del màxim o *complete linkage* i el mètode de Ward.

3.3.1 Fórmula de Lance-Williams

Totes les alternatives per la funció f provenen de la mateixa expressió, la fórmula de Lance-Williams, que permet definir les distàncies d' entre un punt i un clúster a partir de fer variar quatre coeficients. Ve donada per

$$d'(k,\{i,j\}) = \alpha d(k,i) + \beta d(k,j) + \gamma d(i,j) + \delta\,|d(k,i) - d(k,j)|\,, \tag{3.2}$$

on els coeficients poden variar sota les restriccions $\alpha+\beta+\gamma = 1$, $\alpha = \beta$, $\gamma < 1$.
Els següents apartats introdueixen diferents mètodes de calcular la distància entre un element i un clúster, que són tots derivats de la igualtat (3.2).

3.3.2 Mètode del mínim

Aquest mètode defineix

$$d'\left(c_k, c_{i,j}\right) = f\left(d(c_i,c_k), d(c_j,c_k)\right) = \min\{d(c_i,c_k), d(c_j,c_k)\}.$$

El mètode del mínim deriva de la fórmula de Lance-Williams utilitzant $\alpha = \beta = \frac{1}{2}$, $\gamma = 0$, $\delta = -\frac{1}{2}$. En efecte, aplicant aquests coeficients es té

$$d'(k,\{i,j\}) = \frac{1}{2}d(k,i) + \frac{1}{2}d(k,j) - \frac{1}{2}|d(k,i) - d(k,j)|, \tag{3.3}$$

i es poden donar dues situacions diferents:

a. Si $d(k,i) \geq d(k,j)$, llavors

$$d'(k,\{i,j\}) \overset{(3.3)}{=} \frac{1}{2}d(k,i) + \frac{1}{2}d(k,j) - \frac{1}{2}\left(d(k,i) - d(k,j)\right) = d(k,j) = \min\{d(k,i), d(k,j)\}.$$

b. Si $d(k,i) \leq d(k,j)$, llavors

$$d'(k,\{i,j\}) \overset{(3.3)}{=} \frac{1}{2}d(k,i) + \frac{1}{2}d(k,j) - \frac{1}{2}\left(d(k,j) - d(k,i)\right) = d(k,i) = \min\{d(k,i), d(k,j)\}.$$

Com es veu en la següent proposició, el mètode del mínim genera una distància ultramètrica que és la millor aproximació inferior de d d'entre totes les distàncies ultramètriques:

Proposició 3.15 (Distància ultramètrica generada pel mètode del mínim)**.** *Sigui $\underline{U}$ el conjunt de distàncies u tals que són ultramètriques i compleixen que $\forall i,j$, $u(i,j) \leq d(i,j)$. Aleshores, la distància $\underline{u}$ generada pel mètode del mínim és l'element màxim de $\underline{U}$, és a dir,*

$$\underline{u}(i,j) \geq u(i,j) \quad \forall u \in \underline{U} \;\; \forall i,j \in \Omega. \tag{3.4}$$

El mètode del mínim es pot construir d'una manera alternativa. Per a fer-ho cal definir el concepte de cadena:

Definició 3.16 (Cadena)**.** *Sigui (Ω, d) un espai mètric. Es defineix una cadena com*

$$[i,j]_m = \{i = i_1, i_2, ..., j = i_m\},$$

on $i_1, ..., i_m \in \Omega$.

I amb aquesta definició es té el següent resultat:

Teorema 3.17. *Sigui (Ω, d) un espai mètric i $[i,j]_m$ una cadena en Ω. Es defineixen*

$$\sup[i,j]_m := \sup_{1\leq p<m} d(i_p, i_{p+1})$$

i

$$\underline{\tilde{u}}(i,j) := \inf_m \sup[i,j]_m.$$

Aquesta distància és igual a la distància ultramètrica obtinguda pel mètode del mínim.

Demostració. N'hi ha prou amb veure que $\underline{\tilde{u}}$ és distància ultramètrica, compleix que $\underline{\tilde{u}}(i,j) \leq d(i,j)\ \forall i,j \in \Omega$ i si hi ha una altra distància ultramètrica u tal que $u(i,j) \leq d(i,j)\ \forall i,j \in \Omega$, aleshores $\underline{\tilde{u}}(i,j) \geq u(i,j)$. Amb això, per la Proposició 3.15, es té que $\underline{\tilde{u}} = \underline{u}$.

Evidentment, com que tota cadena que uneix i amb j uneix també j amb i, la propietat simètrica de la distància ultramètrica es satisfà. Per altra banda, com que d és distància, $\sup[i,j]_m \geq 0$, i per tant $\underline{\tilde{u}}(i,j) = \inf_m \sup[i,j]_m \geq 0$.
Falta només veure que $\forall i,j,k \in \Omega$, $\underline{\tilde{u}}(i,k) \leq \sup\{\underline{\tilde{u}}(i,j), \underline{\tilde{u}}(j,k)\}$. Es consideren dues cadenes, una que vagi de i a j, $[i,j]_m$, i una altra que vagi de i a j passant per k, $[i,k,j]_n$. Com que el conjunt de cadenes van de i a j passant per k són un subconjunt de les cadenes que van de i a j,

$$\inf_m \sup[i,j]_m \leq \inf_n \sup[i,k,j]_n.$$

Ara, totes les cadenes que van de i a j passant per k es poden crear "enganxant" dues cadenes: una $[i,k]$ i una $[k,j]$, disjuntes excepte per k. Per tant, per qualsevol cadena $[i,k,j]$ es té

$$\sup[i,k,j] = \sup\left\{\sup[i,k], \sup[k,j]\right\},$$

i per tant

$$\underline{\tilde{u}}(i,j) = \inf_m \sup[i,j]_m \leq \inf_n \sup[i,k,j]_n,$$

de forma que $\forall n' \in \mathbb{N}$, $n' \geq 3$,

$$\underline{\tilde{u}}(i,j) \leq \sup[i,k,j]_{n'} = \sup\left\{\sup[i,k]_\alpha, \sup[k,j]_\beta\right\}. \tag{3.5}$$

Ara es poden donar dos casos:

- $\sup\left\{\sup[i,k]_\alpha, \sup[k,j]_\beta\right\} = \sup[i,k]_\alpha$. Aleshores, per (3.5),

  $$\underline{\tilde{u}}(i,j) \leq \sup[i,k]_\alpha,$$

  i prenent ínfims per α a banda i banda

  $$\underline{\tilde{u}}(i,j) = \inf_\alpha \underline{\tilde{u}}(i,j) \leq \inf_\alpha \sup[i,k]_\alpha = \underline{\tilde{u}}(i,k).$$

- $\sup\left\{\sup[i,k]_\alpha, \sup[k,j]_\beta\right\} = \sup[k,j]_\beta$. Per un argument similar, prenent ínfims respecte β en aquest cas, es té

  $$\underline{\tilde{u}}(i,j) = \inf_\beta \underline{\tilde{u}}(i,j) \leq \inf_\beta \sup[k,j]_\beta = \underline{\tilde{u}}(k,j).$$

De manera que

$$\underline{\tilde{u}}(i,j) \leq \sup\left\{\underline{\tilde{u}}(i,k), \underline{\tilde{u}}(k,j)\right\},$$

provant la propietat que faltava per assegurar que $\underline{\tilde{u}}$ és distància ultramètrica.

Per altra banda, sigui una altra distància ultramètrica u tal que $u(i,j) \leq d(i,j)\ \forall i,j \in \Omega$. Per veure que $\underline{\tilde{u}}(i,j) \geq u(i,j)$ cal provar primer la següent desigualtat per inducció sobre m: donat un conjunt d'elements $\{i_1, ..., i_m\}$ i u una distància ultramètrica, es compleix que

$$u(i_1, i_m) \leq \sup_{p\in\{1,\ldots,m-1\}} \left\{u(i_p, i_{p+1})\right\}. \tag{3.6}$$

- $m = 2$:
$$u(i_1, i_2) \leq \sup_{p=1} \{u(i_p, i_{p+1})\} = \sup \{u(i_1, i_2)\} = u(i_1, i_2).$$

- Es suposa cert per m. Ara, per $m+1$ es té:
$$\begin{aligned} u(i_1, i_{m+1}) &\overset{(*)}{\leq} \sup \{u(i_1, i_m), u(i_m, i_{m+1})\} \overset{(HI)}{\leq} \sup \left\{ \sup_{p \in \{1,...,m-1\}} u(i_p, i_{p+1}), u(i_m, i_{m+1}) \right\} \\ &= \sup \{u(i_1, i_2), ..., u(i_{m-1}, i_m), u(i_m, i_{m+1})\} = \sup_{p \in \{1,...,m\}} u(i_p, i_{p+1}), \end{aligned} \tag{3.7}$$

  on en $(*)$ s'aplica que u és distància ultramètrica i en HI la hipòtesi d'inducció.

Sigui ara $\{i = i_1, ..., j = i_m\}$ una cadena que va de i a j. Segons (3.6),

$$u(i, j) = u(i_1, i_m) \leq \sup_{p \in \{1,...,m-1\}} \{u(i_p, i_{p+1})\},$$

i com que aquesta cadena uneix i i j, aleshores

$$u(i, j) \leq \sup_{p \in \{1,...,m-1\}} \{u(i_p, i_{p+1})\} \leq \sup[i, j]_m,$$

donat que $\forall i, j \in \Omega$ es té $u(i, j) \leq d(i, j)$. Per últim, prenent ínfims per m als dos costats,

$$u(i, j) \leq \inf_m \sup[i, j]_m = \underline{\tilde{u}}(i, j).$$

D'aquesta manera, $\underline{\tilde{u}}$ satisfà:

- És una distància ultramètrica.
- $\underline{\tilde{u}}(i, j) \leq d(i, j)$. No s'ha provat fins ara, però considerant la cadena $\{i, j\}$ és evident que
$$\underline{\tilde{u}}(i, j) = \inf_m \sup[i, j]_m \leq \sup[i, j]_2 = d(i, j).$$
- Qualsevol altra distància ultramètrica u tal que $u(i, j) \leq d(i, j)\ \forall i, j \in \Omega$ compleix $u(i, j) \leq \underline{\tilde{u}}(i, j)$.

Per tant, com $\underline{u}$ és ultramètrica i compleix $\underline{u}(i, j) \leq d(i, j)\ \forall i, j \in \Omega$, per la Proposició 3.15 $\forall i, j \in \Omega$,

$$\underline{u}(i, j) \leq \underline{\tilde{u}}(i, j),$$

però per altra banda, $\underline{\tilde{u}}$ també és ultramètrica i compleix que $\underline{\tilde{u}}(i, j) \leq d(i, j)$. Per tant, per la mateixa proposició, $\forall i, j \in \Omega$,

$$\underline{\tilde{u}}(i, j) \leq \underline{u}(i, j).$$

I amb això queda provat que les dues distàncies ultramètriques són equivalents. □

3.3.3 Mètode del màxim

En aquest cas, la distància entre un punt i un conjunt de dos elements es defineix com

$$d'(k, \{i, j\}) = \max \{d(i, k), d(j, k)\}.$$

De manera anàloga a la Proposició 3.15, aquest mètode genera una distància ultramètrica $\overline{u}$ que compleix la següent propietat:

Proposició 3.18 (Distància ultramètrica generada pel mètode del màxim)**.** *Sigui (Ω, d) un espai mètric i $\overline{U}$ el conjunt de distàncies ultramètriques u tals que $u(i, j) \geq d(i, j)\ \forall i, j \in \Omega$. Aleshores, la distància ultramètrica definida pel mètode del màxim és un element minimal de $\overline{U}$.*

Observació 3.19. Així com la distància generada pel mètode del mínim sempre és única, la generada pel màxim no ho és. Només quan les entrades de la matriu de distàncies entre els elements (la matriu que a la posició i, j té el valor $d(i, j)$) que es troben fora de la diagonal són totes diferents, la distància generada és única i és l'element mínim de $\overline{U}$.

Aquesta distància d' també es pot obtenir a partir de (3.2), utilitzant $\alpha = \beta = \delta = \frac{1}{2}$, $\gamma = 0$. En aquest cas,

$$d'(k, \{i, j\}) = \frac{1}{2}\left(d(k, i) + d(k, j) + |d(k, i) - d(k, j)|\right), \tag{3.8}$$

i hi ha la mateixa situació que al mètode del mínim:

a. Si $d(k, i) \geq d(k, j)$, llavors

$$d'(k, \{i, j\}) \overset{(3.8)}{=} \frac{1}{2}\left(d(k, i) + d(k, j) + (d(k, i) - d(k, j))\right) = d(k, i) = \max\{d(k, i), d(k, j)\}$$

b. Si $d(k, i) \leq d(k, j)$, llavors

$$d'(k, \{i, j\}) \overset{(3.8)}{=} \frac{1}{2}\left(d(k, i) + d(k, j) + (d(k, j) - d(k, i))\right) = d(k, j) = \max\{d(k, i), d(k, j)\}$$

Per acabar, existeix una propietat que relaciona la distància definida en l'espai mètric i les ultramètriques $\underline{u}$ i $\overline{u}$:

Proposició 3.20. *Sigui (Ω, d) un espai mètric, i $\underline{u}$ i $\overline{u}$ les distàncies ultramètriques definides pel mètode del mínim i del màxim respectivament. Aleshores es compleix que $\forall i, j \in \Omega$,*

$$\underline{u}(i, j) \leq d(i, j) \leq \overline{u}(i, j),$$

i es satisfà la igualtat si, i només si, d és ultramètrica.

Demostració. Pel Teorema 3.12 i la Proposició 3.15, les desigualtats es compleixen. Només cal comprovar la doble implicació:

$\Rightarrow$) Si $\underline{u}(i, j) = d(i, j) = \overline{u}(i, j)$ $\forall i, j \in \Omega$, com que tant $\underline{u}$ com $\overline{u}$ són ultramètriques, d també ho és.

$\Leftarrow$) Si d és ultramètrica, com que $d(i, j) \leq d(i, j)$, es té que $d \in \underline{U} \cap \overline{U}$. Amb això,

- $d \in \underline{U} \Rightarrow \forall i, j \in \Omega$, $\underline{u}(i, j) \geq d(i, j)$. Però per definició de $\underline{U}$ i per construcció de la ultramètrica $\underline{u}$, $\forall i, j \in \Omega$, $\underline{u}(i, j) \leq d(i, j)$. Per tant, $\underline{u}(i, j) = d(i, j)$.
- $d \in \overline{U} \Rightarrow \forall i, j \in \Omega$, $\overline{u}(i, j) \leq d(i, j)$. Però per definició de $\overline{U}$ i per construcció de la ultramètrica $\overline{u}$, $\forall i, j \in \Omega$, $\overline{u}(i, j) \geq d(i, j)$. Per tant, $\overline{u}(i, j) = d(i, j)$.

I amb això queda demostrada la doble implicació. □

3.3.4 Mètode de Ward

L'últim mètode que recull el treball és el mètode de Ward. Introduït per J. H. Ward Jr. l'any 1963, busca en cada pas ajuntar els dos clústers que minimitzin l'increment de la suma per tot individu de la distància entre ell i el centre del grup al qual pertany.
La funció que es vol optimitzar està relacionada amb

$$E = \sum_{h=1}^{H} E_h, \tag{3.9}$$

on E_h és la suma de les distàncies euclidianes al quadrat entre cada individu i el centre del clúster h. Posant x_{ij}^h el valor de la variable j en l'individu i del clúster h i $m^h = (m_1^h, ..., m_n^h)$ el centre del clúster, es té que

$$E_h = \sum_{i=1}^{n_h} d_{Eucl}^2(x_i^h - m^h) = \sum_{i=1}^{n_h}\sum_{j=1}^{n} \left(x_{ij}^h - m_j^h\right)^2 = \sum_{i=1}^{n_h}\sum_{j=1}^{n} \left(x_{ij}^h\right)^2 - n_h \sum_{j=1}^{n} \left(m_j^h\right)^2.$$

Si en un pas de l'algoritme s'ajunten dos clústers C_k i C_l, es genera un nou grup $C_p := C_k \cup C_l$ que provoca un canvi en E. El mètode de Ward busca minimitzar ΔE, aquest increment, que com que només canvien els clústers k, l i p es calcula segons

$$\Delta E = E_p - (E_k + E_l) = E_p - E_k - E_l. \tag{3.10}$$

Aquesta expressió es pot escriure equivalentment de manera molt més senzilla, involucrant només el nombre d'individus dels grups que s'ajunten i els seus centres. La demostració d'aquest resultat es troba a l'annex A.

Proposició 3.21. *El mètode de Ward ajunta a cada pas els clústers C_k i C_l tals que*

$$\frac{n_k n_l}{n_p} \sum_{j=1}^{n} (m_j^k - m_j^l)^2$$

és mínim.

Observació 3.22. La distància generada per aquest mètode també es pot obtenir amb la fórmula de Lance-Williams fent servir $\alpha = \beta = \frac{n_k+n_i}{n_i+n_j+n_k}$, $\gamma = -\frac{n_k}{n_i+n_j+n_k}$ i $\delta = 0$.

3.4 Clústering no jeràrquic: l'algoritme *k-means*

Un cop acabat l'estudi dels mètodes de clústering jeràrquic, en aquesta secció es presenten els algoritmes no jeràrquics. A diferència dels anteriores, aquests generen una única partició, de manera que convergeixen més ràpidament, però necessiten com a paràmetre inicial el nombre de grups que hi haurà al clústering.
Els algoritmes no jeràrquics parteixen d'una mesura de qualitat associada amb cada partició (veure l'apartat 3.4.3 per un exemple d'aquests criteris), i cal trobar la partició dels elements de Ω que fa millor aquesta mesura. Teòricament això és molt senzill, només es necessita generar totes les particions possibles dels n individus en k clústers, on k és el nombre de grups fixat inicialment i trobar quina és la que millora la mesura de qualitat d'entre totes elles. Aquest enfoc teòric és evidentment correcte, però a la pràctica no és possible. Per Liu (1968) se sap que el nombre de particions de n elements en k grups és

$$N(n,k) := \frac{1}{k!} \sum_{i=1}^{k} (-1)^{k-i} \binom{k}{i} i^n,$$

i per tenir una idea del valor que pot prendre aquesta expressió, només havent de classificar 50 individus en 4 grups diferents ja es té

$$N(50,4) \simeq 5.3 \times 10^{28},$$

de manera que cal buscar alguna opció alternativa per reduir el nombre de càlculs. Un d'aquests mètodes és l'algoritme *k-means*, que s'estudia més endavant en el capítol. Per ara, el següent apartat analitza els possibles criteris que es poden fer servir en dades multivariants contínues com les que es fan servir en aquest projecte.

3.4.1 Criteris d'optimització en dades multivariants contínues

Sigui $X \in \mathcal{M}_{n\times p}$ la matriu de dades. En la Proposició A.8 es prova que la variància d'una variable es pot trencar en la variància dins dels grups i entre els grups. Aquesta idea es pot replicar en $\mathbb{R}^p$ com es veu a continuació.
Es denota per $x_i^m = (x_{i1}^m, ..., x_{ip}^m)$ l'individu i del clúster m, i $\overline{x}$ el vector de mitjanes globals, és a dir, $\overline{x} = (\overline{x}_1, ..., \overline{x}_p)$, i aleshores es posa la matriu de dispersió (juga el paper de la variància en dimensió u) com

$$T = \sum_{j=1}^{k}\sum_{l=1}^{n_j}(x_l^j - \overline{x})(x_l^j - \overline{x})^T, \tag{3.11}$$

on n_j és el nombre d'elements que hi ha al grup j. Aquesta matriu es pot descomposar en

$$T = B + W, \tag{3.12}$$

on, posant $\overline{x}^m = (\overline{x}_1^m, ..., \overline{x}_p^m)$ com el vector de mitjanes al clúster m,

$$B := \sum_{j=1}^{k} n_j(\overline{x}^j - \overline{x})(\overline{x}^j - \overline{x})^T \tag{3.13}$$

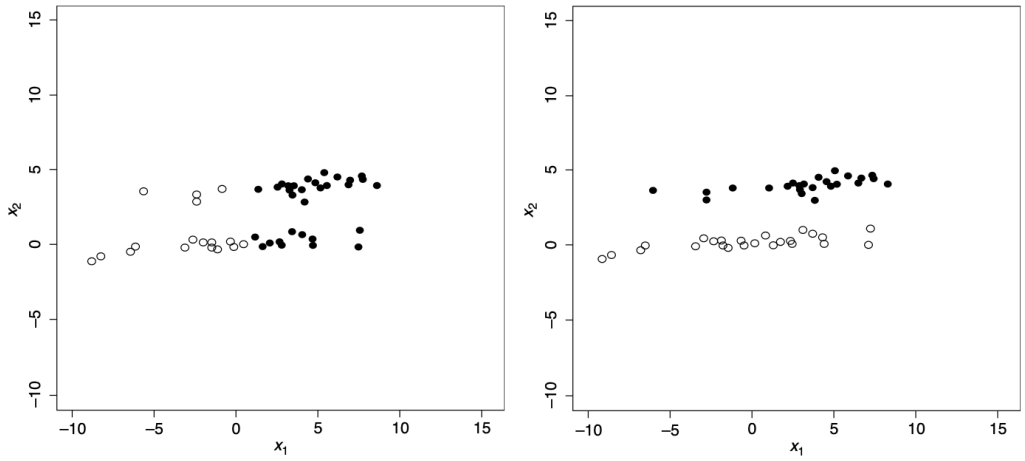
i

$$W := \sum_{j=1}^{k}\sum_{l=1}^{n_j}(x_l^j - \overline{x}^j)(x_l^j - \overline{x}^j)^T. \tag{3.14}$$

Com ja s'ha comentat, l'objectiu dels mètodes de clústering és minimitzar V_{intra} per tal de construir grups el més homogenis possible. Però extrapolar aquesta idea a casos en més dimensions no és tan directe com pot semblar, ja que la relació d'ordre que es té en $\mathbb{R}$ no s'esté a l'espai de matrius. És per això que apareixen diferents criteris que poden servir per traslladar la idea de minimitzar la variància dins dels grups al context de més variables, cadascun amb les seves propietats. A continuació es llisten els més usats[7]:

- Minimització de $tr(W)$. És la traducció més directa de la minimització de V_{intra}, donat que la traça de W és igual a la suma de V_{intra} per cadascuna de les p variables de la matriu. Aquest criteri provoca una reducció de la suma de les distàncies entre cada individu i el centre del clúster on es troba.
  És el criteri més usat per dur a terme algoritmes de clústering, però té una sèrie d'inconvenients: en primer lloc, la traça de W és dependent de la mida de la matriu, i per altra banda el mètode imposa una forma esfèrica als clústers que pot introduir un biaix en els resultats, fent que els grups siguin més artificials degut a aquesta tendència. Un molt bon exemple d'això ve donat per [14], pàg. 117, reproduït en les imatges 1a i 1b, on es veu clarament que els dos clústers tenen més sentit en la segona imatge, però que el biaix que presenta la minimització de $tr(W)$ provoca que els grups que apareguin siguin els de la primera imatge, amb forma més similar a un cercle.
  Una solució per evitar aquest problema i poder seguir fent ús d'aquest criteri és el que es du a terme a la part pràctica del treball: la normalització. El fet de normalitzar les dades (centrar i dividir per la desviació típica) porta a clústers amb formes més similars a una n-esfera, que és el que es necessita per poder fer servir aquest criteri.
  Per últim, es pot provar que la minimització de la traça de W és equivalent a minimitzar la suma de distàncies al quadrat de cada punt amb el centre del seu clúster. És per això que, per exemple, l'algoritme *k-means* (veure capítol 3.4.2) fa ús indirecte d'aquest criteri.

[7] S'han plantejat altres criteris al llarg dels anys, intentant solucionar els problemes existents en algoritmes anteriors, però no es comentaran en aquest treball per dos motius: en primer lloc caldria introduir notacions i resultats que provocarien que l'apartat s'estengués innecessàriament, i per altra banda s'ha considerat no tractar-los perquè no es fan servir a la pràctica.



(a) Minimització $tr(W)$ (b) Minimització $\det(W)$

Figura 1: Clústering segons dos criteris.

- Minimització de $det(W)$. En l'anàlisi multivariant hi ha un test que permet comprovar, donat un conjunt d'individus repartits en diferents grups, si les seves mitjanes són significativament diferents. A partir de la descomposició vista en (3.12), el test postula que valors grans de $\Delta := \frac{det(T)}{det(W)}$ determinen que les mitjanes són prou diferenciades. El segon criteri per determinar la millor partició fa ús d'aquest test i busca fer el mateix, augmentar el valor d'aquest quocient. Per a fer-ho, però, cal notar que T es manté constant sigui quina sigui la partició, i per tant també ho és $det(T)$, així que trobar una partició que faci màxim el valor de Δ equival a trobar la classificació que minimitzi $det(W)$.
  Aquest criteri, tot i que més complicat d'aplicar a la pràctica amb R, soluciona el problema dels clústers esfèrics. De fet, en la imatge 1b, el criteri usat ha estat aquesta minimització de $det(W)$.

- Maximització de $tr(BW^{-1})$. Aquest criteri té el seu origen de nou en l'anàlisi de la variància en més dimensions: en concret prové d'un altre test per a comprovar si els vectors de mitjanes dels grups són diferents entre elles, fet que passa quan el valor $tr(BW^{-1})$ és prou elevat. És per això que buscant la partició que fa màxima aquesta traça n'hi ha prou per construir el clústering.

3.4.2 Algoritme *k-means*

En aquest apartat es tracta el segon dels algoritmes que apareixen al codi de la part pràctica del treball, *k-means*. És un dels algoritmes més coneguts en l'àmbit de la classificació no jeràrquica, i es pot fer servir en un clústering seqüencial per a consolidar la partició que els algoritmes jeràrquics generen. L'algoritme és realment senzill, i es pot explicar en aquests quatre passos:

1. Es fixa un valor $k \in \mathbb{N}$ a priori, el nombre de clústers en els quals es vol dividir els individus.

2. S'escullen k punts per inicialitzar l'algoritme, a l'atzar o a partir del que retorna un clústering jeràrquic previ, calculant el centre de cada grup.

3. S'assigna cada punt del núvol de punts al grup que tingui el centre més a prop d'ell. Un cop s'ha fet això amb tots els punts, es redefineixen els grups i es calculen els centres de cadascun d'ells.

4. Es denota C_j^l el centre del clúster j al pas l. Es fixa també $\epsilon > 0$. Aleshores, es tenen dues opcions per aquest pas de l'algoritme:

a. Si $\exists j_0$ tal que $\left|C_{j_0}^{l} - C_{j_0}^{l-1}\right| \geq \epsilon$, es torna al pas 3.

b. Altrament, l'algoritme acaba i es té la partició definitiva.

Es pot provar que la convergència d'aquest algoritme és sempre a un òptim local, però no es garanteix la convergència a òptims globals. Per a millorar la qualitat de les particions, diversos autors han proposat algunes solucions, involucrant també algoritmes jeràrquics. Les dues més comunament usades són les següents:

- Es repeteix l'algoritme un nombre molt gran de vegades, com a mínim 5000, i s'escull la que millor valor assoleix en la forma d'avaluar la classificació. Això no garanteix assolir l'òptim global, però provoca que la partició que s'esculli sigui millor que la que s'obtindria executant només una iteració.
- Sobretot per bases de dades grans, es pot seguir un procediment que combina els dos tipus d'algoritmes que s'han tractat durant el treball. Es comença amb dues repeticions d'un algoritme no jeràrquic amb 10 clústers, classificant els individus en una taula creuada en funció del grup on han estat assignats en cadascuna de les repeticions. A partir d'aquí es calculen els centres dels grups formats pels individus que han estat classificats en cadascuna de les cel·les de la taula i s'executa una classificació jeràrquica amb aquests punts, que ofereix el nombre k que es requereix per inicialitzar l'algoritme no jeràrquic. Per acabar, es consolida la partició amb una altra repetició de *k-means* o algun procediment similar, per obtenir la classificació definitiva.

A partir d'aquests algoritmes s'obté una classificació dels individus en k grups diferents, que es pot avaluar (això és, comprovar que la classificació és prou bona) a partir del procediment que es tracta a la secció següent. Per ara, però, les següents línies entren en dues de les modificacions del mètode, per veure la seva evolució. Cal tenir en compte que tot i ser introduït fa més de 40 anys, segueix sent molt estudiat i utilitzat en disciplines molt diverses.

Fast k-means Aquesta primera variant de l'algoritme assegura convergència en com a màxim n iteracions. L'inici del procés és exactament el mateix: cal escollir $k \in \mathbb{N}$ com a nombre de clústers i inicialitzar l'algoritme amb k centres. En aquest cas, però, només es treballa amb un individu en cada pas, i es repeteixen aquests dos passos amb cadascun d'ells:

1. S'assigna aquest individu al centre que tingui més a prop, i no es treballa amb la resta dels individus.
2. Es redefineixen els grups i es recalculen els centres. Evidentment, l'únic centre que canviarà de posició és el del grup on s'ha afegit l'individu.

Com s'ha comentat a l'inici, amb n passes s'acaba aquest algoritme. La convergència està assegurada a òptims locals, igual que convergeix el *k-means* original, però és possible que la qualitat de la classificació sigui lleugerament pitjor.

k-medioids Una altra variant és l'algoritme *k-medioids*, on en lloc de calcular o escollir els centres com a punts de l'espai $\mathbb{R}^p$, han de ser necessàriament punts corresponents a algun dels individus. Es poden traduir els quatre passos de l'algoritme original de la següent manera:

1. Es fixa un valor $k \in \mathbb{N}$ a priori. Aquest és el nombre de clústers en els quals es vol separar els individus.
2. S'escullen k elements de la base de dades per inicialitzar l'algoritme.
3. S'assigna cada punt del núvol de punts al grup que tingui el centre més a prop d'ell. Un cop s'hagi fet això amb tots els punts, es redefineixen els grups i es calculen els centres de

cadascun d'ells. Aquest càlcul es fa de la següent manera: suposant que s'està treballant sobre el clúster C_j, es busca el punt $p \in C_j$ tal que es faci mínim

$$\sum_{p' \in C_j} d(p, p')$$

on d és la distància sobre la qual s'estigui treballant (molt sovint és la distància euclídia).

4. Es denota C_j^l el centre del clúster j al pas l. Es fixa també $\epsilon > 0$. Aleshores, es tenen dues opcions per aquest pas de l'algoritme:
   a. Si $\exists j_0$ tal que $\left|C_{j_0}^l - C_{j_0}^{l-1}\right| \geq \epsilon$, es torna al pas 3.
   b. Altrament, l'algoritme acaba i es té la partició definitiva.

3.4.3 El *silhouette plot*: una manera d'avaluar la classificació

Un cop realitzada la classificació no jeràrquica cal buscar alguna forma de veure si el clústering és suficientment bo. S'han plantejat un gran nombre de maneres de fer-ho, moltes d'elles basades en tests multivariants igual que passava en les funcions objectiu pels mètodes de clústering no jeràrquic (veure pp. 23-24), però hi ha una manera més senzilla i fàcil d'executar en R que és l'anomenat *silhouette plot*, una representació gràfica de l'anomenat coeficient de silhouette. Aquesta noció, introduïda per Kaufman i Rousseew, determina com de ben classificat es troba un individu en un determinat clústering.
Per tal de calcular-lo, es suposen els n individus repartits en k clústers, $\Omega = c_1 \vee ... \vee c_k$, i es considera $i \in c_j$ un individu dins d'un clúster determinat. Siguin aleshores

$$a(i) = \frac{1}{|c_j| - 1} \sum_{l \in c_j - \{i\}} d(i, l)$$

la mitjana de totes les distàncies de i a la resta d'elements del clúster on pertany i

$$b(i) = \min_{m \neq j} \frac{1}{|c_m|} \sum_{l \in c_m} d(i, l)$$

la menor de les mitjanes de les distàncies de i a tots els punts d'un clúster diferent al que pertany[8].
A partir d'aquests dos valors es defineix

$$s(i) = \begin{cases} \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} & \text{si } |c_j| > 1, \\ 0 & \text{altrament.} \end{cases} \tag{3.15}$$

i es satisfà $-1 \leq s(i) \leq 1$, com es prova en l'annex A.
Amb això ja es pot definir el coeficient de silhouette del clústering. Es fa de la següent manera:

$$CS = \max_k \overline{s}(k), \tag{3.16}$$

amb

$$\overline{s}(k) = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} s(i), \tag{3.17}$$

on k és el nombre de clústers i n_j el nombre d'elements del clúster c_j. A la pràctica, doncs, l'objectiu és obtenir $\arg\max_k \overline{s}(k)$ per construir un clústering amb aquest nombre de grups.

[8]En aquests casos es sol dir que el clúster que satisfà aquest mínim és el clúster veí de i, perquè és el segon millor clúster per aquest individu.

En la Figura 2 es veu un exemple del gràfic de silhouette extret de l'estudi de les dades de la part pràctica del treball. Aquest consisteix en una línia per cadascun dels individus representant el valor de $s(i)$ de cadascun d'ells, així com sovint una línia vertical indicant el valor $\overline{s}(k)$:

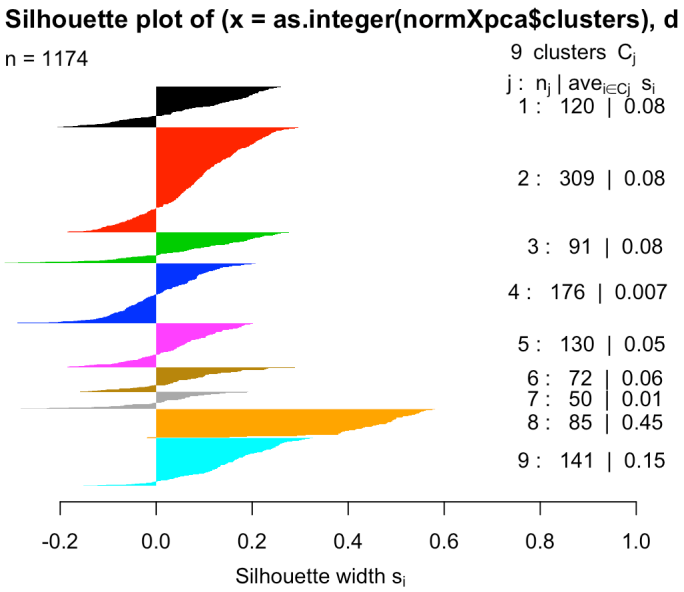


Figura 2: Exemple de *silhouette plot*.

Amb aquesta imatge del *silhouette plot* es clou aquesta secció dedicada als algoritmes de clústering. Al quart capítol del treball es realitza un estudi sobre els mètodes de classificació, K-Nearest Neighbors i l'anàlisi discriminant, una manera diferent de classificar els individus.

4 Classificació

Després de parlar durant el capítol anterior del clústering d'individus, cercant unes categories en les quals poden ser assignats, l'objectiu d'aquest capítol és classificar de manera automàtica en funció d'una variable resposta Y categòrica una nova observació de les variables que construeixen la matriu de dades. A tal efecte s'estudien conceptes com el classificador de Bayes o l'anàlisi discriminant per arribar a construir l'anàlisi discriminant quadràtica.
Les referències principals pel capítol són [4], [7] i [10].

Dos conceptes importants que s'introdueixen en aquest tema són la base de dades *training* i la base de dades *testing*. Donada la matriu de dades amb els seus n individus, es divideixen en dues matrius més petites de mida n_{tr} i n_{te}[9], cadascuna amb un objectiu diferent: *training* es fa servir per desenvolupar el model, i *testing* per comprovar la qualitat del mateix. A més, durant tot el capítol, es suposa que es vol estimar una funció de densitat f a partir d'una base de dades *training* formada per $\{(x_1, y_1), ..., (x_n, y_n)\}$ on y_i són els valors que pren la variable resposta i x_i els valors de les variables presents a la matriu de dades, que en algunes situacions són observacions univariants i en d'altres multivariants. Es denota també l'estimació d'aquesta densitat com $\hat{f}$, posant $\hat{y}_i = \hat{f}(x_i)$.

Abans de començar a analitzar la teoria subjacent als algoritmes de classificació hi ha un aspecte important a destacar. Al llarg del capítol anterior s'ha vist que els mètodes de clústering busquen assignar cada individu a un grup, i com el seu propi nom indica els algoritmes de classificació persegueixen el mateix objectiu. Hi ha, però, una diferència clau que és el motiu pel qual interessa treballar amb els dos tipus d'algoritmes. Els mètodes de clústering classifiquen els individus d'una matriu de dades en funció de les seves variables en un nombre determinat de grups que no existeixen abans d'aquesta classificació, és a dir, que el propi mètode crea aquests clústers, mentre que els algoritmes de classificació reparteixen els individus en grups ja existents, representats per les modalitats de la variable resposta. A més, els mètodes de classificació permeten fer prediccions amb individus nous, mentre que si es vol classificar una nova observació de les variables a partir d'un clústering cal afegir les observacions a la matriu de dades i tornar a repetir l'algoritme.

Fet l'apunt, per tal d'iniciar l'estudi d'aquests algoritmes de classificació és important tenir clara la següent definició:

Definició 4.1 (Ràtio d'error *training*)**.** *Sigui $\hat{f}$ una densitat estimada i $\{(x_1, y_1), ..., (x_{n_{tr}}, y_{n_{tr}})\}$ una base de dades training. Sigui també $\hat{y}_i = \hat{f}(x_i)$ la classe predita per cada individu $i \in \{1, ..., n_{tr}\}$. Es defineix la ràtio d'error d'aquesta base de dades com*

$$ER_{tr} = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \mathbb{1}_{y_i \neq \hat{y}_i}. \tag{4.1}$$

Observació 4.2. Aquesta mateixa definició es pot replicar per la base de dades *testing*, obtenint una ràtio per aquesta base de dades. Es considera que un classificador és bo si és capaç de minimitzar la ràtio d'error ER_{te}, i es pot demostrar que el classificador que compleix això és el classificador de Bayes. En l'apartat següent se n'estudia la teoria.

Una bona pregunta és per què en una situació com la del treball, on es volen classificar individus en funció d'una variable resposta categòrica, no funcionaria una regressió lineal. Es respondrà amb un exemple: es considera un estudi on es volen classificar jugadors de futbol segons si són davanters, defensors o porters. Es podria codificar la variable resposta Y, la posició del jugador, segons

$$Y = \begin{cases} 1 & \text{si és porter,} \\ 2 & \text{si és defensor,} \\ 3 & \text{si és davanter.} \end{cases}$$

[9] És una pràctica general fer servir $n_{tr} = \frac{2}{3}n$ i $n_{te} = \frac{1}{3}n$

Això permetria fer ús d'algoritmes de regressió lineal per tal de predir Y, però aquest enfoc provocaria fer dues suposicions que desvirtuarien els resultats: en primer lloc suposaria un ordre, és a dir, que el fet de ser defensor cau entre ser porter i ser davanter, que és una afirmació molt artificial. Per altra banda, i més important, aquesta codificació suposa que la diferència entre ser porter i defensor és de la mateixa magnitud que la diferència entre ser defensor i davanter, que de nou no té perquè ser cert a més, de nou, de no tenir massa sentit.
Aquesta idea només es podria fer servir en situacions on Y fos binària, codificant

$$Y = \begin{cases} 1 & \text{si } Y = y_0, \\ 0 & \text{altrament,} \end{cases}$$

i considerant que un individu pertany a la classe y_0 si $\mathbb{P}(Y = y_0|X = x_0) > 0.5$ i a y_1 altrament[10], però quan Y tingui més de dues opcions, com és el cas de la classificació dels jugadors que es du a terme en aquest treball, no es pot estendre. Per aquest motiu és necessari construir altres mètodes que permetin fer classificacions en funció d'una variable categòrica, i aquí rau la importància dels algoritmes de classificació. En aquest treball se'n tractaran dos: K-Nearest Neighbors a l'apartat 4.1, basat en el classificador de Bayes que també es tracta, i l'anàlisi discriminant, tant lineal com quadràtic.

4.1 L'algoritme K-Nearest Neighbors

4.1.1 Classificador de Bayes

El classificador de Bayes, basat en la regla del mateix nom relacionada amb la probabilitat condicionada, és un classificador molt simple que permet reduir la ràtio ER_{te}. La idea que hi ha darrere d'aquest mètode és que el classificador de Bayes assigna cada individu a la classe on té més probabilitat d'estar a partir dels valors dels predictors (les variables de la matriu de dades), això és, assignar l'individu a la classe

$$\arg\max_j \mathbb{P}(Y = j|X = x_0),$$

on $x_0 \in \mathbb{R}^p$ és el vector fila de la matriu de dades corresponent a aquell individu.
Com s'ha comentat, es busca maximitzar

$$\mathbb{P}(Y = j|X = x_0), \tag{4.2}$$

i per tant l'error que es dona en la classificació d'un determinat individu és

$$ER|_{X=x_0} = 1 - \max_j \mathbb{P}(Y = j|X = x_0).$$

Així, l'error global del classificador de Bayes és

$$ER = 1 - \mathbb{E}_X \left(\max_j \mathbb{P}(Y = j|X) \right). \tag{4.3}$$

Aquest és l'error mínim que es pot obtenir, i s'anomena ràtio d'error de Bayes.
A nivell teòric aquest classificador és molt interessant per la seva simplicitat i la qualitat dels resultats que s'obtenen, però a la pràctica és impossible d'aplicar. Això és perquè la distribució (4.2) no és coneguda en general, i per aquesta raó cal construir mètodes que generin estimacions de (4.2). El més senzill és K-Nearest Neighbors, que s'estudia a continuació.

4.1.2 K-Nearest Neighbors

Per tal d'aplicar aquest mètode cal definir a priori dos elements: per una banda la base de dades *training*, on cadascun dels individus ja està assignat a alguna de les categories de la variable resposta Y, i un valor $K \in \mathbb{N}$ que s'usa per determinar el nombre de punts que es fan servir durant l'algoritme, que funciona com segueix:

[10] En funció de la situació, aquest 0.5 pot variar. Com que durant el treball no es fa més ús d'aquest mètode no s'hi aprofundirà molt, però en [10], p. 131, se'n troba un exemple

Algoritme K-Nearest Neighbors

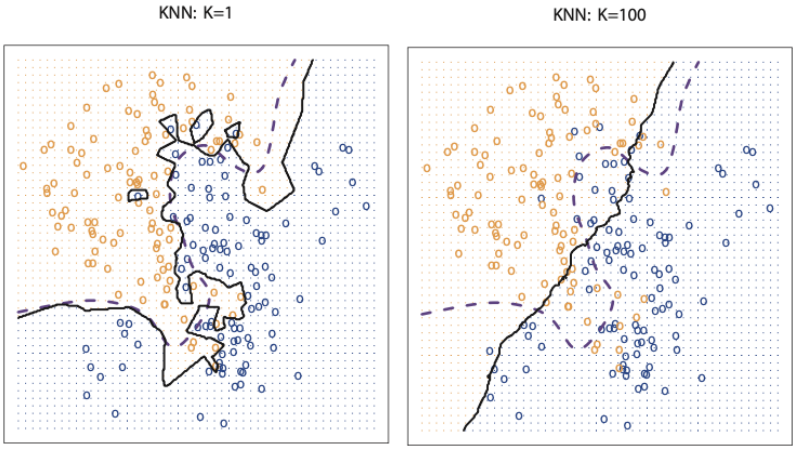
1. Donada una observació x_0 de la base de dades *testing*, es troben els K punts de la base *training* més propers a x_0. Es denota aquest conjunt per K_0.

2. Es calcula per cada categoria j el valor

$$\mathbb{P}(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in K_0} \mathbb{1}_{y_i = j}. \tag{4.4}$$

3. S'assigna cada individu a una classe seguint la regla de Bayes: x_0 s'assigna a la categoria j que maximitza (4.2).

Cal tenir en compte que l'elecció de $K \in \mathbb{N}$ és crítica pel bon funcionament del mètode. Com es mostra en les imatges 3a i 3b, extretes de [10], p. 41, valors massa elevats de K provoquen una frontera de decisió -això és, donades dues classes $Y = y_1$ i $Y = y_2$, els punts tals que $\mathbb{P}(Y = y_1|X = x_0) = \mathbb{P}(Y = y_2|X = x_0)$- excessivament lineal, mentre que valors massa petits generen fronteres massa flexibles:



(a) K massa petita (b) K massa gran

Figura 3: KNN amb diferents eleccions per K.

Amb una bona elecció de K, però, el mètode de KNN pot produir classificacions molt properes a les òptimes donades pel classificador de Bayes.

K-Nearest Neighbors és el mètode més senzill d'aplicació d'una aproximació del classificador de Bayes. En el capítol següent es tracta l'anàlisi discriminant, una tècnica que va un pas més enllà.

4.2 Anàlisi discriminant

Un mètode que no s'estudia en aquest treball és la regressió logística, una manera de modelitzar directament $\mathbb{P}(Y = y|X = x_0)$ quan es tenen dues classes resposta. L'anàlisi discriminant és una via alternativa de donar valors a aquestes probabilitats, extensible a casos on la variable resposta Y té més de dues categories.
El teorema de Bayes és bàsic en aquests algoritmes. Cal, doncs, recordar-lo (a l'annex A se'n veu la prova):

Teorema 4.3 (Teorema de Bayes)**.** *Sigui* $(\Omega, \mathcal{F}, \mathbb{P})$ *un espai de probabilitat. Siguin* $A \in \mathcal{F}$ *un esdeveniment amb probabilitat no nul·la i* $\{B_j\}_{j=1}^n \subset \mathcal{F}$ *una partició de* Ω*. Aleshores,*

$$\mathbb{P}\left(B_i|A\right) = \frac{\mathbb{P}(B_i)\mathbb{P}(A|B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B_i)\mathbb{P}(A|B_i)}{\sum_{j=1}^n \mathbb{P}(A|B_j)\mathbb{P}(B_j)}. \tag{4.5}$$

Primer és necessari establir la notació. Sigui Y la variable resposta categòrica segons la qual es volen classificar els individus, amb $n \geq 2$ opcions diferents i no ordenades. Per cada $y \in \{1, ..., n\}$ es denota per π_y la probabilitat *a priori* de la classe y de la variable resposta, que es pot obtenir de diverses formes, sempre partint de la base de dades *training* (en aquest treball s'obté a partir del clústering que es du a terme prèviament). Es defineix també la funció de densitat condicionada a la classe $Y = y$ com

$$f_y(x) := \mathbb{P}(X = x|Y = y). \tag{4.6}$$

Aquesta pren valors alts si és molt probable que X sigui proper a x sabent que està a la classe y i baixos si passa el contrari. A partir d'això, recuperant el Teorema 4.3 es té que

$$p_y(x) := \mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(Y = y)\mathbb{P}(X = x|Y = y)}{\sum_{i=1}^{n} \mathbb{P}(X = x|Y = i)\mathbb{P}(Y = i)} = \frac{\pi_y f_y(x)}{\sum_{i=1}^{n} \pi_i f_i(x)}, \tag{4.7}$$

de manera que trobant π_y i $f_y(x)$ $\forall y \in \{1, ..., n\}$ ja es poden calcular les probabilitats buscades, i assignar llavors cada individu a la classe que faci més gran (4.7). Estimar π_y és relativament senzill a partir de la base *training*, però $f_y(x)$ és més complicada d'aproximar. No obstant això, si s'aconsegueix, es pot trobar un classificador que faci tan petit com es pugui (4.3), aproximant així el classificador òptim de Bayes. El que es persegueix al llarg de les següents seccions és arribar a un mètode que no imposi massa restriccions sobre les dades com és l'anàlisi discriminant quadràtica, però per tal d'arribar-hi fa falta passar primer per l'anàlisi discriminant lineal, a l'inici amb una sola variable predictiva per comprendre el procés i després estenent-ho a més predictors. A això es dediquen els següents apartats, on cal tenir sempre en compte que es suposa normalitat a totes les variables predictives[11].

4.2.1 Anàlisi discriminant lineal amb una variable predictiva

En aquesta secció es tracta una versió simplificada de l'anàlisi discriminant lineal, o LDA per les seves inicials en anglès. S'assumeix una única variable predictiva X i una variable resposta Y amb n modalitats satisfent que $X|Y = y \sim N(\mu_y, \sigma^2)$.
Partint de (4.7) és clar que l'únic que cal trobar és una estimació de $f_y(x)$ per cada $y \in \{1, ..., n\}$, però amb aquestes dues assumpcions ja es té que $X|Y = y \sim N(\mu_y, \sigma^2)$, és a dir,

$$f_y(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_y)^2}{\sigma^2}}. \tag{4.8}$$

Substituïnt (4.8) en (4.7) s'obté el valor de $p_y(x)$ per cada classe y, i per tant s'assigna l'individu a la classe y que maximitza aquest valor:

$$p_y(x) = \frac{\pi_y \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_y)^2}{\sigma^2}}}{\sum_{i=1}^{n} \pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_i)^2}{\sigma^2}}} \tag{4.9}$$

o, equivalentment[12],

$$\delta_y(x) := \ln(\pi_y) - \frac{x^2}{2\sigma^2} + \frac{x\mu_y}{\sigma^2}. \tag{4.10}$$

Ara bé, amb això no n'hi ha prou per poder construir l'estimador del classificador de Bayes a la pràctica, ja que amb dades reals no es tenen els paràmetres $\pi_1, ..., \pi_n$, $\mu_1, ..., \mu_n$ ni σ, de manera que cal estimar-los segons

$$\hat{\mu}_y = \frac{1}{m_y} \sum_{i:y_i=y} x_i, \tag{4.11}$$

$$\hat{\sigma}^2 = \frac{1}{m-n} \sum_{y=1}^{n} \sum_{i:y_i=y} (x_i - \hat{\mu}_y)^2, \tag{4.12}$$

[11] Això no afecta els resultats de la pràctica del treball. Es disposa d'una base de dades molt gran i per tant, pel Teorema del Límit Central, es pot assumir aquesta distribució Gaussiana en totes les variables.
[12] L'equivalència entre aquestes dues expressions es troba provada a l'annex A.

i

$$\hat{\pi}_y = \frac{m_i}{m}, \tag{4.13}$$

on m és el nombre total d'individus de la base de dades *training*, n el nombre de classes resposta de Y i m_k el nombre d'individus a cada classe de la variable resposta (així, $m = \sum_{k=1}^{n} m_k$)[13]. Aplicant aquests estimadors a (4.8), a la pràctica es classifica l'individu $X = x_0$ a la classe y tal que

$$\hat{\delta}_y(x) := \ln(\hat{\pi}_y) - \frac{x^2}{2\hat{\sigma}^2} + \frac{x\hat{\mu}_y}{\hat{\sigma}^2} \tag{4.14}$$

és màxim.

4.2.2 Anàlisi discriminant lineal amb més d'una variable predictiva

En aquest apartat s'estén el que s'ha vist a la secció anterior a una situació on es vol classificar els individus en funció d'un vector p-dimensional de variables, un cas més proper a la realitat. Igual que en l'LDA amb un sol predictor, cal suposar que $X \sim N(\mu, \Sigma)$, on $\mu = (\mu_1, ..., \mu_p)$ és el vector de mitjanes per cada modalitat de Y i $\Sigma \in \mathcal{M}_{p\times p}$ és la matriu de covariàncies de X. En aquest cas, els individus de cada classe $y \in \{1, ..., n\}$ segueixen una distribució $X|Y = y \sim N(\mu^{(y)}, \Sigma)$, on $\mu^{(y)} = (\mu_1^{(y)}, ..., \mu_p^{(y)})$ és el vector de mitjanes de la classe $Y = y$. Així doncs

$$f_y^{(n)}(x) = \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu^{(y)})^T\Sigma^{-1}(x-\mu^{(y)})\right), \tag{4.15}$$

i substituint (4.15) en (4.7) es té que

$$p_y^{(n)}(x) = \frac{\pi_y \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu^{(y)})^T\Sigma^{-1}(x-\mu^{(y)})\right)}{\sum_{i=1}^{n} \pi_i \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu^{(i)})^T\Sigma^{-1}(x-\mu^{(i)})\right)}. \tag{4.16}$$

El mètode busca classificar l'individu x en la categoria $Y = y$ que maximitza (4.16). De nou es pot substituir aquesta expressió per una més senzilla i equivalent:

$$\delta_y^{(n)}(x) = \ln \pi_y + x^T\Sigma^{-1}\mu^{(y)} - \frac{1}{2}(\mu^{(y)})^T\Sigma^{-1}\mu^{(y)}. \tag{4.17}$$

A la pràctica, però, la situació torna a no ser tan senzilla com en la teoria. Els paràmetres π_y i $\mu^{(y)}$ no es tenen en general, i per tant cal fer servir estimadors anàlegs a (4.11), (4.12) i (4.13) però tenint en compte que ara $x_i, \hat{\mu}^{(y)} \in \mathbb{R}^p$:

$$\hat{\mu}^{(y)} = \frac{1}{m_y}\sum_{i:y_i=y} x_i, \quad \hat{\sigma}^2 = \frac{1}{m-n}\sum_{y=1}^{n}\sum_{i:y_i=y}(x_i - \hat{\mu}^{(y)})^2, \quad \hat{\pi}_y = \frac{m_i}{m}.$$

Tant en l'LDA amb un com amb més predictors, hi ha una regió de l'espai on no es pot classificar els individus en una classe en concret, perquè hi ha dues classes y_i i y_j tals que $p_{y_i}(x) = p_{y_j}(x)$. Aquests punts s'anomenen fronteres de decisió de Bayes, i com que hi ha una frontera d'aquest estil per cada parella de classes, per un determinat problema de classificació en n classes es tenen $\binom{n}{2}$ fronteres de decisió. Recuperant (4.17), donades dues classes y_i i y_j, la frontera de decisió entre elles són els punts $x \in \mathbb{R}^p$ tals que

$$\ln \pi_{y_i} + x^T\Sigma^{-1}\mu^{(y_i)} - \frac{1}{2}(\mu^{(y_i)})^T\Sigma^{-1}\mu^{(y_i)} = \ln \pi_{y_j} + x^T\Sigma^{-1}\mu^{(y_j)} - \frac{1}{2}(\mu^{(y_j)})^T\Sigma^{-1}\mu^{(y_j)}.$$

[13] Només cal fer ús de (4.13) en les situacions on no es tingui coneixement d'aquestes probabilitats a priori, ja que si es coneixen aquests valors només cal usar-los directament.

4.2.3 Anàlisi discriminant quadràtic

Es presenta en aquest apartat l'anàlisi discriminant quadràtic o QDA, l'objectiu final d'aquest capítol. El motiu pel qual s'ha estudiat l'LDA fins ara és perquè el QDA beu directament de la mateixa idea, amb una diferència clau: tot i que ambdós suposen que en cada classe $Y = y$ de la variable resposta els individus provenen d'una distribució multinormal, l'LDA necessita que la covariància (o la matriu de covariàncies) d'aquestes distribucions sigui igual en tots els grups, mentre que el QDA no requereix aquesta restricció. Així doncs es considera que per cada classe y, $X|Y = y \sim N(\mu^{(y)}, \Sigma^{(y)})$, és a dir, que

$$f_y(x) = \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma^{(y)})}} \exp\left(-\frac{1}{2}(x-\mu^{(y)})^T(\Sigma^{(y)})^{-1}(x-\mu^{(y)})\right), \tag{4.18}$$

i substituint (4.18) en (4.7) es té que el mètode de QDA classifica cada individu $X = x$ a la classe $Y = y$ tal que es maximitza

$$\frac{\pi_y \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma^{(y)})}} \exp\left(-\frac{1}{2}(x-\mu^{(y)})^T(\Sigma^{(y)})^{-1}(x-\mu^{(y)})\right)}{\sum_{i=1}^{n} \pi_i \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma^{(i)})}} \exp\left(-\frac{1}{2}(x-\mu^{(i)})^T(\Sigma^{(i)})^{-1}(x-\mu^{(i)})\right)} \tag{4.19}$$

que donat que el logaritme és monòton creixent, el denominador d'aquesta expressió és constant en variar y i hi ha factors constants, és equivalent a maximitzar

$$\begin{aligned} &\ln \pi_y - \frac{1}{2}(x-\mu^{(y)})^T(\Sigma^{(y)})^{-1}(x-\mu^{(y)}) \\ &= \ln \pi_y - \frac{1}{2}x^T(\Sigma^{(y)})^{-1}x + x^T(\Sigma^{(y)})^{-1}\mu^{(y)} - \frac{1}{2}(\mu^{(y)})^T(\Sigma^{(y)})^{-1}\mu^{(y)}, \end{aligned}$$

i substituint $\mu^{(1)}, ..., \mu^{(n)}$, $\pi_1, ..., \pi_n$ i $\Sigma^{(1)}, ..., \Sigma^{(n)}$ pels estimadors corresponents es pot solucionar el problema de classificació.

Abans d'entrar en les maneres amb què es pot determinar la qualitat d'una classificació determinada, una pregunta lògica és la següent: en quines situacions és millor usar LDA i en quines QDA? La resposta a això rau en un concepte anomenat *trade-off* entre biaix i variància[14], del qual no s'ha parlat al llarg del treball però és important. Per entendre'l cal definir els dos conceptes, encara que no sigui d'una forma absolutament rigorosa:

Definició 4.4 (Biaix)**.** *El biaix s'entén com l'error que apareix en intentar aproximar un problema real a partir d'un model més simple.*

Definició 4.5 (Variància)**.** *Sigui $\hat{y}_i = \hat{f}(x_i)$ l'estimació de $y_i = f(x_i)$. Es determina la variància com la variació que tindria $\hat{f}$ en cas que es fes servir una base de dades training diferent. Idealment aquest valor ha de ser petit, perquè valors grans de variància indiquen que un canvi molt petit en les dades provoca un canvi gran en el model.*

En termes generals, a models més flexibles, la variància és major i el biaix és menor. El concepte del *trade-off* s'anomena així pel fet que quan el biaix es redueix la variància tendeix a augmentar, i viceversa. L'objectiu de les aproximacions que es fan, doncs, és trobar un model amb biaix i variància el més baixos possible, permetent així que

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{f}(x_i))^2, \tag{4.20}$$

l'error quadràtic, sigui el mínim possible. En general, l'LDA tendeix a generar classificadors menys flexibles que el QDA, de manera que la seva variància és més baixa i per tant genera classificacions millors i major capacitat de predicció del model. No obstant, cal recordar que l'LDA requereix suposar que la matriu de covariàncies és comuna a totes les classes, i que aquesta suposició sigui

[14] En [10], pp. 29-36, s'amplia la informació sobre el *trade-off* entre variància i biaix.

incorrecta porta problemes, sovint en forma d'un model molt esbiaixat. Generalitzant, és millor fer servir LDA quan la base de dades *training* és relativament petita -provocant que reduir la variància sigui clau-, mentre que si la base de dades és més gran aquesta variància no preocupa tant i es pot assumir el seu augment fent servir QDA.

Amb aquesta anàlisi del QDA acaba el capítol dedicat als mètodes de classificació, i amb ell la part teòrica del treball. A continuació es tracta un cas pràctic d'aplicació de tots aquests mètodes: la classificació de jugadors de Liga EBA, la quarta categoria del bàsquet estatal.

5 Estudi pràctic: classificació dels jugadors de Liga EBA

El cinquè i últim capítol de la memòria està dedicat a l'estudi pràctic realitzat durant el treball: la classificació dels jugadors de la Liga EBA de bàsquet. En aquestes pàgines es presenten les conclusions de l'estudi així com algunes consideracions importants del procés que s'ha seguit. Prèviament és recomanable llegir l'annex B per veure la construcció de la matriu de dades, així com el significat de les variables que es fan servir.

5.1 Anàlisi de components principals i biplot

La primera part de l'estudi busca reduir la dimensió de la matriu de dades per tal d'obtenir una visualització més significativa del núvol de punts que representen els individus. El procés de l'ACP es realitza seguint la mateixa seqüència que es troba al capítol 2, quan s'ha estudiat la teoria darrere del mètode. Abans, però, s'usa una funció del paquet `BasketballAnalyzeR` de R que permet generar una matriu i una representació gràfica amb les correlacions entre les variables de la matriu. Aquesta funció és `corranalysis`, i part de la seva sortida es veu en la imatge 4:

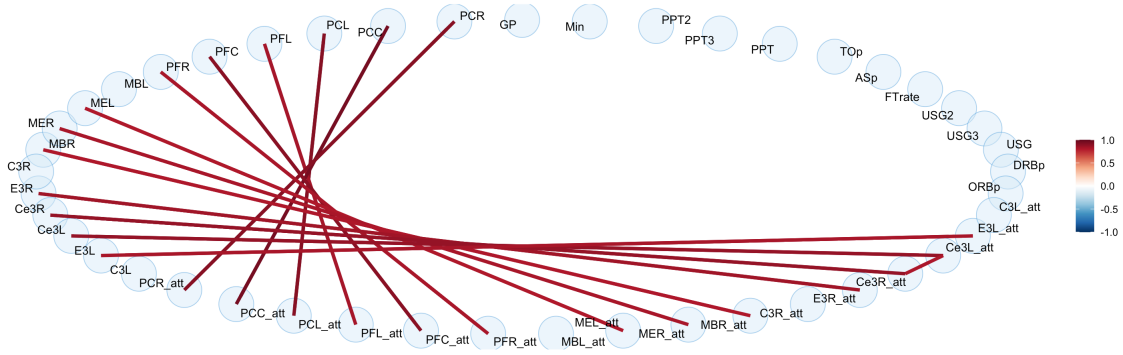


Figura 4: Sortida de la funció `corranalysis` amb les correlacions entre les variables de la matriu de dades.

Es pot observar com les úniques correlacions elevades es donen entre els tirs tirats i anotats des de la mateixa posició, i tot i que això no és ideal no ha reportat problemes durant el procés de clústering, de manera que es pot mantenir aquesta matriu de dades[15].

Un cop construïda aquesta matriu de dades ja es pot començar el procés de l'anàlisi de components principals. El primer que cal notar és que aquestes variables no es troben definides totes en el mateix rang, com es veu en la imatge 5, i que les variàncies de les variables de la matriu de dades són bastant diferents, com es veu en la imatge 6. Aquests dos fets són dos motius de pes per tal de plantejar-se normalitzar la matriu de dades abans de començar l'ACP, procediment que es realitza a continuació.

```
      Min             PPT2            PPT3             PPT             TOp              ASp       
 Min.   :10.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.1333   Min.   : 2.551   Min.   : 0.000  
 1st Qu.:16.10   1st Qu.:0.8064   1st Qu.:0.5714   1st Qu.:0.8088   1st Qu.:12.237   1st Qu.: 7.923  
 Median :20.40   Median :0.9524   Median :0.8182   Median :0.9318   Median :15.160   Median :11.644  
 Mean   :20.64   Mean   :0.9345   Mean   :0.7623   Mean   :0.9175   Mean   :15.661   Mean   :12.570  
 3rd Qu.:25.00   3rd Qu.:1.0830   3rd Qu.:1.0000   3rd Qu.:1.0278   3rd Qu.:18.417   3rd Qu.:15.898  
 Max.   :36.40   Max.   :2.0000   Max.   :3.0000   Max.   :1.4848   Max.   :51.450   Max.   :41.536  
```

Figura 5: Rang i quartils d'algunes de les variables de la matriu de dades.

[15] S'ha intentat arreglar aquest problema calculant percentatges de tir des de cada zona i fent servir una matriu de dades modificada on en lloc dels tirs anotats s'hi posaven aquests percentages. Això ha solucionat els problemes de correlacions elevades, però el clústering resultant no té sentit, de manera que no s'ha aplicat aquest càlcul a l'algoritme definitiu.

```
 [1] 3.514523e+01 4.818936e-02 1.633323e-01 3.105655e-02 2.616013e+01 3.847159e+01 1.332732e-02 6.042043e+01 1.040865e+02
[10] 4.642218e+01 5.301689e+01 7.240197e+01 1.698071e-02 1.315840e-01 3.625972e-01 3.697423e-01 1.682715e-01 1.961399e-02
[19] 5.213992e-02 6.456977e-02 6.959659e-02 5.259284e-02 5.395002e-02 2.354211e-01 5.939072e-02 1.425058e-01 4.328785e-01
[28] 1.145469e-01 4.354676e-03 2.144368e-02 4.530854e-02 4.532123e-02 2.823339e-02 3.862519e-03 1.367420e-02 1.294170e-02
[37] 1.532067e-02 1.380087e-02 1.591453e-02 7.534447e-02 1.719784e-02 6.151022e-02 2.037526e-01 4.576513e-02
```

Figura 6: Variàncies de les variables de la matriu de dades.

Un cop normalitzada la matriu de dades només cal aplicar el Teorema 2.17 que s'ha vist en la teoria de l'ACP i diagonalitzar la matriu de correlacions de les variables amb les quals s'està treballant. Un cop calculada aquesta matriu només cal fer servir la funció `eigen` de R per obtenir la següent sortida, que recull els VAPs de la matriu:

```
eigen() decomposition
$values
 [1] 10.62720213  7.20863153  2.60055334  2.41137125  2.08396863  1.48848318  1.41257122  1.19594977  1.14192009  1.07087301
[11]  1.01846192  0.99365693  0.93498978  0.87753709  0.81316975  0.79186772  0.76188033  0.74264631  0.71479674  0.68351727
[21]  0.62533107  0.60079549  0.55585740  0.46393926  0.38897134  0.31085662  0.28332795  0.22139852  0.20185136  0.18793922
[31]  0.17005606  0.16560255  0.16046073  0.14556371  0.14267323  0.13086128  0.12154483  0.11369908  0.08633371  0.07908367
[41]  0.07551847  0.06367242  0.05841040  0.04529867  0.02690498
```

Figura 7: Valors propis de la matriu de correlacions.

Es pot fer un breu anàlisi del que explica aquesta imatge 7. S'observa com els dos primers valors propis són molt grans, fet que interessa perquè significa que les dues primeres components principals donen la mateixa informació que gairebé 18 de les variables de la matriu original. Per altra banda, però, s'observa com hi ha nou valors propis més que són majors que 1, és a dir, que hi ha nou components principals més que, tot i no ser tan importants com les dues primeres, no són suficientment residuals com per ser rebutjades. Això és un problema per l'ACP, perquè implica un gran nombre de components principals a retenir i per tant que cada component principal explica molt poca variabilitat. En efecte, el gràfic de la Figura 8 mostra aquest fet: es necessiten 16 components principals per arribar al 80% de variabilitat explicada:

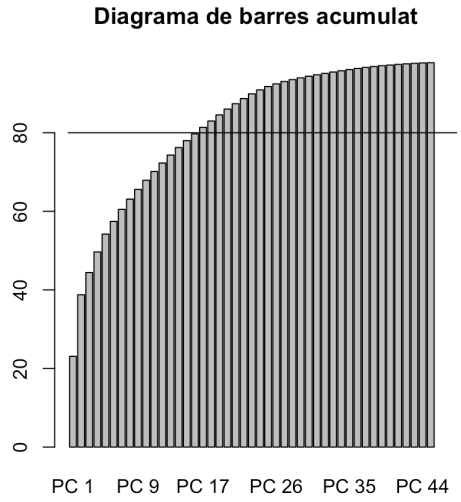


Figura 8: Gràfic de barres amb la variabilitat acumulada de cada component principal.

Això indica que aplicar l'ACP a la matriu completa no té massa sentit si es vol reduir la dimensió de la visualització del núvol de punts, tot i que és útil realitzar-la com a pas previ al clústering. Hi ha dues solucions que s'han provat per tal de poder trobar una visualització amb menys dimensions del núvol de punts:

- ACP amb la matriu no normalitzada. Fent aquesta modificació s'espera que el pes diferent que tenen les variables afecti a la construcció de les components principals fent que les primeres components expliquin més variabilitat del que feien a l'ACP normalitzada. Aquesta solució, no obstant, no ha estat suficient com per arreglar la situació. Com mostra la imatge

10, els resultats són iguals que abans: es tornen a necessitar 16 components principals per assolir el 80% d'inèrcia acumulada, les tres primeres components principals (el límit per a poder fer una representació visual) no arriben a representar el 50% de la variabilitat i a partir de la tercera component principal els increments tornen a ser suficientment grans com per no ser rebutjats però no tant com per aportar informació rellevant. Així doncs, aquesta solució no és útil per tal de visualitzar els individus.

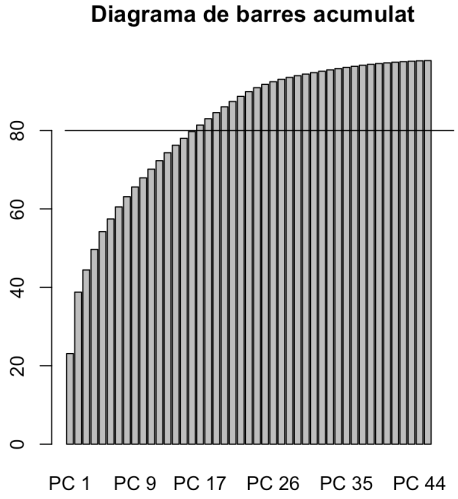


Figura 9: Gràfic de barres amb la variabilitat acumulada de cada component principal en la matriu sense normalitzar.

- ACP amb una matriu amb menys variables. S'intenta en aquest cas buscar quines de les variables de la matriu de dades generen una visualització suficientment representativa per tal de poder observar com es reparteixen els individus en algunes de les variables de la matriu de dades. No és una situació ideal, però poder veure com algunes variables expliquen els jugadors és millor que no poder fer res. Per a fer-ho, es recupera la imatge 6 i es construeix una nova matriu amb les variables que tenen una variància superior a 1. Aquestes són 'GP', 'PPT', 'TO%', 'FTrate', 'USG2', 'USG3', 'USG', 'DRB%' i 'ORB%', i amb elles es construeix una nova matriu que, quan es diagonalitza i es busca la variància acumulada, retorna el següent resultat:

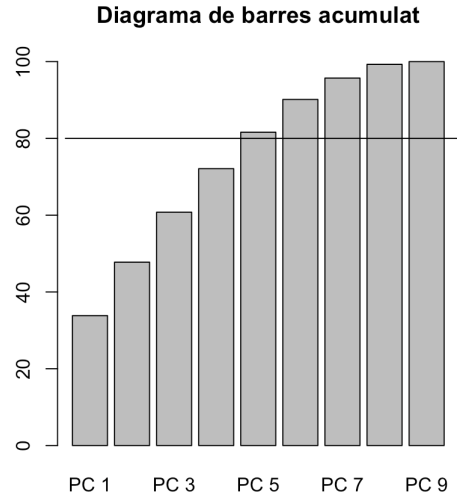


Figura 10: Gràfic de barres amb la variabilitat acumulada de cada component principal en la matriu amb menys variables.

Es pot observar com el resultat segueix sense ésser òptim ja que es necessiten fins a cinc components principals per tal d'assolir el 80% d'inèrcia requerida, però les dues primeres ja expliquen gairebé el 50% d'aquesta inèrcia i els salts en variabilitat d'una component a l'altra són suficientment grans com per aportar informació. És per aquest motiu que l'ACP es realitza amb aquesta matriu simplificada[16], i si es torna a la imatge 4 es pot veure com les correlacions entre aquestes variables són baixes, positiu per l'estudi.

Tot i que el criteri de la inèrcia acumulada aconsella realitzar una ACP amb cinc components principals, es realitza amb només quatre components perquè s'ha usat el criteri de Kaiser (veure secció 2.6). Com que s'ha normalitzat la matriu cal prendre tantes components com VAPs de la matriu de correlacions hi hagi amb valor major a 1, i d'aquests autovalors n'hi ha quatre, com mostra la imatge 11:

```
> eigCr$values
[1] 3.04527283 1.25322399 1.17167647 1.02013681 0.85459608 0.76813824
 0.50128523 0.32180776 0.06386259
```

Figura 11: Valors propis de la matriu de correlacions.

Fent ús de la funció `ACP` del paquet `FactoMineR` es pot obtenir en R tota la informació referent a aquestes components principals, i això és el que es veu a continuació.

Variables millor representades `ACP` permet veure quines són les variables que més pes tenen en cada component principal, és a dir, en funció de quines variables separa millor aquella component principal. Es poden veure aquests resultats en la imatge 12:

USG2	USG	DRBp
26.27830	23.29093	23.21702

(a) Primera CP

TOp	PPT	GP
29.95217	28.15202	20.20753

(b) Segona CP

FTrate	USG3	GP
36.26617	25.90303	15.14633

(c) Tercera CP

FTrate	TOp	USG3
28.18980	22.02168	20.97885

(d) Quarta CP

Figura 12: Variables que millor expliquen cada component principal.

Si s'analitzen només les dues primeres components principals es pot observar com sis de les nou variables que conformen la matriu sobre la qual s'ha executat l'ACP estan ben representades en elles, i les altres dues components principals completen la informació sobre aquestes sis variables a més d'introduir informació extra sobre algunes altres. No es pot fer encara un estudi rigorós de com separen els jugadors aquestes components principals, però aquests valors ja donen una idea del significat de cadascun d'aquests nous eixos.

Representació de les variables En aquest paràgraf es busca determinar la disposició de les variables a partir d'un gràfic anomenat cercle de correlacions: representa cada variable com una fletxa que surt de l'origen de coordenades i acaba al punt $(\rho(CP_1, X), \rho(CP_2, X))$, on ρ denota el coeficient de correlació. Com que s'han conservat quatre components principals, tant per aquesta representació de les variables com pel biplot es realitzaran dues visualitzacions en 2D, una amb les dues primeres components principals que és la que més importància té donat que explica gairebé un 50% de la variabilitat total de la matriu de dades i una altra amb la tercera i la quarta components que afegeix informació a la visualització anterior.
La Figura 13 mostra les correlacions de les variables amb les dues primeres components principals. Observant la imatge es pot veure com la primera component principal separa els individus majoritàriament en funció de les variables 'USG', 'USG2', 'ORB%' i 'DRB%', mentre que la segona component discrimina segons les variables 'PPT' i 'TO%', a més de pel nombre de partits

[16] S'ha provat de realitzar un clústering amb aquesta matriu per veure si, per coherència, es podia realitzar tot el treball amb la mateixa matriu. El resultat no ha estat satisfactori, de manera que es realitzarà el clústering amb la matriu original a la qual s'ha aplicat també una ACP al principi del capítol.

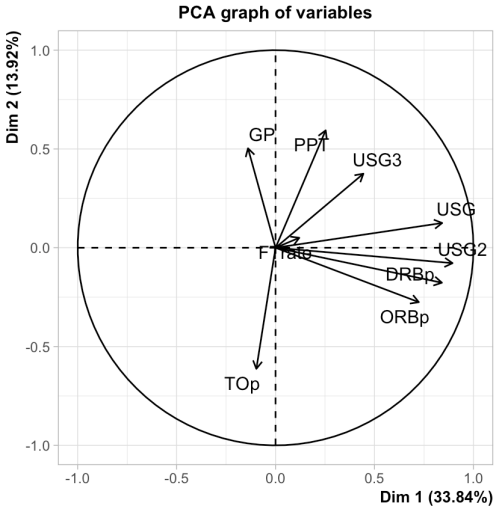


Figura 13: Projecció dels individus sobre les dues primeres CP.

jugats. Fent un anàlisi conjunt de les dues primeres components principals es pot observar que els jugadors més ofensius dels equips i que encistellen amb més efectivitat són els que es troben al primer quadrant del gràfic, donat que en aquesta regió les possessions absorbides i els tirs de tres absorbits són més alts i els tirs de dos estan pràcticament sobre l'eix de la primera component principal. A més, els jugadors en aquest quadrant tenen un baix 'TO%', establint una relació clara (i esperable) entre la qualitat ofensiva d'un jugador i la poca quantitat de pilotes perdudes. Un altre quadrant interessant és el quart quadrant. Segons la disposició de les variables, els jugadors que es troben en aquesta regió del pla tenen valors alts en percentatges de rebot, tirs de 2 absorbits i possessions absorbides, sent més alts com més a la dreta es troben. En la part més propera a l'eix d'ordenades d'aquest quadrant hi ha un altre perfil de jugadors: jugadors amb pocs partits jugats i un alt percentatge de pilotes perdudes, és a dir, amb un rol molt més secundari.

Aquest anàlisi es pot completar amb l'estudi de les dues següents components principals, per arribar a explicar gairebé el 70% de la variabilitat de la matriu de dades. La Figura 14 mostra la distribució de les variables sobre aquests eixos:

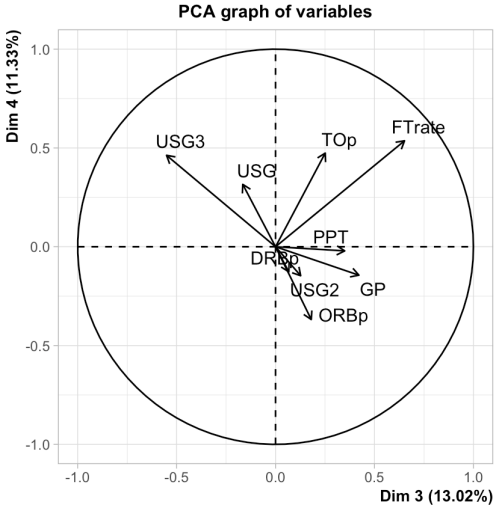


Figura 14: Projecció dels individus sobre la tercera i quarta CP.

Les variables en aquestes dues components principals estan considerablement pitjor representa-

des[17] que en les dues primeres, però tot i això hi ha aspectes interessants a comentar. En primer lloc es pot observar com els jugadors interiors que no són estrelles semblen estar al quart quadrant, ja que és allà on hi ha valors alts de percentatges de rebot i tirs de dos absorbits, però amb baix valor en possessions absorbides, és a dir, jugadors que juguen a prop de cistella però que no tiren gaire. Per altra banda, al segon quadrant s'hi troben els jugadors que tenen com a missió principal dins l'equip el tir de tres, més eficaços com més a la dreta es troben, degut a que la variable 'PPT' (punts per tir) es troba apuntant cap a la dreta gairebé paral·lela a la tercera component principal. S'ha pogut observar, doncs, com la distribució d'aquestes variables ja genera uns quants grups de jugadors més o menys diferenciats, que s'acabaran de completar a la següent secció quan es tracti el clústering de la matriu de dades sencera, ja que cal recordar que aquesta ACP s'està realitzant només amb algunes de les variables.

Biplot Per acabar l'estudi d'aquesta ACP es mostra la representació biplot del núvol de punts. Ja s'ha comentat extensament el que implica la distribució de les variables en el gràfic en el paràgraf anterior, de manera que aquest apartat es fa servir només per poder visualitzar a la pràctica el biplot que s'ha treballat a la secció de teoria, a més de poder observar la distribució dels jugadors en el gràfic. El biplot sobre les dues primeres components principals es mostra en la Figura 15:

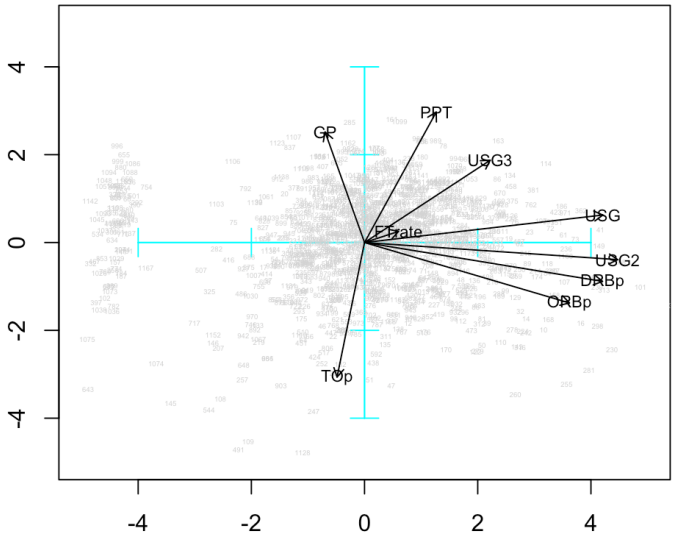


Figura 15: Biplot sobre les dues primeres components principals.

Els jugadors estan representats per cadascun dels números en color gris, mentre que les fletxes corresponen al gràfic de la Figura 13. Es poden observar coses interessants d'aquesta representació, més enllà del seu propi interès com a biplot pel fet que representa individus i variables dins d'un mateix gràfic. Es pot veure en primer lloc com una gran majoria dels jugadors de la lliga es troben dins d'un cercle amb centre a l'origen i radi 2, cosa que indica que aquests jugadors no destaquen en cap dels apartats estadístics que s'expliquen en aquestes dues components principals. Per altra banda, es pot observar un altre grup més o menys nombrós de jugadors a la banda esquerra del gràfic, amb molt poques possessions i tirs absorbits, i amb poc percentatge de rebot. Aquest perfil estadístic s'associa amb jugadors amb un rol molt secundari, i segurament exteriors. És complicat, però, estudiar tots els jugadors del gràfic, de manera que d'ara endavant es farà l'anàlisi dels jugadors d'un sol equip, el GERMANS HOMS - U.E. MATARÓ. En la Figura 16 es pot veure la mateixa imatge que en 15 amb els jugadors d'aquest equip ressaltats. Com es pot observar, es poden repartir els jugadors de l'equip en cinc grups diferents, a jutjar per les dues primeres

[17] A nivell pràctic es pot veure com de ben representada està una variable en funció de com de propera es troba la fletxa al cercle de correlacions. Com més propera és la punta de la fletxa a aquesta corba, més ben representada es troba.

components principals[18]:

- Jugadors 239 i 236. Jugadors molt importants en atac per l'equip, absorbint moltes possessions i gairebé totes elles en tirs de dos punts. Els dos jugadors tenen percentatges de rebot alts, tant en ofensiu com en defensiu, però pel fet d'estar més avall i a la dreta els valors pel jugador 239 són lleugerament més alts que pel 236.
- Jugador 232. Important també per l'equip a nivell ofensiu, però amb més amenaça exterior que els jugadors del grup anterior, i absorbeix menys possessions que ells. A nivell de percentatges de rebot, són significativament més baixos que els jugadors del primer grup.
- Jugadors 234 i 238. Jugadors amb més punts per tir (i per tant millor eficiència ofensiva) de l'equip, consumeixen menys possessions i la majoria dels seus tirs són de tres punts.
- Jugador 237. Jugador amb un perfil ofensiu més secundari, amb menys possessions absorbides que la mitjana i amb la majoria dels seus tirs sent triples. La diferència amb els jugadors de l'últim grup és, a més dels triples, el percentatge de pilotes perdudes, que és baix igual que amb tots els jugadors que s'han analitzat fins ara.
- Jugadors 240, 241, 233 i 235. Els jugadors amb rol més residual dins l'equip en termes de les variables estudiades. Poques possessions, tirs de 2 i tirs de 3 absorbits, i amb una eficiència menor que la mitjana.

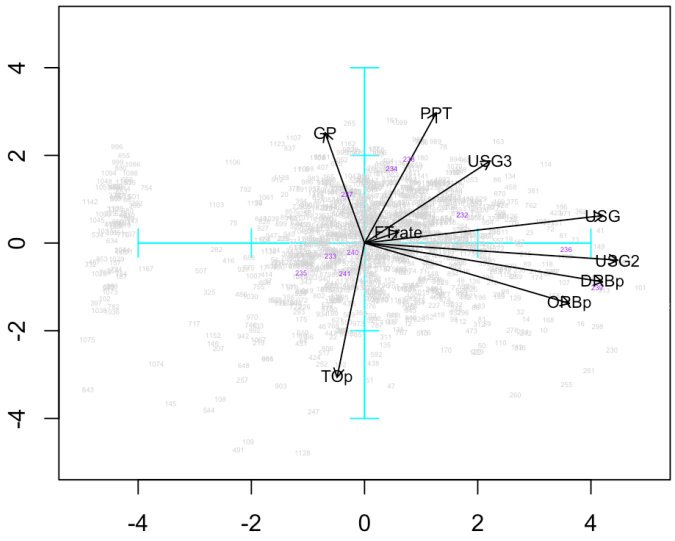


Figura 16: Biplot sobre les dues primeres components principals amb els jugadors de la UE Mataró ressaltats en lila.

Per acabar l'estudi de l'ACP, la Figura 17 mostra els individus sobre la tercera i la quarta components principals. Aquesta imatge permet afirmar, per exemple, que el jugador 241 és molt probablement un jugador interior secundari, ja que es troba a la part de sota de l'eix d'ordenades i cap allà apunten fletxes representant variables com 'ORB%' i 'DRB%', relacionades amb el rebot i per tant normalment característiques de jugadors interiors. També podem veure com hi ha dos jugadors de l'equip, els jugadors 235 i 236, que tiren els tirs lliures millor que la resta dels seus companys, com denota la variable 'FTrate'.

Amb això acaba l'estudi de l'ACP realitzat a la matriu de dades. En el següent apartat, fent ús de mètodes de clústering, es busca consolidar aquesta primera idea dels grups de jugadors als quals s'ha fet esment durant aquesta secció.

[18]Els jugadors estan representats per nombres per conservar el seu anonimat.

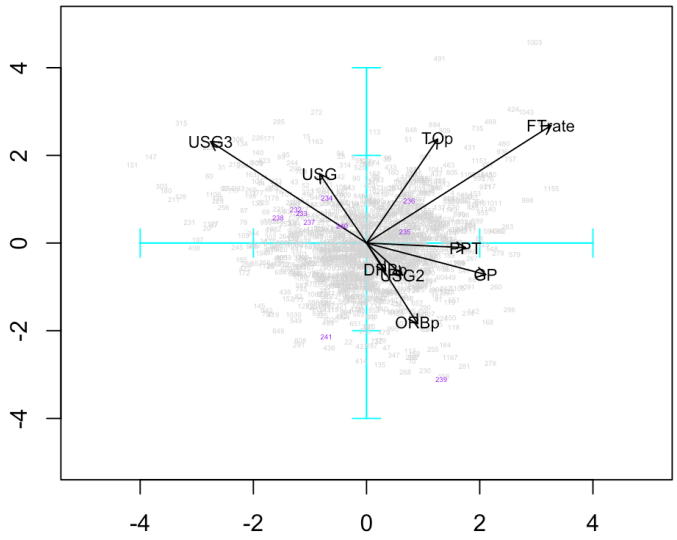


Figura 17: Biplot sobre la tercera i quarta components principals amb els jugadors de la UE Mataró ressaltats en lila.

5.2 Clústering

5.2.1 Selecció del nombre de clústers

La segona part de l'estudi té com a objectiu generar una classificació dels jugadors de la matriu de dades en un nombre de grups segons les seves estadístiques. Durant el bloc teòric de la memòria s'ha comentat alguna vegada el concepte de clústering seqüencial, basat en un clústering jeràrquic que determina el nombre de grups i els centres que inicialitzen el posterior algoritme no jeràrquic, normalment una iteració de *k-means*. Aquest procés és el que s'ha seguit durant el treball, i en el present apartat se'n comenten els resultats.
Muthu Alagappan, en la seva xerrada TED *The new positions of basketball* [27], explica com en la NBA, la lliga de bàsquet americana, es poden diferenciar els jugadors en 10 grups diferenciats fent un estudi similar al del treball, però donat que la NBA és una lliga de nivell extremadament elevat és possible que hi hagi grups de jugadors que existeixin en el seu estudi i no en el que s'està realitzant en aquest projecte, o viceversa. Per aquest motiu es fa ús de la funció `NbClust`, que determina el nombre òptim de clústers entre uns intervals fixats per l'usuari. En el cas d'aquest estudi, com que se sap que el nombre de grups és probable que es trobi entre 10 i 11, es fixaran com a límit inferior nou clústers i com a superior 12. Com mostra la imatge 18, el programa retorna com a nombre òptim nou grups, així que aquest és el valor que es fixa pel clústering jeràrquic que es vol dur a terme a continuació.

```
*******************************************************************
* Among all indices:
* 11 proposed 9 as the best number of clusters
* 9 proposed 10 as the best number of clusters
* 1 proposed 11 as the best number of clusters
* 3 proposed 12 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  9


*******************************************************************
```

Figura 18: Funció `NbClust` amb el nombre òptim de clústers per l'estudi.

5.2.2 Clústering jeràrquic

El primer pas del clústering seqüencial que segueix aquest treball consisteix en un clústering jeràrquic, que es fa seguint el mètode de Ward per calcular les distàncies entre conjunts. R disposa també d'una funció que genera automàticament clústerings jeràrquics, `hclust`, que donada una matriu de distàncies entre els individus calculada prèviament i un mètode de càlcul de distàncies genera aquest clústering. El resultat d'aquest algoritme es mostra en la imatge 19, on s'observa el dendrograma sencer de tots els individus així com les línies blaves que representen la partició generada.

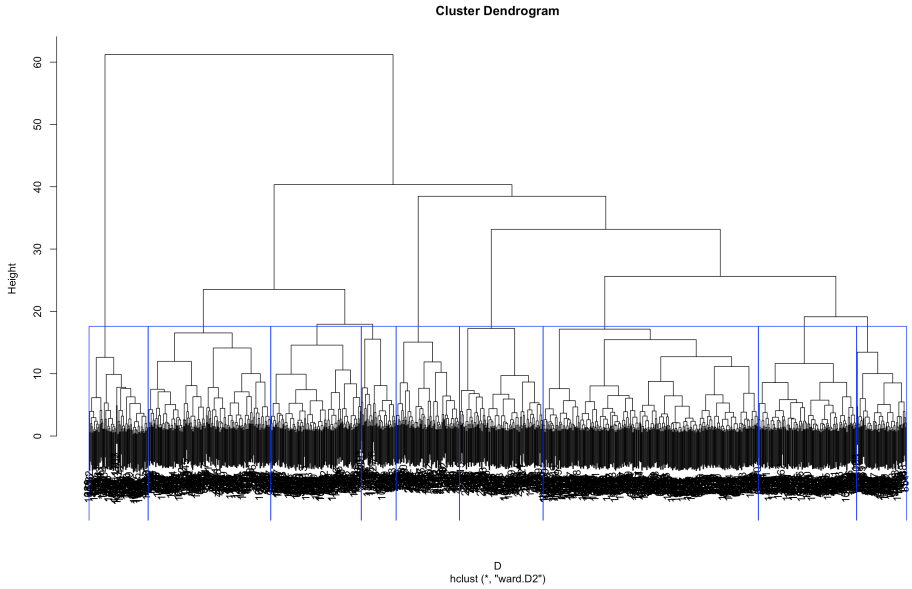


Figura 19: Dendrograma generat per la funció `hclust`.

5.2.3 Consolidació de la partició: clústering no jeràrquic

El clústering no jeràrquic, i en concret l'algoritme *k-means*, és l'última fase del clústering seqüencial que s'ha seguit durant el projecte. Com s'ha comentat al capítol 3, *k-means* necessita dos paràmetres per tal d'ésser inicialitzat: un valor $k \geq 1$ de clústers i un vector de k punts corresponents als centres dels grups en la primera iteració de l'algoritme. Tornant a la imatge 19, s'observen els nou grups diferents que ha retornat el clústering jeràrquic. El codi calcula els centres de cadascun d'aquests grups per tal d'inicialitzar l'algoritme, que es pot reproduir a R gràcies a la funció `k-means`.
Evidentment l'algoritme retorna nou grups de jugadors, ja que *k-means* no varia el nombre de clústers durant les iteracions. Aquests clústers han rebut un nom relacionat amb les seves tendències estadístiques, i es resumeixen en la següent taula[19] [20]:

[19] Les variables representatives estan denotades amb les seves abreviatures per raons d'espai. Veure annex B, taules 2 i 3 pel seu significat.

[20] La taula amb tots els jugadors i el clúster on corresponen és massa extensa per incloure-la a la memòria. És per això que s'ha generat una visualització fent ús de l'aplicatiu Google Data Studio a partir d'una base de dades a Google Drive. S'hi pot accedir a partir dels links https://datastudio.google.com/reporting/db877745-b0c0-4f47-9a6e-f9068a51404c i https://docs.google.com/spreadsheets/d/1JJC1ArHwy8VNzjKtVlN56boOC6IhTccu2vDNjU_duy0/edit?usp=sharing, respectivament.

#	Nom	Variables representatives	
1	Elite back-court	MIN, PPT2, PPT3, PPT, USG2, USG3, USG, DRB%, tirs intentats de fora la zona i amb alt percentatge en tirs de 3 i tirs frontals de 2 punts.	Els millors jugadors exteriors de la lliga. Absorbeixen molts atacs i que són molt eficients en anotació. Capaços d'anotar pràcticament de tot arreu amb alts percentatges, però també d'ajudar en el rebot, especialment defensiu.
2	Outside thread	PPT2, PPT3, PPT, triples intentats de tot arreu i anotats amb alt percentatge.	Jugadors exteriors que destaquen per la seva amenaça en el tir de tres.
3	Mobile bigs	MIN, PPT2, PPT, TO%, FTrate, USG2, USG, DRB%, ORB%, tirs intentats i amb bon percentatge de tot arreu de 2 punts.	Jugadors interiors que tiren molt i de manera molt efectiva des de dins del triple, però fora del triple no tiren -i quan ho fan és amb baix percentatge. Destaquen també en percentatge de rebot, tant ofensiu com defensiu.
4	Space-creating backcourt	AS%, triples tirats i anotats per sota del break[21].	Jugadors exteriors amb perfil de creador d'espais. Destaquen en el percentatge d'assistències i absorbeixen pocs tirs. La majoria d'aquests es donen per sota del break, amb percentatge molt bo.
5	Second unit back-court	TO%, USG3.	Jugadors exteriors amb menys minuts. La gran majoria dels tirs que prenen, que no són gaires, són triples, amb percentatges correctes. Destaca també el percentatge de pilotes perdudes.
6	Assistant ball-handler	AS%, FTrate.	Exteriors amb capacitat de passar però poca amenaça en el tir, motiu pel qual absorbeixen poques possessions. Destaquen per la capacitat de donar assistències i l'alt FTrate donat pel fet que prenen pocs tirs de camp.
7	Second-sword assistant ball-handler	TO%, AS%, FTrate, USG.	Exteriors amb el mateix perfil que els del clúster 6, amb la diferència que absorbeixen més possessions en forma de pilotes perdudes.
8	Assistant ball-handler with shooting ability	MIN, AS%, FTrate, tirs intentats i amb alt percentatge en zones frontals de triple.	Tercer clúster on la principal característica dels jugadors és la capacitat de donar assistències. En aquest cas, però, trobem jugadors amb més minuts de joc que són capaços de tirar de tres, especialment des de zones frontals de la pista.
9	All-around offensive thread	MIN, PPT2, PPT3, PPT, FTrate, USG2, USG3, USG, DRB%, ORB%, tirs tirats i amb percentage alt d'anotació en triple de posicions frontals, tirs intentats i amb alt percentatge de tot arreu de 2 punts.	Aquest clúster és el més especial dels nou, ja que aquí s'hi troba una barreja molt heterogènia de perfils actuals de jugadors, però s'agrupen per la seva capacitat d'anotar tirs de 2 punts amb alts percentatges, així com l'amenaça del seu triple frontal i l'ajuda que aporten a l'equip en el rebot, tant ofensiu com defensiu.

Taula 1: Els nou grups resultants d'aplicar els mètodes de clústering.

[21] El break es defineix com el punt de tall entre la prolongació de la línia de tir lliure i la línia de triple

És també interessant veure el *silhouette plot* d'aquesta classificació, per tal de confirmar si és correcta i té sentit.

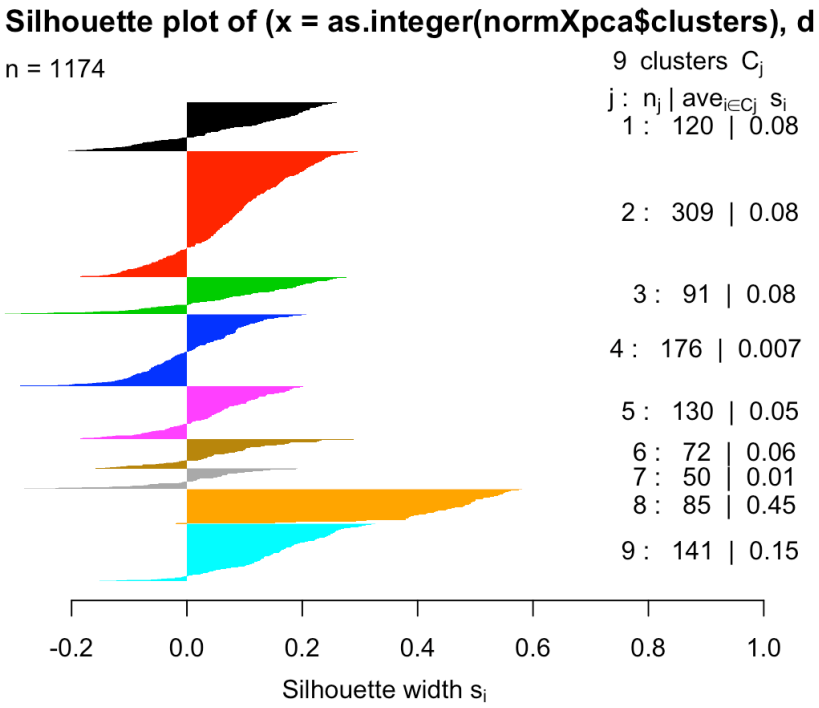


Figura 20: *Silhouette plot* de la classificació en nou grups.

Com s'ha comentat en el capítol 3.4.3 del treball, valors positius indiquen classificacions correctes, sent millor quant més proper sigui a 1. El gràfic mostra, doncs, com la majoria d'individus estan classificats a un grup on corresponen, sent la millor classificació la dels clústers 8 i 9. Així doncs, es pot seguir endavant amb aquests resultats.

5.3 Classificació i anàlisi discriminant

5.3.1 K-Nearest Neighbors

En aquest últim apartat del treball es realitza una introducció a l'aplicació de mètodes de classificació a la base de dades fent una classificació a partir del mètode de KNN. Per tal de simplificar el procés i no haver de crear nou classificadors s'ha executat un estudi previ i s'ha vist que una bona classificació dels individus en menys clústers consisteix en una partició en tres clústers, de manera que es realitzarà un KNN amb l'objectiu de classificar els individus en aquests tres grups[22].
El primer que cal fer per aquest mètode, més enllà de preparar les dades i normalitzar-les d'alguna manera[23], és decidir quin és el valor de k a fer servir. Per això s'ha fet un bucle que executa KNN amb k des d'1 fins a 30 i s'ha estudiat la proporció d'individus mal classificats. El resultat ha estat el següent gràfic, on es determina que la k que proporciona el mínim nombre d'individus mal classificats és $k = 9$, amb prop d'un 86% d'èxit:

[22] S'ha provat de fer una classificació en funció dels nou grups, però per la manca de dades quedaven grups massa petits que no permetien a R completar l'algoritme.

[23] Al codi es fa servir l'anomenada 'normalització min-max', on es fa que totes les variables prenguin valors entre 0 i 1. Una altra opció seria estandarditzar cada variable amb el procés habitual, restant la mitjana i dividint per la desviació típica.

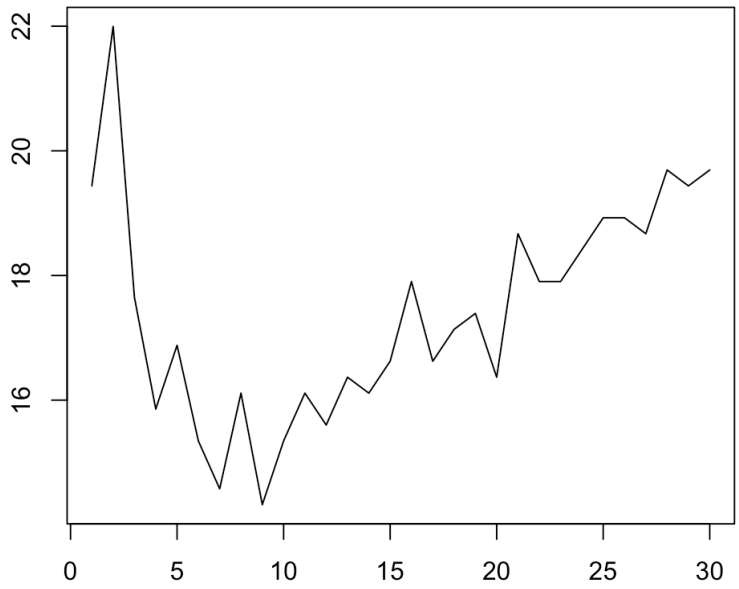


Figura 21: Proporció d'individus mal classificats en funció de k.

La funció `knn` de R també retorna una taula creuada on es pot comparar la predicció feta pel mètode de KNN amb el grup real on pertany l'individu:

	c1[-aleat]			
pred	1	2	3	Row Total
1	229	28	4	261
	0.877	0.107	0.015	0.668
	0.920	0.267	0.108	
	0.586	0.072	0.010	
2	20	77	5	102
	0.196	0.755	0.049	0.261
	0.080	0.733	0.135	
	0.051	0.197	0.013	
3	0	0	28	28
	0.000	0.000	1.000	0.072
	0.000	0.000	0.757	
	0.000	0.000	0.072	
Column Total	249	105	37	391
	0.637	0.269	0.095	

Figura 22: Taula creuada de la classificació real i la predicció amb $k = 9$.

A nivell de resultats no hi ha molt a comentar, ja que la taula els dona de manera bastant clara. Sí es pot, però, passar per sobre de tota la informació que ofereix cada cel·la d'aquesta *crosstable*. En cada cel·la s'hi troben quatre valors:

- El nombre d'individus classificats al clúster que marca la columna on es troba i alhora predits al clúster que indica la fila. Per exemple, hi ha 229 individus que pertanyen al clúster 1 i han estat predits al clúster 1, i 20 individus que pertanyen al grup 1 han estat predits al clúster 2.
- La proporció d'individus respecte del total de la fila. Mirant la fila 1, es té que el 87.7% dels individus predits al clúster 1 han estat predits correctament, el 10.7% pertanyen en realitat

al clúster 2 i l'1.5% restant al grup 3.

- La proporció d'individus respecte del total de la columna. Mirant la primera columna, el 92% dels individus que pertanyen al grup 1 han estat predits correctament, per exemple.
- La proporció d'individus respecte del total d'individus de la base de dades. De nou a la cel·la superior esquerra es pot observar que els 234 individus classificats allà corresponen al 58.6% de la base de dades *testing*.

5.3.2 Anàlisi discriminant quadràtic

Per acabar el treball, s'aplica en aquest apartat el mètode de QDA a la matriu de dades. Cal primer recordar quines són les condicions que la teoria reclama per aquest tipus de mètode: les dades en cada categoria han de seguir una distribució normal, no necessàriament totes amb la mateixa matriu de covariàncies. En l'annex B es veu la distribució de les variables en cada clúster, així com la corba de la normal que haurien de seguir, i com es veu en les imatges cal descartar les variables 'GP', 'USG2', 'USG3', 'USG', 'DRB%' i 'ORB%', ja que en algun dels grups no s'ajusta a la distribució normal que es requereix.
Un cop arreglada la matriu de dades es pot ja dur a terme l'anàlisi discriminant quadràtic. Fent servir la funció `predict` de R es realitza la predicció de la classificació, i s'obté la següent taula:

	Predicció		
Real	1	2	3
1	203	45	1
2	34	71	0
3	25	12	0

Figura 23: Taula creuada de la classificació amb el mètode de QDA.

Aquesta classificació té un error rate (percentatge d'individus mal classificats) de vora un 30%, de manera que el mètode és bastant pitjor que el KNN, i es pot observar també com el grup que pitjor es classifica és el tercer, degut principalment a la poca quantitat d'individus que hi pertanyen. La funció genera també una sèrie de gràfics on apareixen parelles de variables i una coloració del pla en regions segons quin clúster correspon. Per veure'n un exemple[24] es pot observar la Figura 24, on es classifiquen els individus segons les variables 'PPT' i 'PPT2'. Els individus estan representats segons el nombre del clúster on pertanyen, i estan pintats en negre o en vermell segons si la classificació és correcta o no. Pel que fa a les regions del pla, la regió pintada en color blau correspon als individus del clúster 2, el color verd al tercer clúster i la regió taronja al primer grup. Com es pot observar, i com també mostra la taula de la Figura 23, pràcticament no hi ha individus predits en la regió corresponent al tercer clúster, i es pot veure també com un 75% dels jugadors estan ben predits, un resultat no òptim però bastant bo.

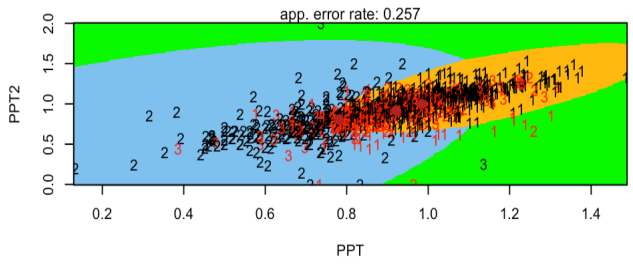


Figura 24: Coloració del pla segons les variables 'PPT' i 'PPT2'.

[24]Degut a la gran quantitat d'imatges que es generen, no es poden incloure totes al projecte. És per això que s'ha creat una carpeta a Google Drive, al link https://drive.google.com/drive/folders/1dnPDoiAvPwW3lsUAKdVL8Bi-H5KeBSs4?usp=sharing, on s'han emmagatzemat totes les imatges per si el lector les vol consultar.

El següent pas, que no es du a terme en aquest treball per falta de dades, és considerar els valors de les variables per individus encara no classificats i, a partir d'aquests, posicionar l'individu en gràfiques com el de la Figura 24 per cada parella de variables.

Amb aquesta introducció al QDA finalitza l'aplicació pràctica dels mètodes vistos al llarg del projecte, i també el propi projecte. Només falta resumir tot el que s'ha vist en un capítol on es treuran les conclusions pertinents i que comença en la pàgina següent.

6 Conclusions

Un cop finalitzat el treball, aquest capítol final té per objectiu exposar les conclusions que s'han extret de l'estudi realitzat. Primer s'analitzarà el treball des d'un punt de vista acadèmic, per acabar amb unes conclusions de caire més personal.

Pel que fa a la secció teòrica del treball considero que ha estat un èxit, ja que s'han pogut estudiar tots els mètodes des de l'ACP fins a l'anàlisi discriminant quadràtic, exposant-ne des de les idees més bàsiques fins a teoremes fonamentals pels mètodes, amb les seves demostracions. Els problemes han arribat en la part pràctica, i des de diferents fronts.
En primer lloc, l'ACP no s'ha pogut realitzar amb la matriu sencera ja que es necessitava una quantitat massa gran de components principals per tal d'assolir el 80% de variabilitat acumulada necessari perquè el mètode sigui suficientment vàlid. S'ha aconseguit solucionar aquest problema reduint el nombre de variables, i tot i que no és la mateixa ACP que s'hauria realitzat amb la matriu sencera sí ha servit per fer-se una idea del mètode així com del biplot, una altra tècnica molt interessant.
Pel que fa al clústering, tot i no ser un clústering amb un valor de silhouette massa proper a 1, és de llarg el mètode que millor s'ha pogut aplicar a la matriu de dades. Els nou grups generats són relativament coherents amb el que es sap del visionat de partits de la lliga i del coneixement de les plantilles dels equips, i això ho demostra també la Figura 20, on es veu que la majoria d'individus estan ben classificats o, com a mínim, a prop de la frontera del clúster.
Per últim, s'ha treballat amb dos mètodes de classificació, el K-Nearest Neighbors i el QDA. En ambdós casos s'ha pogut aplicar l'algoritme, però els resultats no han estat del tot satisfactoris sobretot en termes de l'error rate. Els valors prop del 25% d'error rate indiquen que la classificació és significativament millor que fer-ho a l'atzar, però no és un valor suficientment bo com per assumir que el mètode funciona de manera correcta.

Vist com han funcionat els mètodes a la pràctica, doncs, es pot dir que la segona part del treball ha sigut una bona manera de veure com aplicar els algoritmes estudiats durant la primera part, més teòrica, i com funcionen en un cas real, però l'anàlisi dels resultats, més enllà potser dels que retorna l'algoritme de clústering, no són traslladables a la realitat per la seva falta de consistència. És altament probable que l'error provingui de la selecció de les variables que formen la base de dades, ja que tots els algoritmes s'han aplicat a partir de funcions ja programades. Una altra opció que s'ha valorat és la possibilitat que aquests resultats no òptims vinguin donats perquè s'han usat dades de només una temporada, i més curta de l'habitual, però la Federación Española de Baloncesto ha començat a publicar les dades de tir i del *play-by-play* dels seus partits tot just aquesta temporada.

A nivell personal aquest treball m'ha servit en dos sentits. Per una banda, el rigor requerit en un projecte d'aquestes característiques m'ha permès aplicar moltes de les tècniques de demostració vistes durant el grau, així com recordar resultats i conceptes d'assignatures de cursos previs. Així mateix, m'ha ajudat a seguir millorant els meus coneixements sobre àmbits de les matemàtiques com l'àlgebra lineal i l'anàlisi multivariant, havent de completar algunes de les demostracions que apareixen al llarg de la memòria, que en la seva font d'origen queden com a exercicis pel lector.
Per un altre costat, i més enfocat al meu futur professional, m'ha ajudat a veure diverses maneres d'aplicar mètodes matemàtics a dades relacionades amb l'esport, i més en concret amb el bàsquet. He pogut veure com millorar el rigor dels anàlisis de jugadors que ja he estat elaborant a la UE Mataró aquest any, amb jugadors d'aquesta mateixa lliga. La meva vocació professional és l'anàlisi de dades aplicat a l'esport, i penso que aquest treball és una petita introducció a aquest mon en auge en les recents temporades.

Amb això acaba el treball. Considero que ha estat un èxit relatiu, ja que he pogut estudiar els jugadors de Liga EBA a partir dels mètodes d'anàlisi multivariant i s'han pogut classificar amb un algoritme *k-means* que ha retornat uns resultats que es poden considerar correctes. No obstant això, les dades que s'han escollit per crear la base de dades no han aconseguit generar, i possiblement amb dades diferents els resultats serien millors. També és possible que aquests problemes siguin deguts a que la matriu de dades, amb menys de 2000 jugadors, sigui massa

petita. Això no es podrà saber fins que es juguin més temporades, ja que la Federación Española de Baloncesto ha començat a pujar a la web estadístiques de tir i jugada a jugada aquesta temporada, de manera que les dades de temporades anteriors no serveixen per aquest estudi. Així doncs, considero que aquest treball és un bon punt de partida per, amb una mica més de cura en la selecció de les dades o amb més volum de dades, obtenir resultats molt més satisfactoris.

A Resultats complementaris

Aquest primer annex recull tots els resultats necessaris per tal de completar i/o introduir els que s'han presentat durant el cos principal de la present memòria. Els resultats apareixen en l'ordre corresponent a l'ús que se'n fa durant el treball.

A.1 ACP i biplot

Proposició A.1. *Siguin $X \in \mathcal{M}_{n\times p}$ una matriu de dades, H la matriu de centrat, $1 \in \mathcal{M}_{1\times n}$ un vector columna d'uns i σ la matriu de covariàncies de X. Es compleix:*

(I) $H^T = H$.

(II) $H^2 = H$.

(III) $H1 = 1^T H = 0$.

(IV) Les columnes de HX són centrades.

(V) $\sigma = \frac{1}{n} X^T H X$.

Demostració. (I) és directe de la definició de H (veure Definició 2.1). Pel que fa a la resta de punts:

(II) Cal calcular els valors de $H^2 := (\tilde{h}_{ij})_{1\leq i,j\leq n}$ en dos casos:

a. $i = j$. En aquest cas,

$$\tilde{h}_{ii} = \sum_{k=1}^{n} h_{ik}h_{ki} = \sum_{k\neq i} h_{ik}h_{ki} + h_{ii}h_{ii} = \sum_{k\neq i} \left(-\frac{1}{n}\right)\left(-\frac{1}{n}\right) + \left(\frac{n-1}{n}\right)\left(\frac{n-1}{n}\right)$$
$$= (n-1)\frac{1}{n^2} + \left(\frac{n-1}{n}\right)^2 = \frac{(n-1)+(n-1)^2}{n^2} = \frac{(n-1)(1+n-1)}{n^2} = \frac{n-1}{n} = h_{ii}.$$

b. $i \neq j$. Aleshores,

$$\tilde{h}_{ij} = \sum_{k=1}^{n} h_{ik}h_{kj} = \sum_{k\neq i,j} h_{ik}h_{kj} + h_{ii}h_{ij} + h_{ij}h_{jj} = \sum_{k\neq i,j} \left(-\frac{1}{n}\right)\left(-\frac{1}{n}\right) + 2\left(\frac{n-1}{n}\right)\left(-\frac{1}{n}\right)$$
$$= (n-2)\frac{1}{n^2} - 2\frac{n-1}{n^2} = \frac{(n-2)-2(n-1)}{n^2} = \frac{n-2-2n+2}{n^2} = \frac{-1}{n} = h_{ij}.$$

(III) Cal comprovar els dos productes:

$$H1 = \begin{pmatrix} \sum_{j=1}^n h_{1j} \\ \vdots \\ \sum_{j=1}^n h_{nj} \end{pmatrix} = \begin{pmatrix} \sum_{j\neq 1} h_{1j} + h_{11} \\ \vdots \\ \sum_{j\neq n} h_{nj} + h_{nn} \end{pmatrix} = \begin{pmatrix} \sum_{j\neq 1} -\frac{1}{n} + \frac{n-1}{n} \\ \vdots \\ \sum_{j\neq n} -\frac{1}{n} + \frac{n-1}{n} \end{pmatrix} = \begin{pmatrix} -\frac{n-1}{n} + \frac{n-1}{n} \\ \vdots \\ \frac{n-1}{n} + \frac{n-1}{n} \end{pmatrix} = 0.$$

I per altra banda,

$$1^T H = \begin{pmatrix} \sum_{i=1}^n h_{i1} & \cdots & \sum_{i=1}^n h_{in} \end{pmatrix} = \begin{pmatrix} \sum_{i\neq 1} h_{i1} + h_{11} & \cdots & \sum_{i\neq n} h_{in} + h_{nn} \end{pmatrix}$$
$$= \begin{pmatrix} \sum_{i\neq 1} -\frac{1}{n} + \frac{n-1}{n} & \cdots & \sum_{i\neq n} -\frac{1}{n} + \frac{n-1}{n} \end{pmatrix} = \begin{pmatrix} -\frac{n-1}{n} + \frac{n-1}{n} & \cdots & \frac{n-1}{n} + \frac{n-1}{n} \end{pmatrix} = 0.$$

(IV) $HX = \left(I - \frac{1}{n}A\right)X = X - \frac{1}{n}AX$. Posant $B = AX$ es té que $\forall i, j$,

$$b_{ij} = \sum_{k=1}^{n} a_{ik}x_{kj} = \sum_{k=1}^{n} x_{kj}.$$

Aleshores, fixant la columna j es té que

$$\sum_{i=1}^{n} h_{ij} = \sum_{i=1}^{n}\left(x_{ij} - \frac{1}{n}\sum_{k=1}^{n} x_{kj}\right) = \sum_{i=1}^{n} x_{ij} - \frac{1}{n}\left(\sum_{i=1}^{n}\sum_{k=1}^{n} x_{kj}\right) = \sum_{i=1}^{n} x_{ij} - \frac{1}{n} n \sum_{k=1}^{n} x_{kj} = 0.$$

(V) Primer cal observar que totes dues matrius són en $\mathcal{M}_{p\times p}(\mathbb{R})$. Amb això, σ té a la posició i, j el valor

$$\begin{aligned} Cov(x_i, x_j) &= \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \overline{X_i})(x_{kj} - \overline{X_j}) = \frac{1}{n}\sum_{k=1}^{n}(x_{ki}x_{kj} - x_{ki}\overline{X_j} - \overline{X_i}x_{kj} + \overline{X_iX_j}) \\ &= \frac{1}{n}\sum_{k=1}^{n} x_{ki}x_{kj} - \overline{X_j}\frac{\sum_{k=1}^{n} x_{ki}}{n} - \overline{X_i}\frac{\sum_{k=1}^{n} x_{kj}}{n} + \overline{X_iX_j} = \frac{1}{n}\sum_{k=1}^{n} x_{ki}x_{kj} - \overline{X_iX_j}, \end{aligned}$$

i la matriu $\frac{1}{n}X^THX = \frac{1}{n}X^T\left(I - \frac{1}{n}A\right)X = \frac{1}{n}X^TX - \frac{1}{n^2}X^TAX = \frac{1}{n}X^TX - \frac{1}{n^2}X^TB$ té a la posició i, j:

$$\begin{aligned} &\frac{1}{n}\sum_{k=1}^{n} x_{ik}^T x_{kj} - \frac{1}{n^2}\sum_{k=1}^{n} x_{ik}^T b_{kj} = \frac{1}{n}\sum_{k=1}^{n} x_{ki}x_{kj} - \frac{1}{n^2}\sum_{k=1}^{n}\left(x_{ki}\sum_{l=1}^{n} x_{lj}\right) \\ &= \frac{1}{n}\sum_{k=1}^{n} x_{ki}x_{kj} - \frac{1}{n^2}\left(\sum_{k=1}^{n} x_{ki}\right)\left(\sum_{l=1}^{n} x_{lj}\right) = \frac{1}{n}\sum_{k=1}^{n} x_{ki}x_{kj} - \overline{X_iX_j}. \end{aligned}$$

□

Lema A.2 (Algunes propietats de la covariància)**.** *Siguin $X = (X_1, ..., X_n)$, $Y = (Y_1, ..., Y_n)$, $Z = (Z_1, ..., Z_n)$ i $W = (W_1, ..., W_n)$ quatre variables aleatòries, i $a, b, c, d \in \mathbb{R}$. Aleshores:*

(I) $Cov(X, X) = \mathbb{V}(X)$.

(II) $Cov(X, Y) = Cov(Y, X)$.

(III) $Cov(aX + bY, cZ + dW) = ac \cdot Cov(X, Z) + ad \cdot Cov(X, W) + bc \cdot Cov(Y, Z) + bd \cdot Cov(Y, W)$.

Demostració. Les tres propietats es poden deduir gairebé directament de les definicions de variància i covariància:

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n} \tag{A.1}$$

i

$$\mathbb{V}(X) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} \tag{A.2}$$

A partir de (A.1) i (A.2), (I) i (II) són directes. Només cal provar (III). Per a fer-ho es necessita aquest càlcul: donades A i B variables aleatòries amb n valors presos i α, $\beta \in \mathbb{R}$,

$$\overline{\alpha A + \beta B} = \frac{1}{n}\sum_{i=1}^{n}(\alpha A + \beta B)_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha A_i + \beta B_i) = \alpha\overline{A} + \beta\overline{B}, \tag{A.3}$$

i aleshores

$$\begin{aligned} Cov(aX + bY, cZ + dW) &= \frac{\sum_{i=1}^{n}\left[(aX + bY)_i - \overline{aX + bY}\right]\left[(cZ + dW)_i - \overline{cZ + dW}\right]}{n} \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(aX + bY)_i(cZ + dW)_i - (aX + bY)_i\overline{cZ + dW} - \overline{aX + bY}(cZ + dW)_i\right. \\ &\left. + \left(\overline{aX + bY}\right)\left(\overline{cZ + dW}\right)\right] = \frac{1}{n}\sum_{i=1}^{n}(aX_i + bY_i)(cZ_i + dW_i) - \frac{1}{n}\sum_{i=1}^{n}(aX_i + bY_i)\overline{cZ + dW} \\ &- \frac{1}{n}\sum_{i=1}^{n}\overline{aX + bY}(cZ_i + dW_i) + \frac{1}{n}\sum_{i=1}^{n}\left(\overline{aX + bY}\right)\left(\overline{cZ + dW}\right) = \frac{1}{n}\sum_{i=1}^{n}(acX_iZ_i + adX_iW_i \end{aligned}$$

$$+bcY_iZ_i + bdY_iW_i) - \overline{cZ+dW}\frac{\sum_{i=1}^n(aX_i+bY_i)}{n} - \overline{aX+bY}\frac{\sum_{i=1}^n(cZ_i+dW_i)}{n}$$
$$+\left(\overline{aX+bY}\right)\left(\overline{cZ+dW}\right) = ac\frac{\sum_{i=1}^n X_iZ_i}{n} + ad\frac{\sum_{i=1}^n X_iW_i}{n} + bc\frac{\sum_{i=1}^n Y_iZ_i}{n}$$
$$+bd\frac{\sum_{i=1}^n Y_iW_i}{n} - \left(\overline{aX+bY}\right)\left(\overline{cZ+dW}\right) \tag{A.4}$$

Donat que $\left(\overline{aX+bY}\right)\left(\overline{cZ+dW}\right) = \left(a\overline{X}+b\overline{Y}\right)\left(c\overline{Z}+d\overline{W}\right)$ per A.3, es té que

$$\left(\overline{aX+bY}\right)\left(\overline{cZ+dW}\right) = ac\overline{X}\cdot\overline{Z} + ad\overline{X}\cdot\overline{W} + bc\overline{Y}\cdot\overline{Z} + bd\overline{Y}\cdot\overline{W}, \tag{A.5}$$

i es pot substituir (A.5) en (A.4) per trobar que

$$Cov(aX+bY, cZ+dW) = ac\frac{\sum_{i=1}^n X_iZ_i}{n} + ad\frac{\sum_{i=1}^n X_iW_i}{n} + bc\frac{\sum_{i=1}^n Y_iZ_i}{n}$$
$$+bd\frac{\sum_{i=1}^n Y_iW_i}{n} - ac\overline{X}\cdot\overline{Z} - ad\overline{X}\cdot\overline{W} - bc\overline{Y}\cdot\overline{Z} - bd\overline{Y}\cdot\overline{W}, \tag{A.6}$$

de manera que només cal comprovar que per qualsevol parell de variables -per exemple X i Z- es té que

$$\sum_{i=1}^n X_iZ_i - n\overline{X}\cdot\overline{Z} = \sum_{i=1}^n (X_i-\overline{X})(Z_i-\overline{Z})$$

o, equivalentment, que

$$-n\overline{X}\cdot\overline{Z} = \sum_{i=1}^n\left(-X_i\overline{Z} - \overline{X}Z_i + \overline{X}\,\overline{Z}\right) \Leftrightarrow \sum_{i=1}^n (X_i\overline{Z} + \overline{X}Z_i) = 2n\overline{X}\cdot\overline{Z},$$

i això és fàcil de veure perquè

$$\sum_{i=1}^n X_i\overline{Z} + \sum_{i=1}^n \overline{X}Z_i = n\overline{Z}\frac{\sum_{i=1}^n X_i}{n} + n\overline{X}\frac{\sum_{i=1}^n Z_i}{n} = 2n\overline{X}\cdot\overline{Z}.$$

Aleshores,

$$\sum_{i=1}^n X_iZ_i - n\overline{X}\cdot\overline{Z} = \sum_{i=1}^n (X_i-\overline{X})(Z_i-\overline{Z}), \tag{A.7}$$

$$\sum_{i=1}^n X_iW_i - n\overline{X}\cdot\overline{W} = \sum_{i=1}^n (X_i-\overline{X})(W_i-\overline{W}), \tag{A.8}$$

$$\sum_{i=1}^n Y_iZ_i - n\overline{Y}\cdot\overline{Z} = \sum_{i=1}^n (Y_i-\overline{Y})(Z_i-\overline{Z}), \tag{A.9}$$

i

$$\sum_{i=1}^n Y_iW_i - n\overline{Y}\cdot\overline{W} = \sum_{i=1}^n (Y_i-\overline{Y})(W_i-\overline{W}), \tag{A.10}$$

i recuperant (A.6) es troba que

$$Cov(aX+bY, cZ+dW) = ac\left(\frac{\sum_{i=1}^n X_iZ_i}{n} - \overline{X}\cdot\overline{Z}\right) + ad\left(\frac{\sum_{i=1}^n X_iW_i}{n} - \overline{X}\cdot\overline{W}\right)$$
$$+bc\left(\frac{\sum_{i=1}^n Y_iZ_i}{n} - \overline{Y}\cdot\overline{Z}\right) + bd\left(\frac{\sum_{i=1}^n Y_iW_i}{n} - \overline{Y}\cdot\overline{W}\right) = ac\frac{\sum_{i=1}^n X_iZ_i - n\overline{X}\cdot\overline{Z}}{n}$$
$$+ad\frac{\sum_{i=1}^n X_iW_i - n\overline{X}\cdot\overline{W}}{n} + bc\frac{\sum_{i=1}^n Y_iZ_i - n\overline{Y}\cdot\overline{Z}}{n} + bd\frac{\sum_{i=1}^n Y_iW_i - n\overline{Y}\cdot\overline{W}}{n}$$
$$\overset{(*)}{=} ac\frac{\sum_{i=1}^n (X_i-\overline{X})(Z_i-\overline{Z})}{n} + ad\frac{\sum_{i=1}^n (X_i-\overline{X})(W_i-\overline{W})}{n} + bc\frac{\sum_{i=1}^n (Y_i-\overline{Y})(Z_i-\overline{Z})}{n}$$
$$+bd\frac{\sum_{i=1}^n (Y_i-\overline{Y})(W_i-\overline{W})}{n} = acCov(X,Z) + adCov(X,W) + bcCov(Y,Z) + bdCov(Y,W),$$

on en $(*)$ s'apliquen les igualtats (A.7), (A.8), (A.9) i (A.10). □

Proposició A.3. *Sigui $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ una matriu simètrica. Aleshores, qualssevol dos vectors propis v_1 i v_2 de valors propis λ_1 i λ_2 amb $\lambda_1 \neq \lambda_2$ són ortogonals.*

Demostració. Donat que v_1 és VEP de VAP λ_1 i v_2 ho és de VAP λ_2, es té que $Av_1 = \lambda_1 v_1$ i que $Av_2 = \lambda_2 v_2$. Aleshores

$$(Av_1)v_2 = (\lambda_1 v_1)v_2,$$

i per tant

$$\lambda_1 v_1 v_2 = Av_1 v_2 = (Av_1)^T v_2 = v_1^T A^T v_2 = v_1^T A v_2 = v_1(Av_2) = v_1\lambda_2 v_2 = \lambda_2 v_1 v_2.$$

Aleshores,

$$(\lambda_1 - \lambda_2)v_1 v_2 = 0,$$

però $\lambda_1 - \lambda_2 \neq 0$ per hipòtesi, de manera que $v_1 v_2 = 0 \Leftrightarrow v_1 \perp v_2$. □

Teorema A.4 (Teorema espectral per a matrius simètriques)**.** *Sigui $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ simètrica. Aleshores, A és diagonalitzable ortogonalment.*

Demostració. Abans de provar el teorema s'ha de veure un lema previ:

Lema A.5. *Si $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ és una matriu simètrica, aleshores tots els seus valors propis són reals.*

Demostració. Es considera $A \in \mathcal{M}_{n\times n}(\mathbb{C})$. Si $A \neq 0$ es pot considerar un valor propi de A, $\lambda \in \mathbb{C}$. Es pren també $v \in \mathbb{C}^n$ com el vector propi de valor propi λ, i com a tal es satisfà que

$$Av = \lambda v.$$

Conjugant-ho es té que

$$\overline{Av} = \overline{\lambda v} = \overline{\lambda}\overline{v}.$$

Per altra banda, però, A és una matriu real, i per tant $\overline{A} = A$. Amb això,

$$A\overline{v} = \overline{A}\overline{v} = \overline{Av} = \overline{\lambda v} = \overline{\lambda}\overline{v},$$

i transposant aquesta última igualtat recordant que A és simètrica s'arriba a

$$\overline{v}^T A = \overline{v}^T A^T = (A\overline{v})^T = (\overline{\lambda}\overline{v})^T = \overline{\lambda}\overline{v}^T,$$

de manera que

$$\lambda(\overline{v}^T v) = \overline{v}^T(\lambda v) = \overline{v}^T(Av) = (\overline{v}^T A)v = (\overline{\lambda}\overline{v}^T)v$$

i per tant

$$(\lambda - \overline{\lambda})\overline{v}^T v = 0. \tag{A.11}$$

Com que $v = u + iw = \begin{pmatrix} u_1 + iw_1 \\ \vdots \\ u_n + iw_n \end{pmatrix} \in \mathbb{C}^n$, aleshores

$$\overline{v}^T v = \begin{pmatrix} u_1 - iw_1 & \cdots & u_n - iw_n \end{pmatrix} \begin{pmatrix} u_1 + iw_1 \\ \vdots \\ u_n + iw_n \end{pmatrix} = \sum_{j=1}^{n}(u_j - iw_j)(u_j + iw_j)$$

$$= \sum_{j=1}^{n}(u_j^2 + iu_j w_j - iw_j u_j - i^2 w_j^2) = \sum_{j=1}^{n}(u_j^2 + w_j^2) \neq 0$$

ja que $v \neq 0$ per ser un vector propi, de manera que (A.11) és equivalent a

$$\lambda - \overline{\lambda} = 0,$$

és a dir, que

$$\lambda = \overline{\lambda}$$

i per tant $\lambda \in \mathbb{R}$. □

Ara cal provar el teorema per inducció sobre n, la dimensió de la matriu.

- Si $n = 1$, és clar que la matriu $A = \begin{pmatrix} a_{11} \end{pmatrix} \in \mathcal{M}_{1\times 1}(\mathbb{R})$ és diagonalitzable sobre els reals, amb $Spec(A) = \{a_{11}\} \subset \mathbb{R}$.

- Es suposa cert per $k = n - 1$, i cal provar-ho per n. Es considera la transformació lineal f amb matriu A, això és,

$$\begin{array}{rccc} f: & \mathbb{R}^n & \longrightarrow & \mathbb{R}^n \\ & x & \longmapsto & Ax' \end{array}$$

  i es considera un valor propi $\lambda \in \mathbb{R}$ pel Lema A.5, i el seu vector propi associat v, que es pot considerar unitari.
  Com que $\|v\|_2 = 1$ es pot trobar una base de $\mathbb{R}^n$ completant el conjunt $\{v\}$ amb vectors linealment independents $w_2, ..., w_n$, i es pot transformar en ortonormal per mitjà de la normalització de Gram-Schmidt arribant a $\mathbb{R}^n = \langle v, y_2, ..., y_n \rangle$. Així, es pot construir també una matriu P ortogonal que té en la columna 1 el vector v i en la columna j el vector y_j, $2 \leq j \leq n$. Com que les columnes de P són un conjunt de vectors ortonormals, P és ortogonal.
  Ara, A és simètrica i P ortogonal, i per tant $(Q^{-1}AQ)^T = Q^T A^T (Q^{-1})^T = Q^{-1}AQ$. Així, la primera columna d'aquesta matriu és igual a la primera columna de $Q^{-1}AQ$, és a dir,

$$\begin{pmatrix} \lambda \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

  Així doncs, $Q^{-1}AQ$ es pot escriure per blocs com

$$Q^{-1}AQ = \begin{pmatrix} \lambda & 0^T \\ 0 & B \end{pmatrix},$$

  on $0 \in \mathbb{R}^{n-1}$ és una columna de zeros. Ara, per inducció, com que B és simètrica ($Q^{-1}AQ$ ho és), $\exists R \in \mathcal{M}_{(n-1)\times(n-1)}(\mathbb{R})$ ortogonal i $C \in \mathcal{M}_{(n-1)\times(n-1)}(\mathbb{R})$ simètrica tals que $B = R^{-1}CR$, i per tant

$$Q^{-1}AQ = \begin{pmatrix} \lambda & 0^T \\ 0 & B \end{pmatrix} = \begin{pmatrix} \lambda & 0^T \\ 0 & R^{-1}CR \end{pmatrix} = \begin{pmatrix} 1 & 0^T \\ 0 & R^{-1} \end{pmatrix} \begin{pmatrix} \lambda & 0^T \\ 0 & C \end{pmatrix} \begin{pmatrix} 1 & 0^T \\ 0 & R \end{pmatrix}. \tag{A.12}$$

  Ara, és clar que la matriu del centre és diagonal ja que C ho és, i

$$\begin{pmatrix} 1 & 0^T \\ 0 & R^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0^T \\ 0 & R \end{pmatrix} = \begin{pmatrix} 1 & 0^T \\ 0 & R^{-1}R \end{pmatrix} = I_{n\times n},$$

  i per tant, si es posa $S := \begin{pmatrix} 1 & 0^T \\ 0 & R \end{pmatrix}$ i $D := \begin{pmatrix} \lambda & 0^T \\ 0 & C \end{pmatrix}$, es té de A.12 que

$$Q^{-1}AQ = SDS^{-1} \Leftrightarrow A = QSDS^{-1}Q^{-1} = (QS)D(QS)^{-1},$$

  de manera que A diagonalitza ortogonalment, perquè D és diagonal i com que tant Q com S són ortogonals es té que $(QS)^{-1} = S^{-1}Q^{-1} = S^TQ^T = (QS)^T$, de forma que QS és també ortogonal.

□

Teorema A.6 (Construcció de les components principals II)**.** *Sigui X una matriu de dades de dimensions $n \times p$ amb matriu de covariàncies σ amb algun(s) VAP(s) repetit(s). Aleshores,*

(I) Si $\lambda_1 = ... = \lambda_r > \lambda_{r+1} > ... > \lambda_p$, $\exists! U \subset \mathbb{R}^p$ subespai vectorial de dimensió r tal que $U = Span[\alpha_1, ..., \alpha_r]$, $\forall i$ $\alpha_i \alpha_i^T = 1$ i que maximitza $\mathbb{V}(\varphi_1) + ... + \mathbb{V}(\varphi_r)$.

(II) Si $\lambda_1 > ... > \lambda_s = ... = \lambda_{r+s-1} > \lambda_{r+s} > ... > \lambda_p$, $\exists! U \subset \mathbb{R}^p$ *subespai vectorial de dimensió* r *tal que* $U = Span[\alpha_s, ..., \alpha_{s+r-1}]$, $\forall i \in \{s, ..., s+r-1\}$ $\alpha_i \alpha_i^T = 1$, $\forall i \in \{s, ..., s+r-1\}$ $\forall j < s$ $\alpha_i \alpha_j^T = 0$ *i que maximitza* $\mathbb{V}(\varphi_s) + ... + \mathbb{V}(\varphi_{r+s})$, *prenent el màxim entre totes les possibles* $\varphi_s, ..., \varphi_{r+s}$ *incorrelacionades amb* φ_j $\forall j < s$.

Demostració. Per (I) cal veure que en aquesta situació hi ha r VEPs diferents de la matriu de covariàncies que maximitzen $\mathbb{V}(\varphi_1)$. Repetint el procés fet al Teorema 2.4 s'arriba de nou a (2.4), i a que λ_1 és un VAP de la matriu de covariàncies. En aquest cas, però, λ_1 apareix repetit r vegades i, com que σ és simètrica i per tant diagonalitza sobre els reals (veure Teorema A.4), $dim(ker(\sigma - \lambda_1 I)) = r$. Això es tradueix en que hi ha r VEPs de VAP λ_1 $v_1, ..., v_r$, que es poden suposar unitaris. Construint doncs $\varphi_1, ..., \varphi_r$ seguint (2.1) amb $v_1, ..., v_r$ com a vectors de coeficients, es tindrà que $\forall i \in \{1, ..., r\}$

$$\mathbb{V}(\varphi_i) = v_i \sigma v_i^T = \lambda_i = \lambda_1,$$

perquè v_i és VEP de VAP λ_i i $\lambda_i = \lambda_1$ per $i = 1, ..., r$. Per tant, $\mathbb{V}(\varphi_i)$ és màxim per tot i, ja que pren el valor del VAP més gran de σ i $\mathbb{V}(\varphi_i)$ sempre ha de ser un VAP d'aquesta matriu. Així doncs, $\mathbb{V}(\varphi_1) + ... + \mathbb{V}(\varphi_r)$ és màxim. A més, donat que σ és simètrica i λ_1 és un VAP d'ordre r, aleshores $dim(U) = dim(ker(\sigma - \lambda_1 I)) = r$.

Per provar (II) la situació és exactament igual. El Teorema 2.4 permet trobar les $s-1$ primeres components principals, i quan s'arriba a la s-èsima es troben r possibles VEPs de VAP λ_s, $v_s, ..., v_{r+s-1}$, suposats unitaris. Es construeixen llavors $\varphi_s, ..., \varphi_{r+s-1}$ seguint 2.1 i fent servir $v_s, ..., v_{r+s-1}$ com a vectors de coeficients, i s'arriba a que $\mathbb{V}(\varphi_k) = v_k \sigma v_k^T = \lambda_k = \lambda_s$ $\forall k \in \{s, ..., r+s-1\}$. Aquest és el valor màxim que pot prendre $\mathbb{V}(\varphi_k)$ d'entre tots els VAPs que fan que $v_s, ..., v_{r+s-1}$ siguin ortogonals amb $v_1, ..., v_{s-1}$. En efecte, si hi hagués un valor major que λ_s, aquest hauria de ser algun dels $\lambda_1, ..., \lambda_{s-1}$ (els VAPs majors que λ_s), però llavors els VEPs de VAP λ_s no serien ortogonals amb tots els anteriors. Per tant, $\mathbb{V}(\varphi_s) + ... + \mathbb{V}(\varphi_{s+r-1})$ és màxim d'entre totes les possibles components principals incorrelacionades amb $\varphi_1, ..., \varphi_{s-1}$. □

Proposició A.7 (Normalitat de les columnes de la matriu normalitzada)**.** *Siguin* X *una matriu de dades i* $Z = \Phi(X) = (Z_1, ..., Z_p)$ *la corresponent matriu normalitzada, on* Φ *està definida segons 2.9. Aleshores es compleix:*

(I) Les columnes de Z *són centrades.*

(II) Les columnes de Z *tenen variància 1.*

Demostració. (I) és una conseqüència directa de l'Observació 2.12 i la Proposició A.1(IV), on ha quedat provat que $\overline{(HX)_j} = 0$ $\forall j \in \{1, ..., p\}$:

$$\overline{Z_j} = \frac{1}{\sqrt{\mathbb{V}(X_j)}} \overline{(HX)_j} = 0$$

Per provar (II) n'hi ha prou amb calcular la variància de la columna j de HX, ja que $\sqrt{\mathbb{V}(X_j)}$ és una constant i per tant

$$\mathbb{V}(Z_j) = \mathbb{V}\left(\frac{(HX)_j}{\sqrt{\mathbb{V}(X_j)}}\right) = \frac{1}{\left(\sqrt{\mathbb{V}(X_j)}\right)^2} \mathbb{V}((HX)_j).$$

S'ha vist ja que a la posició i, j de HX es té

$$x_{ij} - \frac{1}{n} \sum_{k=1}^{n} x_{kj}, \tag{A.13}$$

i per tant

$$\mathbb{V}((HX)_j) = \frac{1}{n}\sum_{i=1}^{n}\left[(HX)_{ij} - \overline{(HX)_j}\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(HX)_{ij}\right]^2$$
$$\overset{(A.13)}{=} \frac{1}{n}\sum_{i=1}^{n}\left[x_{ij} - \frac{1}{n}\sum_{k=1}^{n}x_{kj}\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{X_j}\right)^2 = \mathbb{V}(X_j),$$

de manera que efectivament

$$\mathbb{V}(Z_j) = \frac{1}{\mathbb{V}(X_j)}\mathbb{V}(X_j) = 1.$$

□

A.2 Clústering

Proposició A.8. *Sigui S^2 la variància total de la variable X sobre la població. Aleshores,*

$$S^2 = V_{intra} + V_{entre}.$$

Demostració.

$$\begin{aligned}
S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_j - \overline{X}\right)^2 = \frac{1}{n}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(x_j - \overline{X}\right)^2 &= \sum_{h=1}^{H}\sum_{j=1}^{n_h}\frac{1}{n}\frac{n_h}{n_h}\left(x_j - \overline{X}_h + \overline{X}_h - \overline{X}\right)^2 \\
&= \sum_{h=1}^{H}\frac{1}{n_h}\frac{n_h}{n}\sum_{j=1}^{n_h}\left[\left(x_j - \overline{X}_h\right)^2 + 2\left(x_j - \overline{X}_h\right)\left(\overline{X}_h - \overline{X}\right) + \left(\overline{X}_h - \overline{X}\right)^2\right] \\
&= \sum_{h=1}^{H}\frac{1}{n_h}\frac{n_h}{n}\sum_{j=1}^{n_h}\left(x_j - \overline{X}_h\right)^2 + \sum_{h=1}^{H}\frac{1}{n_h}\frac{n_h}{n}\sum_{j=1}^{n_h}2\left(x_j - \overline{X}_h\right)\left(\overline{X}_h - \overline{X}\right) \\
&\quad + \sum_{h=1}^{H}\frac{1}{n_h}\frac{n_h}{n}\sum_{j=1}^{n_h}\left(\overline{X}_h - \overline{X}\right)^2. \qquad \text{(A.14)}
\end{aligned}$$

El primer sumand es pot escriure com

$$\sum_{h=1}^{H}\frac{1}{n_h}\frac{n_h}{n}\sum_{j=1}^{n_h}\left(x_j - \overline{X}_h\right)^2 = \sum_{h=1}^{H}\frac{n_h}{n}\left[\frac{1}{n_h}\sum_{j=1}^{n_h}\left(x_j - \overline{X}_h\right)^2\right] = \sum_{h=1}^{H}\frac{n_h}{n}V_h = V_{intra},$$

el segon com

$$\begin{aligned}
\sum_{h=1}^{H}\frac{1}{n_h}\frac{n_h}{n}\sum_{j=1}^{n_h}2\left(x_j - \overline{X}_h\right)\left(\overline{X}_h - \overline{X}\right) &= \sum_{h=1}^{H}2\frac{1}{n_h}\frac{n_h}{n}\sum_{j=1}^{n_h}\left[x_j\overline{X}_h - x_j\overline{X} - \overline{X}_h^2 + \overline{X}_h\overline{X}\right] \\
&= \sum_{h=1}^{H}2\frac{1}{n_h}\frac{n_h}{n}\left[\sum_{j=1}^{n_h}x_j\overline{X}_h - \sum_{j=1}^{n_h}x_j\overline{X} - \sum_{j=1}^{n_h}\overline{X}_h^2 + \sum_{j=1}^{n_h}\overline{X}_h\overline{X}\right] \\
&= \sum_{h=1}^{H}2\frac{1}{n_h}\frac{n_h}{n}\left[\overline{X}_h\sum_{j=1}^{n_h}x_j - \overline{X}\sum_{j=1}^{n_h}x_j - n_h\overline{X}_h^2 + n_h\overline{X}_h\overline{X}\right]
\end{aligned}$$

$$= \sum_{h=1}^{H} 2\frac{1}{n_h}\frac{n_h}{n}\overline{X}_h \sum_{j=1}^{n_h} x_j - \sum_{h=1}^{H} 2\frac{1}{n_h}\frac{n_h}{n}\overline{X}\sum_{j=1}^{n_h} x_j - \sum_{h=1}^{H} 2\frac{1}{n_h}\frac{n_h}{n} n_h \overline{X}_h^2$$

$$+ \sum_{h=1}^{H} 2\frac{1}{n_h}\frac{n_h}{n} n_h \overline{X}_h \overline{X} = \sum_{h=1}^{H}\left[2\frac{n_h}{n}\overline{X}_h\left(\frac{1}{n_h}\sum_{j=1}^{n_h} x_j\right)\right] - \sum_{h=1}^{H}\left[2\frac{n_h}{n}\overline{X}\left(\frac{1}{n_h}\sum_{j=1}^{n_h} x_j\right)\right]$$

$$- \sum_{h=1}^{H}\frac{2n_h}{n}\overline{X}_h^2 + \sum_{h=1}^{H}\frac{2n_h}{n}\overline{X}_h\overline{X} = \sum_{h=1}^{H}\frac{2n_h}{n}\overline{X}_h^2 - \sum_{h=1}^{H}\frac{2n_h}{n}\overline{X}\overline{X}_h - \sum_{h=1}^{H}\frac{2n_h}{n}\overline{X}_h^2$$

$$+ \sum_{h=1}^{H}\frac{2n_h}{n}\overline{X}_h\overline{X} = 0,$$

i l'últim factor és

$$\sum_{h=1}^{H}\frac{1}{n_h}\frac{n_h}{n}\sum_{j=1}^{n_h}\left(\overline{X}_h - \overline{X}\right)^2 = \sum_{h=1}^{H}\frac{1}{n_h}\frac{n_h}{n} n_h\left(\overline{X}_h - \overline{X}\right)^2 = \sum_{h=1}^{H}\frac{n_h}{n}\left(\overline{X}_h - \overline{X}\right)^2 = V_{entre}.$$

Així doncs, (A.14) es pot posar efectivament com

$$S^2 = V_{intra} + V_{entre}.$$

□

Proposició A.9. *Sigui $X \in \mathcal{M}_{n\times p}(\mathbb{R})$ una matriu de dades i σ la seva matriu de covariàncies. Sigui també $Spec(\sigma) = \{\lambda_1, ..., \lambda_p\}$. Aleshores es satisfà*

$$I_X = \sum_{j=1}^{p}\lambda_j.$$

Demostració. σ diagonalitza sobre els reals pel fet de ser simètrica. Per això, $\exists U \in \mathcal{M}_{p\times p}(\mathbb{R})$, la matriu formada pels VEPs de σ en columnes, que satisfà

$$\sigma = U\Lambda U^{-1}.$$

Per altra banda, la traça d'una matriu és la suma dels seus autovalors, i per tant

$$tr(\sigma) = \sum_{j=1}^{p}\lambda_j = tr(\Lambda).$$

Però σ és la matriu de covariàncies de X, i per tant

$$tr(\sigma) = \sum_{j=1}^{p}\mathbb{V}(X_j) = I_X,$$

de manera que

$$I_X = \sum_{j=1}^{p}\lambda_j.$$

□

Proposició A.10. *El mètode de Ward ajunta a cada pas els clústers C_k i C_l tals que*

$$\frac{n_k n_l}{n_p}\sum_{j=1}^{n}(m_j^k - m_j^l)^2$$

és mínim.

Demostració. El primer que cal veure és que $\forall j \in \{1, ..., n\}$

$$n_p m_j^p = n_k m_j^k + n_l m_j^l. \tag{A.15}$$

En efecte, com que per qualsevol clúster h es té

$$m_j^h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{ij}^h$$

i el clúster C_p és la unió dels clústers C_k i C_l,

$$n_p m_j^p = \sum_{i=1}^{n_p} x_{ij}^p = \sum_{i=1}^{n_p} \left[x_{ij}^p \mathbb{1}_{C_k}(x) + x_{ij}^p \mathbb{1}_{C_l}(x) \right] = \sum_{i=1}^{n_k} x_{ij}^k + \sum_{i=1}^{n_l} x_{ij}^l = n_k m_j^k + n_l m_j^l.$$

Elevant al quadrat les dues bandes de (A.15) s'obté

$$n_p^2 (m_j^p)^2 = (n_p m_j^p)^2 = (n_k m_j^k + n_l m_j^l)^2 = (n_k m_j^k)^2 + (n_l m_j^l)^2 + 2 n_k m_j^k n_l m_j^l. \tag{A.16}$$

Per altra banda,

$$(m_j^k - m_j^l)^2 = (m_j^k)^2 - 2 m_j^k m_j^l + (m_j^l)^2 \Rightarrow 2 m_j^k m_j^l = (m_j^k)^2 + (m_j^l)^2 - (m_j^k - m_j^l)^2,$$

de manera que es pot reescriure (A.16) com

$$\begin{aligned} n_p^2 (m_j^p)^2 &= (n_k m_j^k)^2 + (n_l m_j^l)^2 + n_k n_l \left[(m_j^k)^2 + (m_j^l)^2 - (m_j^k - m_j^l)^2 \right] \\ &= n_k^2 (m_j^k)^2 + n_l^2 (m_j^l)^2 + n_k n_l (m_j^k)^2 + n_k n_l (m_j^l)^2 - n_k n_l (m_j^k - m_j^l)^2 \\ &= n_k (n_k + n_l)(m_j^k)^2 + n_l (n_l + n_k)(m_j^l)^2 - n_k n_l (m_j^k - m_j^l)^2. \end{aligned} \tag{A.17}$$

Com que C_p és el clúster format per la unió de C_k i C_l, es compleix que $n_p = n_k + n_l$. Per tant,

$$n_p^2 (m_j^p)^2 = n_k n_p (m_j^k)^2 + n_l n_p (m_j^l)^2 - n_k n_l (m_j^k - m_j^l)^2.$$

Per últim, dividint als dos costats per n_p^2 es té

$$(m_j^p)^2 = \frac{n_k}{n_p} (m_j^k)^2 + \frac{n_l}{n_p} (m_j^l)^2 - \frac{n_k n_l}{n_p^2} (m_j^k - m_j^l)^2, \tag{A.18}$$

i llavors

$$\begin{aligned} \Delta E = E_p - E_k - E_l &= \left[\sum_{i=1}^{n_p} \sum_{j=1}^{n} \left(x_{ij}^p \right)^2 - n_p \sum_{j=1}^{n} \left(m_j^p \right)^2 \right] \\ &- \left[\sum_{i=1}^{n_k} \sum_{j=1}^{n} \left(x_{ij}^k \right)^2 - n_k \sum_{j=1}^{n} \left(m_j^k \right)^2 \right] - \left[\sum_{i=1}^{n_l} \sum_{j=1}^{n} \left(x_{ij}^l \right)^2 - n_l \sum_{j=1}^{n} \left(m_j^l \right)^2 \right] \\ = \sum_{j=1}^{n} \left[\sum_{i=1}^{n_p} (x_{ij}^p)^2 - \sum_{i=1}^{n_k} (x_{ij}^k)^2 - \sum_{i=1}^{n_l} (x_{ij}^l)^2 \right] &- n_p \sum_{j=1}^{n} (m_j^p)^2 + n_k \sum_{j=1}^{n} (m_j^k)^2 \\ + n_l \sum_{j=1}^{n} (m_j^l)^2 &= n_l \sum_{j=1}^{n} (m_j^l)^2 + n_k \sum_{j=1}^{n} (m_j^k)^2 - n_p \sum_{j=1}^{n} (m_j^p)^2, \end{aligned}$$

que fent servir (A.18) es pot escriure com

$$\begin{aligned} \Delta E &= n_l \sum_{j=1}^{n} (m_j^l)^2 + n_k \sum_{j=1}^{n} (m_j^k)^2 - n_p \sum_{j=1}^{n} \left[\frac{n_k}{n_p} (m_j^k)^2 + \frac{n_l}{n_p} (m_j^l)^2 - \frac{n_k n_l}{n_p^2} (m_j^k - m_j^l)^2 \right] \\ &= n_l \sum_{j=1}^{n} (m_j^l)^2 + n_k \sum_{j=1}^{n} (m_j^k)^2 - \sum_{j=1}^{n} n_k (m_j^k)^2 - \sum_{j=1}^{n} n_l (m_j^l)^2 + \sum_{j=1}^{n} \frac{n_k n_l}{n_p} (m_j^k - m_j^l)^2 \\ &= \frac{n_k n_l}{n_p} \sum_{j=1}^{n} (m_j^k - m_j^l)^2. \end{aligned}$$

Per tant, donat que el mètode de Ward ajunta els dos clústers que facin mínim ΔE, es compleix el que demanava la proposició. □

Proposició A.11. $\forall i \in \Omega,\ -1 \leq s(i) \leq 1$.

Demostració. Es pot fer la prova per casos. Evidentment si $|c_j| = 1$ es compleix, per tant es treballarà sempre sota $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$. Es tenen dues opcions:

- $\max\{a(i), b(i)\} = a(i)$. Sota aquesta condició, $b(i) \leq a(i)$, i per tant

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} = \frac{b(i) - a(i)}{a(i)} = \frac{b(i)}{a(i)} - 1,$$

  i es té que $0 \leq \frac{b(i)}{a(i)} \leq 1$, de manera que

$$0 - 1 \leq s(i) \leq 1 - 1 \Leftrightarrow -1 \leq s(i) \leq 0 \Rightarrow -1 \leq s(i) \leq 1.$$

- $\max\{a(i), b(i)\} = b(i)$. En aquest cas $b(i) \geq a(i)$, així que

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} = \frac{b(i) - a(i)}{b(i)} = 1 - \frac{a(i)}{b(i)},$$

  i es té que $0 \leq \frac{a(i)}{b(i)} \leq 1$, de manera que

$$1 - 0 \geq s(i) \geq 1 - 1 \Leftrightarrow 1 \geq s(i) \geq 0 \Rightarrow -1 \leq s(i) \leq 1.$$

□

A.3 Classificació

Teorema A.12 (Teorema de Bayes)**.** *Sigui $(\Omega, \mathcal{F}, \mathbb{P})$ un espai de probabilitat. Siguin $A \in \mathcal{F}$ un esdeveniment amb probabilitat no nul·la i $\{B_j\}_{j=1}^n \subset \mathcal{F}$ una partició de Ω. Aleshores,*

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(B_i)\mathbb{P}(A|B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B_i)\mathbb{P}(A|B_i)}{\sum_{j=1}^n \mathbb{P}(A|B_j)\mathbb{P}(B_j)}. \tag{A.19}$$

Demostració. Cal abans de començar provar un lema previ, la regla de les probabilitats totals:

Lema A.13 (Regla de les probabilitats totals)**.** *Sigui $(\Omega, \mathcal{F}, \mathbb{P})$ un espai de probabilitat. Siguin $A \in \mathcal{F}$ un esdeveniment i $\{B_j\}_{j=1}^n \subset \mathcal{F}$ una partició de Ω. Aleshores,*

$$\mathbb{P}(A) = \sum_{j=1}^n \mathbb{P}(A|B_j)\,\mathbb{P}(B_j). \tag{A.20}$$

Demostració.

$$A = A \cap \Omega = A \cap \left(\bigcup_{j=1}^n B_j\right) = \bigcup_{j=1}^n (A \cap B_j),$$

on $\{A \cap B_j\}_j$ són conjunts disjunts dos a dos donat que ja ho eren els B_j i $\forall j,\ A \cap B_j \subseteq B_j$. Aleshores, com $\mathbb{P}$ és una mesura additiva pel fet de ser una probabilitat,

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{j=1}^n (A \cap B_j)\right) = \sum_{j=1}^n \mathbb{P}(A \cap B_j) \overset{(*)}{=} \sum_{j=1}^n \mathbb{P}(A|B_j)\mathbb{P}(B_j),$$

on en $(*)$ s'usa la definició de probabilitat condicionada. □

Amb el lema provat és senzill de provar el Teorema de Bayes:

$$\mathbb{P}\left(B_i|A\right) = \frac{\mathbb{P}(B_i \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \cap B_i)}{\mathbb{P}(A)} \overset{(*)}{=} \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} \overset{(**)}{=} \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{n}\mathbb{P}(A|B_j)\mathbb{P}(B_j)},$$

on en $(*)$ s'usa la definició de la probabilitat condicionada i en $(**)$ el Lema A.13. □

Lema A.14. *Sigui $f : A \longrightarrow B$ una funció diferenciable en un punt $x_0 \in A$ i sigui $g : f(A) \longrightarrow C$ una funció diferenciable en $f(x_0)$ i monòtona creixent. Aleshores, si x_0 és màxim de f, x_0 és també màxim de $g \circ f$.*

Demostració. Donat que x_0 és màxim de f, es satisfà que

$$f'(x_0) = 0 \tag{A.21}$$

i

$$f''(x_0) < 0. \tag{A.22}$$

A més, donat que g és monòtona creixent, es té que

$$g'(x) > 0 \ \ \forall x \in f(A). \tag{A.23}$$

Aleshores, es compleixen les següents dues propietats:

$$(g \circ f)'(x) = g'(f(x))f'(x) \Rightarrow (g \circ f)'(x_0) = g'(f(x_0))f'(x_0) \overset{(A.21)}{=} g'(f(x_0)) \cdot 0 = 0 \tag{A.24}$$

i

$$\begin{aligned}(g \circ f)''(x) &= (g'(f(x))f'(x))' = g''(f(x))(f'(x))^2 + g'(f(x))f''(x) \\ \Rightarrow (g \circ f)''(x_0) &= g''(f(x_0))(f'(x_0))^2 + g'(f(x_0))f''(x_0) \overset{(A.21)}{=} g'(f(x_0))f''(x_0) \underset{(A.23)}{\overset{(A.22)}{<}} 0,\end{aligned} \tag{A.25}$$

fent d'x_0 un màxim de $g \circ f$. □

Proposició A.15 (Funció a maximitzar en l'LDA amb una variable predictora)**.**

$$\begin{aligned}\tilde{y} = \arg\max_y p_y(x) &= \arg\max_y \frac{\pi_y \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_y)^2}{\sigma^2}}}{\sum_{i=1}^{n} \pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_i)^2}{\sigma^2}}} \\ \Leftrightarrow \tilde{y} = \arg\max_y \delta_y(x) &= \arg\max_y \left[\ln(\pi_y) - \frac{x^2}{2\sigma^2} + \frac{x\mu_y}{\sigma^2}\right].\end{aligned}$$

Demostració. Sigui

$$p_y(x) = \frac{\pi_y \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_y)^2}{\sigma^2}}}{\sum_{i=1}^{n} \pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_i)^2}{\sigma^2}}} = \frac{\pi_y f_y(x)}{\sum_{i=1}^{n} \pi_i f_i(x)} \tag{A.26}$$

tal i com s'ha definit en (4.7). Es busca quina és la classe $Y = y$ que maximitza aquest valor.
El denominador de l'expressió (A.26) és una funció de x, de manera que fixada x el valor és constant i per tant és equivalent maximitzar (A.26) a fer-ho per

$$q_y(x) = \pi_y f_y(x) = \pi_y \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_y)^2}{\sigma^2}}, \tag{A.27}$$

el seu numerador. Ara, el logaritme és una funció monòtona creixent, de manera que pel Lema A.14 es té que els punts on $q_y(x)$ pren un màxim són els mateixos on $\ln(q_y(x))$ pren un màxim[25]. Així doncs, es busca maximitzar

$$\ln\left(\pi_y \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_y)^2}{\sigma^2}}\right) = \ln(\pi_y) - \ln(\sigma) - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\frac{(x-\mu_y)^2}{\sigma^2}. \tag{A.28}$$

A més, $-\ln(\sigma) - \frac{1}{2}\ln(2\pi)$ és una constant, de manera que n'hi ha prou amb trobar la classe y que maximitza

$$\ln(\pi_y) - \frac{1}{2}\frac{(x-\mu_y)^2}{\sigma^2} = \ln(\pi_y) - \frac{1}{2}\frac{x^2 - 2x\mu_y + \mu_y^2}{\sigma^2} = \ln(\pi_y) - \frac{x^2}{2\sigma^2} + \frac{x\mu_y}{\sigma^2} - \frac{\mu_y^2}{2\sigma^2}. \tag{A.29}$$

I, una vegada més, donat que $-\frac{\mu_y^2}{2\sigma^2}$ és constant, es pot eliminar i només cal maximitzar

$$\delta_y(x) := \ln(\pi_y) - \frac{x^2}{2\sigma^2} + \frac{x\mu_y}{\sigma^2}. \tag{A.30}$$

□

Observació A.16. Arguments anàlegs als vistos en la Proposició A.15 proven que:

- En l'LDA amb més d'una variable, és equivalent maximitzar

$$p_y^{(n)}(x) = \frac{\pi_y \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu^{(y)})^T\Sigma^{-1}(x-\mu^{(y)})\right)}{\sum_{i=1}^n \pi_i \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu^{(i)})^T\Sigma^{-1}(x-\mu^{(i)})\right)} \tag{A.31}$$

  a fer-ho amb

$$\delta_y^{(n)}(x) = \ln\pi_y + x^T\Sigma^{-1}\mu^{(y)} - \frac{1}{2}(\mu^{(y)})^T\Sigma^{-1}\mu^{(y)}. \tag{A.32}$$

- En el QDA també es té una equivalència similar, en aquest cas entre maximitzar

$$\frac{\pi_y \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma^{(y)})}} \exp\left(-\frac{1}{2}(x-\mu^{(y)})^T(\Sigma^{(y)})^{-1}(x-\mu^{(y)})\right)}{\sum_{i=1}^n \pi_i \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma^{(i)})}} \exp\left(-\frac{1}{2}(x-\mu^{(i)})^T(\Sigma^{(i)})^{-1}(x-\mu^{(i)})\right)} \tag{A.33}$$

  i

$$\ln\pi_y - \frac{1}{2}x^T(\Sigma^{(y)})^{-1}x + x^T(\Sigma^{(y)})^{-1}\mu^{(y)} - \frac{1}{2}(\mu^{(y)})^T(\Sigma^{(y)})^{-1}\mu^{(y)}. \tag{A.34}$$

[25]Per tal de poder fer aquesta transformació cal assumir que $p_y(x) > 0$, però en cas que per alguna modalitat de la variable resposta això no passés evidentment aquella no podria ser la màxima i per tant no caldria seguir estudiant-la.

B Informació complementària per l'estudi pràctic

B.1 La matriu de dades

La matriu de dades que es fa servir durant el treball consta d'una sèrie d'observacions relacionades amb diferents àmbits estadístics del bàsquet. Per tal que qualsevol persona, independentment del seu coneixement sobre l'esport, pugui seguir el raonament seguit durant aquest anàlisi pràctic, és interessant definir totes les variables de la matriu de dades. Moltes d'elles es calculen a partir de dades recollides durant els partits, que estan representades per les següents abreviatures[26]:

P2	Tirs de 2 punts anotats.	P2A	Tirs de 2 punts intentats.
P3	Tirs de 3 punts anotats.	P3A	Tirs de 3 punts intentats.
FT	Tirs lliures anotats.	FTA	Tirs lliures intentats.
FG	Tirs de camp anotats. ($FG = P2 + P3$)	FGA	Tirs de camp intentats ($FGA = P2A + P3A$).
TO	Pilotes perdudes.	AS	Assistències.

Taula 2: Variables i les seves abreviatures.

A partir d'aquí, les variables que formaran part de la matriu de dades amb la qual es treballa queden recollides en la següent taula:

Variable	
GP	Partits jugats
Min	Minuts jugats per partit.
PPT2	Punts anotats per cada tir de 2 punts tirat pel jugador. $PPT2 = \frac{2*P2}{P2A}$.
PPT3	Punts anotats per cada tir de 3 punts tirat pel jugador. $PPT3 = \frac{3*P3}{P3A}$.
PPT	Punts anotats per cada tir de camp tirat pel jugador. $PPT = \frac{2*P2+3*P3}{FGA}$.
TO%	Percentatge de les possessions que consumeix un jugador (tirs de camp, tirs lliures, assistències donades i pilotes perdudes) que acaben amb pilota perduda. $TO\% = 100 \cdot \frac{TO}{FGA+0.44 \cdot FTA+AS+TO}$.
AST%	Percentatge de les possessions que consumeix un jugador que acaben amb assistència donada. $AS\% = 100 \cdot \frac{AS}{FGA+0.44 \cdot FTA+AS+TO}$.
FTrate	Ràtio entre els tirs lliures anotats i els tirs de camp tirats per un jugador. $FTrate = \frac{FT}{FGA}$.
USG2	Percentatge dels tirs de 2 de l'equip que assumeix el jugador. $USG2 = 100 \cdot \frac{P2A}{P2A_{eq}}$.
USG3	Percentatge dels tirs de 3 de l'equip que assumeix el jugador. $USG3 = 100 \cdot \frac{P3A}{P3A_{eq}}$.
USG	Percentatge de les possessions de l'equip que assumeix el jugador. $USG = 100 \cdot \frac{FGA+0.44 \cdot FTA+AST+TO}{FGA_{eq}+0.44 \cdot FTA_{eq}+AS_{eq}+TO_{eq}}$.
DRB%	Percentatge dels rebots defensius disponibles que recull un jugador quan està en pista. Els rebots defensius disponibles per un equip seran els seus rebots defensius i els rebots ofensius del rival, i així $DRB\% = 100 \cdot \frac{DRB}{DRB_{eq}+ORB_r}$.
ORB%	Percentatge dels rebots ofensius disponibles que recull un jugador quan està en pista. Els rebots ofensius disponibles per un equip seran els seus rebots ofensius i els rebots defensius del rival, i així $ORB\% = 100 \cdot \frac{ORB}{ORB_{eq}+DRB_r}$.

Taula 3: Variables de la matriu de dades.

[26] Les mateixes abreviatures amb el subíndex "eq" indiquen el valor d'aquella estadística per l'equip al qual pertany el jugador, i amb el subíndex "r" el valor per l'equip rival. Les estadístiques d'equip sempre es computen durant el temps que el jugador està en pista.

A més d'aquestes variables, es codifiquen els tirs dels jugadors segons la seva posició al camp, com mostra la imatge 25. Donada una regió del camp X, es denotaran en la matriu de dades com X el nombre de tirs anotats des d'aquella regió i com X_{att} els tirs intentats des d'allà[27]:

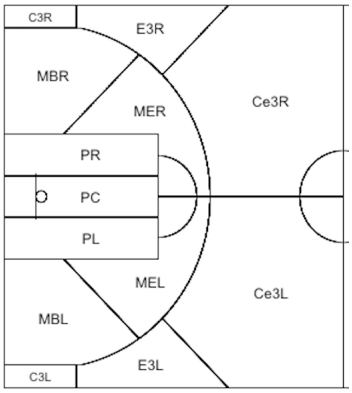


Figura 25: Regions en les quals es divideix la pista.

Ajuntant totes aquestes variables s'obté la matriu de dades de l'estudi, una matriu amb 1395 files, cadascuna corresponent a un jugador de la lliga, i 45 columnes, els noms dels jugadors i 44 variables descriptives.

B.2 Anàlisi discriminant

B.2.1 Distribució de les variables en cada grup

Les següents imatges mostren la distribució d'algunes de les variables en cadascun dels tres grups obtinguts amb el clústering previ, així com la distribució normal a la qual s'haurien d'aproximar: si X és la variable que s'està estudiant, aquesta corba representa la distribució d'una $N(\mu_X, \sigma_X^2)$. Com es pot observar, en les variables 'GP', 'USG2', 'USG3', 'USG', 'DRB%' i 'ORB%' la distribució no s'ajusta a la corba normal, de manera que aquestes no s'usaran per aplicar el QDA.

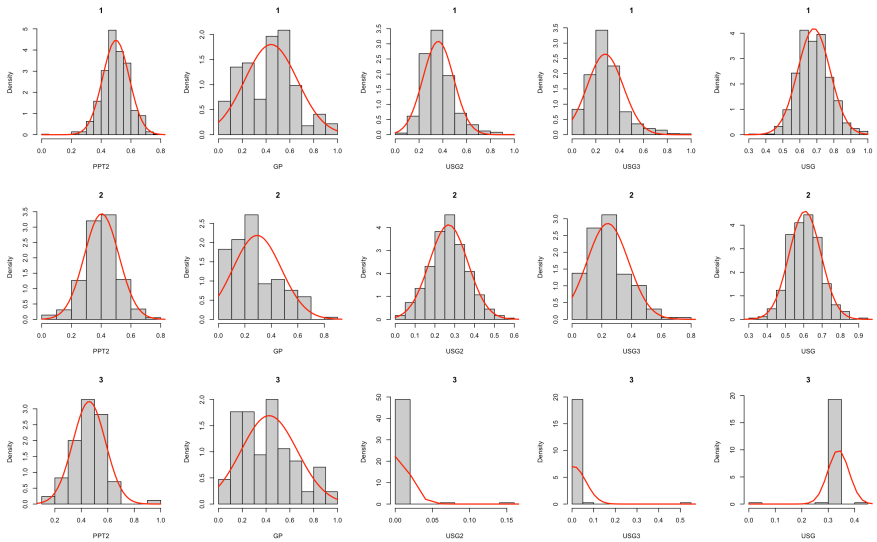


Figura 26: Distribució d'algunes variables en cada clúster i aproximació a la normal corresponent.

[27]Tot i que aquestes variables apareixen a la base de dades, quan es realitzi l'estudi dels resultats no s'estudien una per una, sinó que s'agruparan en regions més grans dins la pista per facilitar la comprensió dels grups.

Aquests gràfics mostren el que s'ha comentat abans: en les variables que es mostren hi ha com a mínim un clúster on les dades no s'ajusten de manera correcta a la distribució normal que haurien de seguir. Per poder comparar, s'han afegit també les distribucions de la variable 'PPT2' en cadascun dels tres clústers, que com es pot veure s'ajusta de manera relativament bona a les distribucions normals. Variables com aquesta són les que es poden fer servir pel QDA.

Referències

[1] Baigorri Matamala, A. (1982). *Contrastes de hipótesis en el modelo de análisis de la varianza multivariante de un factor con efectos aleatorios*. Estadística española, 95, pp.73-83.

[2] Balakrishnama, S. & Ganapathiraju, S. (s. f.). *Linear discriminant analysis - a brief tutorial*. Institute for Signal and Information Processing, MSU.

[3] Burden, R. L., Faires, D. J., & Burden, A. M. (2015). *Numerical Analysis (10th Revised ed.)*. Cengage Learning.

[4] Cuadras, C. M. (2014). *Nuevos métodos de análisis multivariante* (Revisada). CMC Editions.

[5] Cuadras, C. M. (2016). *Problemas de Probabilidades y Estadística. Vol. 1*. Universitat de Barcelona.

[6] Fox, J., & Weisberg, S. (2018). *An R Companion to Applied Regression* (3rd ed.). Sage Publications, Inc.

[7] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* (2nd 2009, Corr. 9th Printing 2017 ed.). Springer.

[8] Husson, F., Le, S., & Pagès, J. (2017). *Exploratory Multivariate Analysis by Example Using R (English Edition) (2a ed.)*. Chapman and Hall/CRC.

[9] Izenman, A. J. (2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning (Springer Texts in Statistics) (English Edition)* (1st ed. 2008, Corr. 2nd printing 2013 ed.). Springer.

[10] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: With Applications in R: 103 (1st ed. 2013, Corr. 7th printing 2017 ed.)*. Springer.

[11] Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis Book 1) (English Edition) (1a ed.)*. STHDA.

[12] Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis: 603*. Wiley-Interscience.

[13] Kostov, B. (2020). *Descripción Multivariante. Mèt. Est. en Mineria de Dades*. DEIO (UPC).

[14] Landau, S., Leese, M., Stahl, D., & Everitt, B. S. (2011). *Cluster Analysis: 848 (5th ed.)*. Wiley.

[15] Lantz, B. (2019). *Machine Learning with R - Third Edition: Expert techniques for predictive modeling*. Packt Publishing.

[16] Montanero, J. (2019). *Manual abreviado de estadística multivariante*. Departamento de Matemáticas. Universidad de Extremadura.

[17] Oliver, D. (2004). *Basketball on Paper: Rules and Tools for Performance Analysis (Illustrated ed.)*. Potomac Books.

[18] Peña, D. (2021). *Análisis multivariante de datos*. McGraw-Hill Interamericana de España S.L.

[19] Ponsa, A. (2019). *El sofá verde. Baloncesto y números: Un paseo por el deporte y la razón (Spanish Edition)*. Independently published.

[20] Redondas, D. (2010). *Análisis multivariante*. UPM.

[21] Rousseeuw, P. (1987). *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20, pp.53-65.

[22] Severini, T. A. (2014). *Analytic Methods in Sports: Using Mathematics and Statistics to Understand Data from Baseball, Football, Basketball, and Other Sports (English Edition) (1a ed.)*. Chapman and Hall/CRC.

[23] Torres, J. (2020). *Python Deep Learning: Introducción práctica con Keras y TensorFlow 2 (1a ed.)*. Marcombo.

[24] Tusell, F. (2005). *Análisis multivariante*. UPV.

[25] Zhang, L., Lu, F., Liu, A., Liu, C. & Guo, P. (2016). Application of K-Means Clustering Algorithm for Classification of NBA Guards. *International Journal of Science and Engineering Applications*, 5(1).

[26] Zuccolotto, P., & Manisera, M. (2020). *Basketball Data Science: With Applications in R*. CRC Press.

[27] The new positions of basketball: Muthu Alagappan at TEDxSpokane. (2013, 8 abril). [Vídeo]. YouTube. https://www.youtube.com/watch?v=E-gpSQQe3w8