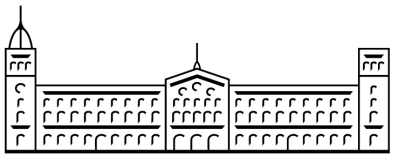

Treball final de grau

GRAU DE MATEMÀTIQUES

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Anàlisi de Components Principals i Discriminant Lineal: una aplicació a resultats acadèmics

---

Autor: David Balboa Gómez

Director:	**Dra. Olga Julià de Ferran**
	Dra. Laura Igual Muñoz
Realitzat a:	**Departament de Probabilitat, Lògica i Estadística**

Barcelona, 29 de juny de 2017

Abstract

Aiming to give an introduction of Data analysis and support the innovation in teaching project of the University of Barcelona which has the purpose of creating an intelligent support system for both the head of studies and the tutors, this work studies the Principal Components Analysis and the Linear Discriminant Analysis; from a theoretical basis in order to understand its mechanics and from a practical implementation, applying them to academic results in order to extract information regarding the abandonment of university studies. The sample used are the degree in Mathematics, Computer Science and Law, for every one of which we considered the grades of the first course taking into account the students that abandoned separated and together with the students that finished. The project has mainly been implemented in Python. For visualization purposes we also used R.

Resum

Amb la intenció d'introduir-se en l'anàlisi de dades, i alhora, de donar suport al projecte d'innovació docent de la Universitat de Barcelona que té com a objectiu crear un sistema de suport intel·ligent tant per al cap d'estudis com pels tutors, aquest treball estudia els mètodes Anàlisi de Components Principals i Anàlisi Discriminant Lineal; tant des d'una basant teòrica per entendre el funcionament d'aquests; com d'una pràctica, implementant-los i aplicant-los a resultats acadèmics per tal d'extreure'n informació sobre l'abandonament als estudis universitaris. La mostra a estudiar són els Graus de Matemàtiques, Enginyeria Informàtica i Dret, pels quals per cada un d'ells s'han considerat les notes de les assignatures de primer i s'han estudiat tant separadament com de forma conjunta els grups d'alumnes que han acabat i els que han abandonat l'ensenyament. La implementació dels mètodes s'ha realitzat mitjançant el llenguatge de programació *Python* i per la visualització s'ha fet servir també l'*R*.

El que no es defineix, no es pot mesurar. El que no es mesura, no es pot millorar. Allò que no es millora, es degrada sempre.

William Thomson, Lord Kelvin, (1824 - 1907) físic i matemàtic britànic.

Índex

1 Introducció

Aquest treball sorgeix de l'interès despertat en l'assignatura *Tallers de Nous Usos de la Informàtica* per la Ciència de les Dades, un camp interdisciplinari que persegueix l'extracció generalitzada de coneixement a partir d'informació representada per dades en totes les seves possibles formes. L'estadística, la mineria de dades, l'aprenentatge automàtic i l'analítica predictiva són alguns dels camps que conformen això que avui dia es coneix com a *Data Science* i que és tant present en el món que ens envolta.

En descobrir-ne que a la facultat existia un projecte d'innovació docent anomenat *Intelligent Support System for Tutor of Studies* [5], del grup consolidat INDOMAIN [6], que treballava amb l'anàlisi de dades multivariants per tal de crear una eina de suport que ajudés als tutors d'estudis a preveure l'abandonament dels seus alumnes, ràpidament el treball va anar agafant forma. Després d'observar els diferents treballs finals de Grau que s'havien fet vinculats a aquest projecte, es va optar per donar-li un enfocament més concret i estudiar-ne només dos dels mètodes més coneguts i importants: l'**Anàlisi de Components Principals** (PCA) i l'**Anàlisi Discriminant Lineal** (LDA).

L'objectiu és per tant estudiar com i per què funcionen aquests mètodes, aplicar-los a les dades que fa servir el projecte d'innovació docent i analitzar-ne els resultats per tal de col·laborar en l'estudi de l'abandonament en els Graus de Matemàtiques, Enginyeria Informàtica i Dret, de la Universitat de Barcelona. Així, amb el mètode PCA reduirem les dades a dues dimensions per tal de visualitzar-les gràficament, estudiar-ne les components principals i fer una primera aproximació a un sistema de classificació que ens permeti predir l'abandonament. I després, amb el mètode LDA buscarem aquesta mateixa classificació i predicció per analitzar-ne els resultats.

Per fer-ho farem servir principalment el llenguatge de programació *Python*, tot i que també ens recolçarem en l'*R* per tal de comprovar o donar suport a les nostres investigacions.

A continuació, en aquest capítol es pretén introduir els conceptes i resultats necessaris per al desenvolupament del treball, així com aproximar al lector a l'estudi de les dades multivariants. Als capítols 2 i 3 s'estudiaran amb profunditat els mètodes PCA i LDA respectivament. Al capítol 4 es comentaran els experiments realitzats i s'analitzaran els resultats. I per acabar, en l'últim capítol s'exposaran les conclusions a les quals s'ha arribat i es proposaran possibles continuacions per aquest treball.

1.1 Què és l'anàlisi multivariant?

L'anàlisi multivariant reuneix l'estadística i el tractament de dades per tal d'estudiar, analitzar, representar i interpretar totes aquelles dades que resulten d'observar més d'una variable estadística sobre una mostra d'individus.

Les tècniques d'anàlisi multivariant van començar a desenvolupar-se per resoldre problemes de classificació en Biologia, però ràpidament es van extendre per trobar variables indicadores (és a dir, un nombre menor de variables construides a partir de les orginials que resumeixin les dades amb la mínima pèrdua d'informació) en Psicometria, Màrqueting i en les Ciències socials. Avui dia tenen aplicacions en tots el camps científics, essent en Enginyeria i Ciències de la computació eines essencials per resumir informació i dissenyar sistemes de classificació automàtica i de reconexiement de patrons.

El seu estudi pot plantejar-se a dos nivells. Al primer, l'objectiu seria utilitzar només les dades disponibles i extreure'n la informació que contenen. Als mètodes que persegueixen aquest objectiu se'ls coneixen com a mètodes d'**exploració de les dades**. Els del segon nivell, anomenats mètodes d'**inferència**, pretenen obtenir conclusions sobre la població que ha generat les dades per tal de construir un model que en pugui fer prediccions.

Així, l'anàlisi multivariant podem definir-lo com un conjunt de mètodes estadístics que tenen com a finalitat analitzar simultàniament conjunts de dades de les quals s'han mesurat un gran nombre de variables.

1.2 Representació de dades multivariants

Suposem que tenim n individus, $\omega_1, \ldots, \omega_n$, dels quals s'han observat p variables, $\chi_1, \ldots, \chi_p$. Per referir-nos a les observacions d'aquestes variables sobre els individus farem servir la notació següent:

$$x_{ij} = \chi_j(\omega_i) \text{ observació de la variable } \chi_j \text{ sobre l'individu } \omega_i,$$
$$\forall i = 1, \ldots, n \; \forall j = 1, \ldots, p$$

Aquesta notació ja indueix que la manera més fàcil i còmode de representar de manera conjunta totes les dades és mitjançant la matriu de dimensió $n \times p$:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1j} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \ldots & x_{ij} & \ldots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nj} & \ldots & x_{np} \end{pmatrix}.$$

A $\mathbf{X}= (x_{ij})$, on les files s'identifiquen amb els individus i les columnes amb les variables, se l'anomena **matriu de dades multivariants**.

Per tal de simplificar i facilitar la lectura, també farem ús de les següents notacions:

1. x_i correspondrà a la fila i-èsima de la matriu $\mathbf{X}$ i l'operarem com un vector columna.

2. X_j denotarà la columna j-èsima d'$\mathbf{X}$.

3. $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_j, \ldots, \bar{x}_p)^t$ serà el vector columna de les mitjanes de les variables, és a dir:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \text{ per } j \in \{1, \ldots, p\}.$$

La seva expressió matricial en funció de la matriu de dades correspon a:

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{1}^t X, \text{ on } \mathbf{1} = (1, \ldots, 1)^t \text{ és el vector columna d'uns d'ordre } n \times 1.$$

1.3 Matrius de dades

A banda de la matriu de dades multivariants ja definida, hi han d'altres matrius vinculades a les dades que apareixeran recurrentment al llarg del treball. Veiem quines són:

1. **Matriu de dades centrades**: és la matriu que resulta de restar a cada observació la mitjana de la variable a la qual pertany:

$$\bar{\mathbf{X}} = (x_{ij} - \bar{x}_j) \ \forall 1 \leq i \leq n, \ \forall 1 \leq j \leq p.$$

2. **Matriu de covariàncies**: és una matriu quadrada i simètrica d'ordre p que recull tant les variàncies de les variables com les covariàncies entre elles:

$$\mathbf{S} = (s_{ij}) \text{ per } i, j \in \{1, \ldots, p\} \text{ on } s_{ij} = \text{cov}(\chi_i, \chi_j) = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j).$$

La matriu de covariàncies també es pot obtenir directament a partir de la matriu de dades centrades:

$$\mathbf{S} = \frac{1}{n}\bar{\mathbf{X}}^t\bar{\mathbf{X}}$$

3. **Matriu de correlacions**: és també una matriu quadrada, simètrica i d'ordre p, que recull les correlacions entre les variables:

$$\mathbf{R} = (r_{ij}) \text{ per } i, j \in \{1, \ldots, p\} \text{ on } r_{ij} = \text{cor}(\chi_i, \chi_j) = \frac{s_{ij}}{s_i s_j}$$

essent s_i, s_j les desviacions típiques de les variables χ_i, χ_j respectivament.

Una forma alternativa d'obtenir la matriu de correlacions, és a partir de la matriu de covariàncies mitjançant l'expressió matricial:

$\mathbf{R} = D^{-1}\mathbf{S}D^{-1}$ on D és una matriu diagonal $p \times p$ amb les desviacions típiques de les variables.

Per tal de remarcar la importància en l'anàlisi multivariant de les dades tant de la matriu de covariàncies $\mathbf{S}$, com de la de correlacions $\mathbf{R}$, anem a veure el següent teorema:

Teorema 1.3.1 (**Teorema de la dimensió**). *Si $r = rang(S) \leq p$, aleshores hi ha r variables linealment independents i les altres $p-r$ són combinació lineal d'aquestes r variables.*

Demostració. Si $\text{rang}(\mathbf{S}) = r$, ordenem adequadament les p variables mesurades per tal que siguin $\chi_1, \ldots, \chi_r$ les que fan que el menor de rang r de $\mathbf{S}$ sigui la matriu de covariàncies S_r:

$$\begin{pmatrix} s_{11} & \ldots & s_{1r} \\ \vdots & \ddots & \vdots \\ s_{r1} & \ldots & s_{rr} \end{pmatrix}.$$

Aleshores, si prenem χ_j amb $j > r$, la fila $(s_{j1}, \ldots, s_{jp})$ és combinació de les primeres r files de la matriu $\mathbf{S}$:

$$(s_{j1}, s_{j2}, \ldots, s_{jp}) = \sum_{k=1}^{r} a_k (s_{k1}, s_{k2}, \ldots, s_{kp})$$

i per tant podem expressar els seus elements com:

$$s_{ji} = \text{cov}(\chi_j, \chi_i) = \sum_{k=1}^{r} a_k \ \text{cov}(\chi_k, \chi_i) = \sum_{k=1}^{r} a_k s_{ki} \quad \forall i \in \{1, \ldots, p\}.$$

Tenint en compte que si una variable té variància zero, aleshores és que és una constant *quasi-segurament*, observem que:

$$\text{var}(\chi_j - \sum_{i=1}^{r} a_i \chi_i) = \text{var}(\chi_j) + \text{var}(\sum_{i=1}^{r} a_i \chi_i) - 2\text{cov}(\chi_j, \sum_{i=1}^{r} a_i \chi_i)$$

$$= s_{jj} + \sum_{i=1}^{r} a_i (\sum_{k=1}^{r} a_k s_{ik}) - 2\sum_{i=1}^{r} a_i s_{ji} = s_{jj} + \sum_{i=1}^{r} a_i (\sum_{k=1}^{r} a_k s_{ki}) - 2\sum_{i=1}^{r} a_i s_{ji}$$

$$= \sum_{i=1}^{r} a_i s_{ji} + \sum_{i=1}^{r} a_i s_{ji} - 2\sum_{i=1}^{r} a_i s_{ji} = 0.$$

Per tant:

$$\chi_j - \sum_{i=1}^{r} a_i \chi_i = c \Rightarrow \chi_j = c + \sum_{i=1}^{r} a_i \chi_i \text{ on } c \text{ és una constant.}$$

□

Corol·lari 1.3.2. *Si totes les variables tenen variància no nul·la (és a dir, cap d'elles es redueix a una constant) i $r = rg(\mathbf{R}) \leq p$, hi han r variables linealment independents i les $p - r$ són combinació lineal d'aquestes r variables.*

Demostració. Del fet que $\mathbf{S} = D\mathbf{R}D$ on D és la matriu diagonal amb les desviaciones típiques de les variables, es té que $r = \text{rg}(\mathbf{R}) = \text{rg}(\mathbf{S})$. □

1.4 Variables compostes i transformacions lineals

Alguns dels mètodes més importants en l'anàlisi multivariant consisteixen a obtenir i interpretar combinacions lineals adequades de les variables observables. És per això que si volem continuar assentant les bases d'aquest, cal que introduïm els següents conceptes.

Definició 1.4.1. *Una **variable composta** Y, és una combinació lineal de les variables observables amb coeficients $a = (a_1, \ldots, a_p)^t$: $Y = a_1\chi_1 + ... + a_p\chi_p$. És a dir, $Y = \mathbf{X}a$.*

Observació 1.4.2. Si $Z = b_1\chi_1 + ... + b_p\chi_p = \mathbf{X}b$, es compleix que:

1. $\bar{Y} = \bar{x}^t a$, $\bar{Z} = \bar{x}^t b$.
2. $\text{var}(Y) = a^t\mathbf{S}a$, $\text{var}(Z) = b^t\mathbf{S}b$.
3. $\text{cov}(Y, Z) = a^t\mathbf{S}b$.

Definició 1.4.3. *Sigui T una matriu $p \times q$. Una **transformació lineal** de la matriu de dades és considerar $Y = \mathbf{X}T$. Les columnes $Y_1, \ldots, Y_q$ de la matriu resultant Y, són les variables transformades de $\mathbf{X}$.*

Proposició 1.4.4. *Les transformacions lineals compleixen que:*

i) $\bar{y}^t = \bar{x}^t T$, on $\bar{y}$ és el vector columna de mitjanes de Y.

ii) $S_Y = T^t\mathbf{S}T$, on S_Y és la matriu de covariàncies de Y.

Demostració. En efecte,

i) $\bar{y}^t = \frac{1}{n}\mathbf{1}^t Y = \frac{1}{n}\mathbf{1}^t XT = \bar{x}^t T$.

ii) $S_Y = \frac{1}{n}Y^t\mathbf{H}Y = \frac{1}{n}(\mathbf{X}T)^t\mathbf{H}(\mathbf{X}T) = \frac{1}{n}T^t\mathbf{X}^t\mathbf{H}\mathbf{X}T = T^t\mathbf{S}T$.

□

1.5 Distribucions multivariants

En l'anàlisi multivariant, les dades acostumen a provenir d'una població caracteritzada per una distribució multivariant. Sigui $\mathbf{X} = (X_1, \ldots, X_p)$ un vector aleatori amb distribució absolutament contínua i funció de densitat $f(x_1, \ldots, x_p)$. És a dir, f compleix:

i) $f(x_1, \ldots, x_p) \geq 0$ per tot $(x_1, \ldots, x_p) \in \mathbb{R}^p$.

ii) $\int_{\mathbb{R}^p} f(x_1, \ldots, x_p) dx_1 \ldots dx_p = 1$.

Si coneixem $f(x_1, \ldots, x_p)$ podrem trobar la funció de densitat de cada variable marginal X_j mitjançant la integral

$$f_j(x_j) = \int f(x_1, \ldots, x_j, \ldots, x_p) dx_1 \ldots dx_{j-1} dx_{j+1} \ldots dx_p.$$

Com en el cas d'una matriu de dades, en les distribucions multivariants també són importants els conceptes següents:

a) vector de mitjanes: $\mu = (E(X_1), \ldots, E(X_p))^t$, on $E(X_j)$ és l'esperança

b) matriu de covariàncies: $\mathbf{\Sigma} = (\sigma_{ij})$ on $\sigma_{ij} = \text{cov}(X_i, X_j)$, $\sigma_{ii} = \text{var}(X_i)$.

Observació 1.5.1. Com que els elements de la matriu $(\mathbf{X} - \mu)(\mathbf{X} - \mu)^t$, d'ordre $p \times p$, són $(X_i - \mu_i)(X_j - \mu_j)$, i $\text{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$, la matriu de covariàncies $\mathbf{\Sigma}$ pot escriure's com:

$$\mathbf{\Sigma} = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^t].$$

Anem a veure dues de les distribucions multivariants més presents en l'anàlisi multivariant.

Definició 1.5.2. *La **distribució normal multivariant no degenerada** d'un vector aleatori $\mathbf{X} = (X_1, \ldots, X_p)$, sorgeix de la generalització de la normal univariant. Així, si la funció de densitat per una variable aleatòria Y amb distribució normal, $Y \sim N(\mu, \sigma^2)$, venia donada per*

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(y-\mu)^2/\sigma^2} = \frac{(\sigma^2)^{-1/2}}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)\frac{1}{\sigma^2}(y-\mu)},$$

la funció de densitat del vector aleatori $\mathbf{X}$ amb distribució normal multivariant, $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$, es defineix:

$$f(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{|\Sigma|^{-1/2}}{(\sqrt{2\pi})^p} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)}.$$

Una manera alternativa de definir la distribució normal multivariant és generalitzant la propietat que si Y és una variable aleatòria amb distribució normal, llavors se satisfà que $Y = \mu + \sigma U$ amb $U \sim N(0,1)$. És a dir que es pot contemplar el vector aleatori $\mathbf{X}$ com una combinació lineal de p variables, $Y_1, \ldots, Y_p$, independents amb distribució $N(0,1)$:

$$\begin{aligned} X_1 &= \mu_1 + a_{11}Y_1 + \cdots + a_{1p}Y_p \\ &\vdots \qquad\qquad\qquad\quad \vdots \\ X_p &= \mu_p + a_{p1}Y_1 + \cdots + a_{pp}Y_p \end{aligned}$$

que expressat matricialment seria

$$\mathbf{X} = \mu + AY,$$

on $Y = (Y_1, \ldots, Y_p)^t$ i $A = (a_{ij})$ és una matriu d'ordre p que compleix $AA^t = \mathbf{\Sigma}$.

Lema 1.5.3. *Les dues definicions anteriors de distribució normal multivariant són equivalents.*

Demostració. Segons la fórmula del canvi de variable

$$f_X(x_1, \ldots, x_p) = f_Y(y_1(x), \ldots, y_p(x))|\frac{\partial y}{\partial x}|,$$

essent $y_i = y_i(x_1, \ldots, x_p)$, $\forall i = 1, \ldots, p$, el canvi i $J = |\frac{\partial y}{\partial x}|$ el jacobià del canvi. Del fet que $\mathbf{X} = \mu + AY$ es té que

$$y = A^{-1}(x - \mu) \Rightarrow |\frac{\partial y}{\partial x}| = |A^{-1}|$$

i com les p variables Y_i són $N(0,1)$ independents:

$$f_X(x_1, \ldots, x_p) = \frac{1}{(\sqrt{2\pi})^p} e^{\frac{1}{2}\sum_{i=1}^p y_i^2} |A^{-1}|$$

Però com $\mathbf{\Sigma}^{-1} = (A^{-1})^t(A^{-1})$, llavors

$$y^t y = (x-\mu)^t (A^{-1})^t (A^{-1})(x-\mu) = (x-\mu)^t \mathbf{\Sigma}^{-1}(x-\mu)$$

i substituint en l'expressió anterior es té la igualtat. □

Observació 1.5.4. Sigui $\mathbf{X}$ un vector aleatori tal que $\mathbf{X} \sim N(\mu, \mathbf{\Sigma})$. Algunes de les propietats més importants que compleix $\mathbf{X}$ per tenir aquesta distribució són les següents:

1. La distribució de cada variable marginal X_i és normal univariant.
2. Tota combinació lineal de les variables $X_1, \ldots, X_p$: $Y = b_0 + b_1X_1 + \cdots + b_pX_p$, és també normal univariant.

3. Si $\mathbf{\Sigma} = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp})$ és una matriu diagonal, aleshores les variables $X_1, \ldots, X_p$ són estocàsticament independents:

$$f(x_1, \ldots, x_p; \mu, \mathbf{\Sigma}) = f(x_1; \mu_1, \sigma_{11}) \times \cdots \times f(x_p; \mu_p, \sigma_{pp})$$

Definició 1.5.5. *Si les files d'una matriu aleatòria $Z_{n\times p}$ són independents amb distribució $N_p(0, \mathbf{\Sigma})$, aleshores direm que la matriu $\mathbf{Q} = Z^t Z$ segueix una* ***distribució Wishart*** *$W_p(\mathbf{\Sigma}, n)$ amb paràmetres $\mathbf{\Sigma}$ i n graus de llibertat.*

Quan $\mathbf{\Sigma}$ és definida positiva i $n \geq p$, la funció de densitat de $\mathbf{Q}$ és:

$$f(\mathbf{Q}) = c|\mathbf{Q}|^{n-p-1} e^{-\frac{1}{2} tr(\mathbf{\Sigma^{-1}Q})}$$

$$\text{on, } c^{-1} = 2^{np/2} \pi^{p(p-1)/4} |\mathbf{\Sigma}|^{n/2} \prod_{i=1}^{p} \Gamma[\tfrac{1}{2}(n+1-i)].$$

Observació 1.5.6. Veiem algunes de les propietats lligades a aquesta distribució:

1. Si Q_1, Q_2 són independents Wishart $W_p(\mathbf{\Sigma}, m), W_p(\mathbf{\Sigma}, n)$, aleshores la suma $Q_1 + Q_2$ també és Wishart $W_p(\mathbf{\Sigma}, m+n)$.

2. Si Q és $W_p(\mathbf{\Sigma}, n)$, separem les p variables en dos conjunts de p_1 i p_2 variables i considerem les particions corresponents de $\mathbf{\Sigma}$ i Q:

$$\mathbf{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix},$$

   aleshores Q_{11} és $W_{p1}(\mathbf{\Sigma}_{11}, n)$ i Q_{22} és $W_{p2}(\mathbf{\Sigma}_{22}, n)$.

3. Si Q és $W_p(\mathbf{\Sigma}, n)$ i T és una matriu $p \times q$ de constants, aleshores $T^t Q T$ és $W_q(T^t \mathbf{\Sigma} T, n)$. En particular, si u és un vector, llavors:

$$\frac{u^t Q u}{u^t \mathbf{\Sigma} u} \sim \chi^2_n$$

   on χ^2_n correspon a la **distribució ki-quadrat de Pearson** amb n graus de llibertat, és a dir, a la suma de n variables aleatòries independents amb distribució $N(0,1)$.

1.6 Descomposició singular

Així com per les matrius simètriques existeix una descomposició en les seves fonts de variació intrínseques, és a dir, en les direccions dels vectors propis amb coeficients que depenen dels valors propis, per les matrius rectangulars també es té una descomposició similar.

Definició 1.6.1. *Sigui A una matriu d'odre $n \times p$ amb $n \geq p$. S'anomena* ***descomposició en valors singulars*** *d'A a:*

$$A = U D^{1/2} V' \text{ on:}$$

i) $D^{1/2}$ *és una matriu diagonal* $p \times p$ *que conté les arrels quadrades dels valors propis no nuls de les matrius* AA^t *o* A^tA. *Aquests termes, són els anomenats* ***valors singulars*** *de la matriu* A.

ii) U *és una matriu* $n \times p$ *que conté en columnes els vectors propis associats als valors propis no nuls d'*AA^t *i a més, satisfà:* $U^tU = \mathbf{I}$.

iii) V *és una matriu* $p \times p$ *que conté en columnes els vectors propis associats als valors propis no nuls d'*A^tA *i a més, satisfà:* $VV^t = V^tV = \mathbf{I}$.

Observació 1.6.2. Si $n = p$ i A és simètrica, aleshores $U = V$ i $A = UD^{1/2}U^t$ és la descomposició espectral de la matriu A. Els valors singulars seran els valors propis d'A.

La descomposició en valors singulars té una gran importància pràctica en l'anàlisi multivariant perquè pot demostrar-se que la millor aproximació de rang $r < p$ a la matriu $\mathbf{X}$, s'obté prenent els r valors propis més grans de $\mathbf{X^tX}$, els corresponents vectors propis de $\mathbf{XX^t}$ i $\mathbf{X^tX}$, i construint:

$$\hat{\mathbf{X}} = U_r D_r^{1/2} V_r^t$$

on $U_r, D_r^{1/2}, V_r$ són tal com s'han definit en la descomposició singular tenint en compte la restricció de la dimensió r: U_r és $n \times r$, $D_r^{1/2}$ és $r \times r$ i V_r és $p \times r$.

La demostració d'aquest fet passa per considerar un espai de dimensió r definit per una base ortonormal U_r, on U_r és $p \times r$ i satisfà $U_r^tU_r = \mathbf{I}$, amb la qual es vol trobar una aproximació de la matriu $\mathbf{X}$. És a dir, es vol expressar cada una de les files $(\mathrm{x}_1, \ldots, \mathrm{x}_n)$ de la matriu, essent x_i el vector $p \times 1$ d'observacions de l'individu i de la mostra, mitjançant els vectors d'U_r.
La predicció de la variable x_i serà la projecció ortogonal sobre l'espai generat per aquests vectors:

$$\hat{\mathrm{x}}_i = U_r U_r^t \mathrm{x}_i$$

per tant es volen determinar els vectors d'U_r tals que l'error quadràtic d'aproximació per totes les files de la matriu sigui mínim. Aquest error ve donat per:

$$E = \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \hat{x}_{ij})^2 = \sum_{i=1}^{n} (\mathrm{x}_i - \hat{\mathrm{x}}_i)^t (\mathrm{x}_i - \hat{\mathrm{x}}_i),$$

però com es pot escriure:

$$E = \sum_{i=1}^{n} \mathrm{x}_i^t \mathrm{x}_i - \sum_{i=1}^{n} \mathrm{x}_i^t U_r U_r^t \mathrm{x}_i,$$

minimitzar-lo és equivalent a maximitzar el segon terme.
Si es té present ara el fet que un escalar és igual a la seva traça i que si A, B són dues matriu $n \times m$, $m \times n$, aleshores se satisfà $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, s'observa que

$$\sum_{i=1}^{n} \mathrm{x}_i^t U_r U_r^t \mathrm{x}_i = \mathrm{tr}(\sum_{i=1}^{n} \mathrm{x}_i^t U_r U_r^t \mathrm{x}_i) = \sum_{i=1}^{n} \mathrm{tr}(\mathrm{x}_i^t U_r U_r^t \mathrm{x}_i) = \sum_{i=1}^{n} \mathrm{tr}(U_r U_r^t \mathrm{x}_i \mathrm{x}_i^t)$$

$= \mathrm{tr}(U_r U_r^t \sum_{i=1}^{n} \mathrm{x}_i \mathrm{x}_i^t)$ i substituint $\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{x}_i \mathrm{x}_i^t$ en l'equació, s'obté

$$\sum_{i=1}^{n} \mathrm{x}_i^t U_r U_r^t \mathrm{x}_i = n\mathrm{tr}(U_r U_r^t \mathbf{S}) = n\mathrm{tr}(U_r^t \mathbf{S} U_r).$$

Que ens diu llavors que per minimitzar l'error s'ha de trobar un conjunt de vectors $U_r = [\mathrm{u}_1, \ldots, \mathrm{u}_r]$ que maximitzin la suma dels elements diagonals de $U_r^t \mathbf{S} U_r$, és a dir, $\sum_{j=1}^{r} \mathrm{u}_j^t \mathbf{S} \mathrm{u}_j$. Veurem al capítol de l'Anàlisi de Components Principals que per maximitzar aquesta expressió s'arriba als vectors propis corresponents als r valors propis més grans, tal com s'havia indicat en la definició.

2 Anàlisi de Components Principals

Un dels principals problemes al qual ha de fer front l'anàlisi de dades multivariants és el de la reducció de la dimensionalitat. Aconseguir reduir el nombre de variables implicades en el problema sense perdre gaire informació sobre aquest, és crucial, no només perquè simplifica l'anàlisi i estalvia temps de càlcul, sinó també perquè ens ofereix la possibilitat de, en cas que siguin poques, representar-les gràficament i comparar fàcilment diferents conjunts de dades o instants en el temps.

L'anàlisi de components principals (PCA, *Principal Component Analysis*) és un d'aquests mètodes que serveixen a la reducció de la dimensionalitat. El seu descobriment se li atribueix a Harold Hotelling (1885 - 1973), que inspirat per Karl Pearson (1857 - 1936) i la seva primera aproximació al 1921 al problema d'obtenir un millor indicador que resumís un conjunt de variables, va generalitzar al 1933 la idea de components principals introduïnt l'anàlisi de correlacions canòniques i permetent així resumir simultàniament dos conjunts de variables.

La idea que hi ha darrere del PCA és realitzar una transformació lineal sobre les dades multivariants per tal de trobar de manera seqüencial les $r \leq p$ variables incorrelacionades que recullin la variabilitat més gran possible, les anomenades *components principals*. Aquestes, ens les podem veure per tant com un nou sistema de coordenades ortogonals sobre el qual les dades podran representar-se en dimensió inferior. Això, i el fet que sempre es farà respectant la màxima variabilitat possible, atorguen a l'anàlisi de components principals la possibilitat tant de ser aplicat com un mètode de classificació no supervisat (és a dir, sense informació a priori de les classes a les quals pertanyen les dades), com de poder interpretar les dades a partir d'aquestes noves variables que s'han obtingut com a combinacions lineals de les originals.

Tot i que el PCA s'aplica partint de la matriu de covariàncies, com que aquesta no és invariant per canvis d'escala, es recomana realitzar-ho sobre la matriu de correlacions en cas que les variables no provinguin de la mateixa naturalesa, és a dir, que no siguin dimensionalment homogènies o que l'ordre de la magnitud de les variables aleatòries mesurades no sigui el mateix.

2.1 Obtenció de les components principals

D'acord amb la definició donada de components principals, la manera de trobar-ne la primera d'elles és cercant la combinació lineal de les variables originals $\chi_1, \ldots, \chi_p$ que tingui màxima variància.

Per tal de facilitar els càlculs sense perdre informació sobre les dades, les estandarditzarem restant–li a cada una d'elles la seva mitjana. És a dir, en comptes de treballar amb la matriu de dades multivariants, $\mathbf{X}$, farem ús de la matriu de dades centrades $\bar{\mathbf{X}}$.

Els valors que tindran els n invidius en aquesta primera component vindran representats per un vector $p \times 1$: $Z_1 = \bar{\mathbf{X}} a_1$. On a_1 recollirà els coeficients pertinents per a aquesta transformació.

Com que hem centrat les dades, les variables ara tenen mitjana zero. Si diem $\bar{\mathbf{X}} = (y_{ij})$:

$$\begin{aligned}\bar{Y_j} &= \frac{1}{n}\sum_{i=1}^{n} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \frac{1}{n}\sum_{k=1}^{n} x_{kj}) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}[nx_{ij} - \sum_{k=1}^{n} x_{kj}) = \frac{1}{n^2}(n\sum_{i=1}^{n} x_{ij} - \sum_{i=1}^{n}\sum_{k=1}^{n} x_{kj}] \\ &= \frac{1}{n^2}(n\sum_{i=1}^{n} x_{ij} - n\sum_{k=1}^{n} x_{kj}) = 0.\end{aligned}$$

Veiem com això també s'ha traslladat a la nova variable:

$$\bar{Z_1} = \frac{1}{n}\sum_{i=1}^{n} y_i^t a_1 = \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p} y_{ij}a_{1j}) = \sum_{j=1}^{p} a_{1j}(\frac{1}{n}\sum_{i=1}^{n} y_{ij}) = 0.$$

La variància de Z_1 que volem que sigui màxima serà:

$$\text{var}(Z_1) = \tfrac{1}{n}Z_1^t Z_1 = \tfrac{1}{n}a_1^t \bar{\mathbf{X}}^{\mathbf{t}}\bar{\mathbf{X}}a_1 = a_1^t(\tfrac{1}{n}\bar{\mathbf{X}}^{\mathbf{t}}\bar{\mathbf{X}})a_1 = a_1^t\mathbf{S}a_1,$$

essent $\mathbf{S}$, tal com hem vist anteriorment, la matriu de covariàncies de les variables $\chi_1, \ldots, \chi_p$.

Per tal de donar solució a aquest sistema indeterminat (podríem maximitzar la variància de Z_1 augmentant el mòdul d'a_1 tant com volguéssim), imposem, sense perdre generalitat, la condició $a_1^t a_1 = 1$. Així, per maximitzar l'expressió considerada i contemplar alhora la condició esmentada farem servir els multiplicadors de Lagrange:

$$M = a_1^t\mathbf{S}a_1 - \lambda(a_1^t a_1 - 1).$$

Derivant respecte de les components d'a_1 de la forma habitual i igualant l'equació a zero obtenim:

$$\frac{\partial M}{\partial a_1} = 2\mathbf{S}a_1 - 2\lambda a_1 = 0 \text{ que té per solució: } \mathbf{S}a_1 = \lambda a_1.$$

És a dir que a_1 és vector propi de valor propi λ de la matriu de covariàncies $\mathbf{S}$. Així, recuperant la variància de Z_1 calculada tenim:

$$\text{var}(Z_1) = a_1^t\mathbf{S}a_1 = a_1^t\lambda a_1 = \lambda a_1^t a_1 = \lambda.$$

Que ens diu per tant que el vector propi de valor propi més gran de la matriu de covariàncies, defineix els coeficients de cada una de les variables en la primera component principal.

Amb això, hem volgut mostrar quin és el procés que porta a descobrir quines són les components principals. Si volguéssim trobar ara la segona component, Z_2, hauríem de procedir d'una manera semblant afegint-li la condició que aquesta fos incorrelacionada amb Z_1.

Generalitzarem aquest procés mitjançant les definicions i el teorema següents:

Definició 2.1.1. *Sigui* $\mathbf{X} = [X_1, \ldots, X_p]$ *una matriu de dades multivariants. Es defineixen les* ***components principals*** *com les variables compostes* $Z_1 = \bar{\mathbf{X}}a_1, \ldots, Z_p = \bar{\mathbf{X}}a_p$*, tals que:*

i) $var(Z_1)$ *és màxima condicionat a* $a_1^t a_1 = 1$*.*

ii) D'entre totes les variables compostes Z *tals que* $cov(Z_1, Z) = 0$*, la variable* Z_2 *és tal que* $var(Z_2)$ *és màxima condicionat a* $a_2^t a_2 = 1$*.*

iii) Si $p \geq 3$*, la component* Z_3 *és una variable incorrelacionada amb* Z_1, Z_2 *amb màxima variància. Anàlogament per la resta de components.*

Definició 2.1.2. *Si* $T = [a_1, \ldots, a_p]$ *és la matriu que té per columnes els vectors que defineixen les components principals, aleshores a la transformació lineal* $Z = \bar{\mathbf{X}}T$ *se l'anomena* ***transformació per components principals****.*

Observació 2.1.3. Com que $T = [a_1, \ldots, a_p]$ té per columnes els vectors propis normalitzats de la matriu de covariàncies i aquesta és simètrica, es té que els $a_1, \ldots, a_p$ són vectors ortonormals (en tant que els vectors propis d'una matriu simètrica són sempre ortogonals) i per tant la matriu T és una matriu ortogonal: $TT^t = T^tT = \mathbf{I}$. Això ens permet la tranformació lineal inversa, és a dir, podem expressar les variables originals normalitzades $\tilde{\chi_1}, \ldots, \tilde{\chi_p}$ en funció de les components principals $Z_1, \ldots, Z_p$: $\bar{\mathbf{X}} = ZT^t$.

Teorema 2.1.4. *Siguin* $a_1, \ldots, a_p$*,* $r \leq p$*, els vectors propis normalitzats de la matriu de covariàncies* $\mathbf{S}$ *de dimensió* $p \times p$*:* $\mathbf{S}a_i = \lambda_i a_i$ *amb* $a_i^t a_i = 1 \ \forall i = 1, \ldots, p$*. Aleshores:*

i) Les variables compostes $Z_i = \bar{\mathbf{X}}a_i$*, per* $i = 1, \ldots, p$*, són les components principals.*

ii) Se satisfà: $var(Z_i) = \lambda_i \ \forall i = 1, \ldots, p$*.*

iii) Les components principals són variables incorrelacionades: $cov(Z_i, Z_j) = 0$ *per* $1 \leq i, j \leq p$ *si* $i \neq j$*.*

Demostració. Siguin $\lambda_1 > \cdots > \lambda_p$ els valors propis de $\mathbf{S}$. Veiem que les variables $Z_i = \bar{\mathbf{X}}a_i$ són incorrelacionades i verifiquen que $\text{var}(Z_i) = \lambda_i \ \forall i = 1, \ldots, p$:

$$\text{var}(Z_j) = \text{cov}(Z_j, Z_j) = a_j^t \mathbf{S} a_j = a_j^t \lambda_j a_j = \lambda_j a_j^t a_j = \lambda_j$$

Per haver considerat $a_1, \ldots, a_p$ normalitzats.

$$\text{cov}(Z_i, Z_j) = a_i^t \mathbf{S} a_j = a_i^t \lambda_j a_j = \lambda_j a_i^t a_j$$
$$\text{cov}(Z_j, Z_i) = a_j^t \mathbf{S} a_i = a_j^t \lambda_i a_i = \lambda_i a_j^t a_i$$

Com que $a_i^t a_j = a_j^t a_i$, i $\mathbf{S}$ és una matriu simètrica, es té que $a_i^t \mathbf{S} a_j = a_j^t \mathbf{S} a_i$ i llavors:

$$\text{cov}(Z_i, Z_j) - \text{cov}(Z_j, Z_i) = 0 = (\lambda_j - \lambda_i) a_i^t a_j$$

És a dir que per $i \neq j$, $a_i^t a_j = 0$. Per tant: $\text{cov}(Z_i, Z_j) = 0 \ \forall i \neq j$, les variables són incorrelacionades.

Provem ara que aquestes variables $Z_1, \ldots, Z_p$ són efectivament les components principals:

Sigui $U = \sum_{i=1}^{p} c_i \chi_i$ una variable composta tal que $\sum_{i=1}^{p} c_i^2 = 1$. Com que $\bar{\mathbf{X}} = ZT^t$, si diem $t_1, \ldots, t_p$ a les columnes de la matriu T^t, tenim $\tilde{\chi_i} = Zt_i \ \forall i \in \{1, \ldots, p\}$ i per tant podem expressar aquesta nova variable com: $U = \sum_{i=1}^{p} c_i Z t_i = \sum_{i=1}^{p} \alpha_i Z_i$ on es compleix llavors que $\sum_{i=1}^{p} \alpha_i^2 = 1$ pel fet que $1 = c^t c = c^t T^t T c = \alpha^t \alpha$ essent $c = (c_1, \ldots, c_p)^t$ i $\alpha = (\alpha_1, \ldots, \alpha_p)^t$.

Aleshores:

$$\text{var}(U) = \text{var}(\sum_{i=1}^{p} \alpha_i Z_i) = \sum_{i=1}^{p} \alpha_i^2 \text{var}(Z_i) = \sum_{i=1}^{p} \alpha_i^2 \lambda_i \leq (\sum_{i=1}^{p} \alpha_i^2) \lambda_1 = \text{var}(Z_1),$$

que prova que Z_1 té màxima variància.

Considerem ara les variables U incorrelacionades amb Z_1. Podem expressar-les com:

$$Y = \sum_{i=1}^{p} d_i \chi_i = \sum_{i=2}^{p} \beta_i Z_i \text{ condicionat a } \sum_{i=2}^{p} \beta_i^2 = 1.$$

Aleshores:

$$\text{var}(U) = \text{var}(\sum_{i=2}^{p} \beta_i Z_i) = \sum_{i=2}^{p} \beta_i^2 \text{var}(Z_i) = \sum_{i=2}^{p} \beta_i^2 \lambda_i \leq (\sum_{i=2}^{p} \beta_i^2) \lambda_2 = \text{var}(Z_2),$$

i per tant Z_2 està incorrelacionada amb Z_1 i té màxima variància.

Si $p \geq 3$, la demostració de que $Z_3, \ldots, Z_p$ són també components principals és anàloga. □

2.2 Propietats de les components principals

Acabem de veure que la variància d'una certa component principal Z_i ve donada per $\text{var}(Z_i) = \lambda_i$, on λ_i és un dels valors propis de la matriu de covariàncies de les

variables originals, i per tant, la variació final recollida per aquestes noves variables vindrà donada per: $\sum_{i=1}^{p} \text{var}(Z_i) = \sum_{i=1}^{p} \lambda_i$.

Definició 2.2.1. *Es defineix la* ***variància generalitzada*** *d'un conjunt de k variables com el determinant de la matriu de covariàncies associada. Aquesta és una mesura global i escalar que pretén representar l'espai que ocupa el conjunt de dades a estudiar, com ara l'àrea pel cas $k = 2$ o el volum per $k = 3$.*

Lema 2.2.2. *Les components principals conserven tant la variabilitat inicial de les dades com la variància generalitzada.*

Demostració. La variabilitat total de les variables originals ve donada per:

$$\sum_{i=1}^{p} \text{var}(\chi_i) = \text{tr}(\mathbf{S})$$

i com que la suma dels valors propis d'una matriu és igual a la seva traça:

$$\sum_{i=1}^{p} \text{var}(\chi_i) = \text{tr}(\mathbf{S}) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \text{var}(Z_i).$$

Si ara recordem que el determinant d'una matriu equival al producte dels seus valors propis, tenim:

$$|\mathbf{S}| = \prod_{i=1}^{p} \lambda_i.$$

I d'altra banda, si diem $\mathbf{S_Z}$ a la matriu de covariàncies de les components principals, com que aquestes són per definició incorrelacionades, $\mathbf{S_Z}$ serà una matriu diagonal amb coeficients $\text{var}(Z_i)$:

$$|\mathbf{S_Z}| = \prod_{i=1}^{p} \text{var}(Z_i) = \prod_{i=1}^{p} \lambda_i.$$

Tenint per tant la igualtat tant en la variància total com en la generalitzada. □

D'acord al valor de la variabilitat total mostrat en l'observació anterior, es té que la proporció de variabilitat explicada per cada una de les components correspon al quocient entre la variància de la component considerada i la suma dels valors propis de la matriu $\mathbf{S}$. Així, donades les m primeres components principals es defineix el **percentatge de variabilitat explicada** per aquestes com:

$$P_m = 100 \frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p}.$$

Lema 2.2.3. *Les covariàncies entre cada component principal i les variables $\chi_1, \ldots, \chi_p$ normalitzades venen donades pel producte de les coordenades del vector propi que defineix la component, amb el valor propi associat:*

$$cov(Z_i; \tilde{\chi_1}, \ldots, \tilde{\chi_p}) = \lambda_i a_i = (\lambda_i a_{i1}, \ldots, \lambda_i a_{ip}) \text{ on } a_i \text{ és tal que } Z_i = \bar{\mathbf{X}} a_i.$$

Demostració. Prenem per a la demostració una component principal concreta, per exemple, $Z_1 = \bar{\mathbf{X}} a_1$. Com que podem expressar $\tilde{\chi_i} = \bar{\mathbf{X}} e_i$ essent e_i l'i-èssim vector columna de la base canònica, tenim que:

$$\text{cov}(Z_1, \tilde{\chi_i}) = \text{cov}(\bar{\mathbf{X}} a_1, \bar{\mathbf{X}} e_i) = a_1^t \mathbf{S} e_i \text{ tal com hem vist en la } \textit{proposició } 1.4.4.$$

Per ser $a_1 = (a_{11}, \ldots, a_{1p})$ vector propi de la matriu de covàriances, se satisfà $\mathbf{S} a_1 = \lambda_1 a_1$ que és equivalent per ser $\mathbf{S}$ simètrica a $a_1^t \mathbf{S} = \lambda_1 a_1^t$. Aleshores:

$$\begin{aligned} \text{cov}(Z_1; \tilde{\chi_1}, \ldots, \tilde{\chi_p}) &= (a_1^t \mathbf{S} e_1, \ldots, a_1^t \mathbf{S} e_p) = (\lambda_1 a_1^t e_1, \ldots, \lambda_1 a_1^t e_p) \\ &= \lambda_1 (a_{11}, \ldots, a_{1p}) = \lambda_1 a_1. \end{aligned}$$

□

Observació 2.2.4. Les correlacions entre una component principal i una varible χ_i normalitzada és proporcional al coeficient d'aquesta variable en la definició de la component, i el coeficient de proporcionalitat és el quocient entre la desviació típica de la component i la desviació típica de la variable:

$$\text{corr}(Z_i, \tilde{\chi_j}) = \frac{\text{cov}(Z_i, \tilde{\chi_j})}{\sqrt{\text{var}(Z_i)\text{var}(\tilde{\chi_j})}} = \frac{\lambda_i a_{ij}}{\sqrt{\lambda_i s_{jj}}} = a_{ij} \sqrt{\frac{\lambda_i}{s_{jj}}}.$$

Veiem ara com la tranformació per components principals és la que proporciona la representació, en un espai de dimensió reduïda, més òptima de la matriu de dades multivariants.

Suposem que volem representar les files $\mathrm{x}_1, \ldots, \mathrm{x}_n$ de $\mathbf{X}$ en un espai de dimensió $m \leq p$. Per fer-ho, necessitem introduir abans una distància. Com les variables considerades són contínues i l'espai en el qual les voldrem representar és $\mathbb{R}^m$, sembla coherent prendre la euclidiana.

Definició 2.2.5. *La **distància euclidiana** (al quadrat) entre les files de $\mathbf{X}$, $x_i^t = (x_{i1}, \ldots, x_{ip})$, $x_i^t = (x_{j1}, \ldots, x_{jp})$ és:*

$$\delta_{ij}^2 = (x_i - x_j)^t (x_i - x_j) = \sum_{h=1}^{p} (x_{ih} - x_{jh})^2.$$

La matriu $\Delta = (\delta_{ij})$ és la matriu $n \times n$ de distàncies entre les files.

Per tal de mesurar la qualitat de la representació de les dades es defineix el concepte següent:

Definició 2.2.6. *La **variabilitat geomètrica** de la matriu de distàncies Δ és el promig dels seus elements al quadrat:*

$$V_\delta(\mathbf{X}) = \frac{1}{2n^2}\sum_{i,j=1}^{n}\delta_{ij}^2.$$

Així, si $Y = \mathbf{X}T$ és una transformació lineal de $\mathbf{X}$ on T és una matriu $p \times m$,

$$\delta_{ij}^2(m) = (y_i - y_j) = \sum_{h=1}^{m}(y_{ih} - y_{jh})^2$$

és la distància euclidiana entre dues files d'Y. La varibilitat geomètrica en dimensió $m \leq p$ és:

$$V_\delta(Y)_m = \frac{1}{2n^2}\sum_{i,j=1}^{n}\delta_{ij}^2(m).$$

Teorema 2.2.7. *La variabilitat geomètrica de la distància euclidiana és la traça de la matriu de covariàncies*

$$V_\delta(\mathbf{X}) = tr(\mathbf{S}) = \sum_{h=1}^{p} var(\chi_h).$$

Demostració. Per veure-ho, reduïm-nos primer al cas univariant. Si $x_1, \ldots, x_n$ és una mostra univariant amb variància s^2, aleshores es té que:

$$\frac{1}{n^2}\sum_{i,j=1}^{n}(x_i - x_j)^2 = \frac{1}{n^2}\sum_{i,j=1}^{n}(x_i - \bar{x} - (x_j - \bar{x}))^2$$

$$= \frac{1}{n^2}\sum_{i,j=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n^2}\sum_{i,j=1}^{n}(x_j - \bar{x})^2 - \frac{2}{n^2}\sum_{i,j=1}^{n}(x_i - \bar{x})(x_j - \bar{x})$$

$$= \frac{1}{n}\sum_{j=1}^{n}(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2) + \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2) - \frac{2}{n^2}(\sum_{i=1}^{n}(x_i - \bar{x}))(\sum_{j=1}^{n}(x_j - \bar{x}))$$

$$= \frac{1}{n}ns^2 + \frac{1}{n}ns^2 + 0 = 2s^2.$$

És a dir que:

$$\frac{1}{2n^2}\sum_{i,j=1}^{n}(x_i - x_j)^2 = s^2$$

i per tant, aplicant això a cada una de les columnes d'$\mathbf{X}$ i sumant s'obté

$$V_\delta(\mathbf{X}) = \sum_{j=1}^{p} s_{jj} = \mathrm{tr}(\mathbf{S}).$$

□

El que marcarà doncs si una representació en dimensió reduïda és bona, serà la variabilitat geomètrica. Voldrem que aquesta sigui màxima per tal que els punts estiguin el més separats possible en la representació.

Teorema 2.2.8. *La transformació lineal T que maximitza la variabilitat geomètrica en dimensió m és la tranformació per components principals $Z = \bar{\mathbf{X}}T$, és a dir, $T = [t_1, \ldots, t_m]$ conté els m primers vectors propis normalitzats de $\mathbf{S}$.*

Demostració. La variabilitat geomètrica de $Y = \mathbf{X}V$ on $V = [\mathrm{v}_1, \ldots, \mathrm{v}_m]$ és una matriu $p \times m$ qualsevol, és

$$V_\delta(Y)_m = \sum_{j=1}^{m} \mathrm{var}(Y_j)$$

Com que s'assoleix la màxima variància quan Y_j és una component principal: $\mathrm{var}(Y_j) \leq \lambda_j$, i així:

$$màx V_\delta(Y)_m = \sum_{j=1}^{m} \lambda_j.$$

□

Observació 2.2.9. El percentatge de variabilitat geomètrica per $Z = \bar{\mathbf{X}}T$ és

$$P_m = 100\frac{V_\delta(Z)_m}{V_\delta(\mathbf{X})_p} = 100\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p}$$

que correspon exactament al percentatge de variabilitat explicada per les m primeres components principals.

2.3 Nombre de components principals

Una de les maneres més habituals per decidir amb quin nombre m de components principals ens quedem, és predefinir una quantitat de variabilitat que volem que expliquin aquestes i escollir l'm pel qual P_m és més pròxim a aquest valor. D'altra banda, una alternativa clara és escollir l'm pel qual s'estabilitza la variabilitat explicada per les components; si augmentar el valor m no aporta gaire més informació de les dades, aquest ja és un bon estimador.

Veiem un parell de criteris que ajuden a decidir quin nombre de components principals és l'òptim per les dades considerades.

Criteri de Kaiser. Aquest criteri parteix de la idea que en cas d'extreure les components principals de la matriu de correlacions, i per tant suposar que les variables observades tenen variància 1, no interessa quedar-se amb aquelles components que tinguin variància inferior a 1 ja que explicaran menys variabilitat que una de les variables originals. El criteri, per tant, consisteix en:

Retenir les m primeres components tals que $\lambda_m \geq 1$, on $\lambda_1, \ldots, \lambda_p$ són els valors propis d'$\mathbf{R}$, que també corresponen a les variàncies de les components.

La extensió d'aquest criteri a la matriu de covàriancies consisteix en considerar els $\lambda_m \geq v$, on $v = \text{tr}(\mathbf{S})/p$ és la mitjana de les variàncies.
Estudis posteriors al plantejament d'aquest criteri proven que és més correcte el punt de tall $\lambda^* = 0.7$ per la matriu de correlacions i $\lambda^* = 0.7 \times v$ per la de covariàncies.

Test de esfericitat. Aquest criteri suposa que la matiru de dades prové d'una població normal multivariant $N_p(\mu, \sigma)$, i es planteja la hipòtesi següent:

$$H_0^{(m)} : \lambda_1 > \cdots > \lambda_m > \lambda_{m+1} = \cdots = \lambda_p$$

si aquesta és certa, aleshores la distribució de les dades és esfèrica i per tant no s'haurà de considerar més de m components principals perquè no hi hauran direccions de màxima variabilitat a partir de m.
El test per decidir sobre $H_0^{(m)}$ està bassat en l'estadístic χ^2 i s'aplica seqüencialment:

a) Si acceptem $H_0^{(0)}$, és a dir $m = 0$, tots els valors propis són iguals i no hi han direccions principals. En cas de rebutjar-ho, repetim el test amb $H_0^{(1)}$.

b) Si acceptem $H_0^{(r)}$ amb $r < p$, aleshores $m = r$ i es té que:
$\lambda_1 > \cdots > \lambda_r > \lambda_{r+1} = \cdots = \lambda_p$.

c) En cas de rebutjar $H_0^{(r)}$ amb $r < p$, es repeteix el test per $H_0^{(r+1)}$.

2.4 Representació gràfica: Biplot

L'eina que farem servir per representar gràficament els resultats obtinguts pel PCA és el Biplot. Aquest és un mètode general i potent que permet visualitzar alhora tant les observacions com les variables de la matriu de dades multivariants.

El biplot consisteix a aproximar la matriu $\mathbf{X}$ mitjançant la descomposició en valors singulars vista en la definició 1.6.1, per una de rang dos:

$$\mathbf{X} \approx U_2 D_2^{1/2} V_2^t = (U_2 D_2^{1/2-c/2})(D_2^{c/2} V_2^t) = FC$$

on U_2 és $n \times 2$, $D_2^{1/2}$ és diagonal d'ordre 2, V_2^t és $2 \times p$ i totes tres estan definides segons la definició donada en la descomposició singular.

Prenent $0 \leq c \leq 1$ s'obtenen diferents descomposicions de la matriu $\mathbf{X}$ en dues matrius, i per tant diferents biplots, on: F representa les n files d'$\mathbf{X}$ en un espai de dos dimensions, i C representa en el mateix espai les columnes de la matriu. Els valors més utilitzats per aquest paràmetre són $c = 0$, 0.5 i 1. Amb cada un d'ells es prioritza la fidelització en la representació de o bé les variables, o bé les observacions.

A causa de com són les dades amb les quals treballarem, a nosaltres ens interessa estudiar-ne el cas $c = 1$. Representarem per tant les files d'$\mathbf{X}$ per la matriu U_2 i les columnes per $D_2^{1/2} V_2^t$. Per distingir ambdues representacions, les observacions es dibuixen com a punts i les variables com a vectors al pla. Veiem algunes de les propietats d'aquesta representació.

2.4.1 Propietats

Com que en la següent proposició s'esmentarà la distància de Mahalanobis introduim breument la definició d'aquesta:

Definició 2.4.1. *Donada una matriu multivariant de dades* $\mathbf{X}$ *amb matriu de covariàncies* $\mathbf{S}$*, es defineix la* ***distància de Mahalanobis*** *entre les files* x_i, x_j *d'*$\mathbf{X}$ *com:*

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^t \mathbf{S}^{-1} (x_i - x_j)}.$$

La distància de Mahalanobis, a diferència de l'euclidiana, té en compte la correlació de les variables i a més és invariant per canvis d'escala.

Proposició 2.4.2. *El biplot d'una matriu de dades multivariants prenent* $c = 1$ *satisfà:*

i) La representació de les observacions com a punts al pla mitjançant les files de V_2*, equival a projectar les observacions sobre el pla de les dues components principals estandarditzades perquè tinguin variància unitat.*

ii) Les distàncies euclidianes entre els punts del pla són equivalents, aproximadament, a les distàncies de Mahalanobis entre les observacions originals.

iii) La representació de les variables mitjançant vectors de dues coordenades és tal que l'angle entre aquests equival, aproximadament, a la correlació de les variables.

Demostració. Per demostrar aquestes propietats farem ús de la relació que hi ha entre les components principals i els vectors propis de $\bar{\mathbf{X}}\bar{\mathbf{X}}^{\mathbf{t}}$.

(i) Si $Z = \bar{\mathbf{X}}T$ és la transformació per components principals, es té que els vectors que formen les columnes de Z, $\mathrm{Z}_1, \ldots, \mathrm{Z}_p$, són vectors propis sense normalitzar de la matriu simètrica $\bar{\mathbf{X}}\bar{\mathbf{X}}^{\mathbf{t}}$ (i per tant, són vectors ortogonals). Veiem-ho:

Observem que els vectors propis normalitzats $a_1, \ldots, a_p$ de la matriu de covariàncies $\mathbf{S} = \frac{1}{n}\bar{\mathbf{X}}^{\mathbf{t}}\bar{\mathbf{X}}$ són també vectors propis de $\bar{\mathbf{X}}^{\mathbf{t}}\bar{\mathbf{X}}$ ja que satisfan:

$$\bar{\mathbf{X}}^{\mathbf{t}}\bar{\mathbf{X}}a_j = n\lambda_j a_j.$$

Multiplicant ara per $\bar{\mathbf{X}}$ tenim

$$\bar{\mathbf{X}}\bar{\mathbf{X}}^{\mathbf{t}}(\bar{\mathbf{X}}a_j) = n\lambda_j(\bar{\mathbf{X}}a_j)$$

i per tant $\mathrm{Z}_j = \bar{\mathbf{X}}a_j$ és un vector propi de la matriu $\bar{\mathbf{X}}\bar{\mathbf{X}}^{\mathbf{t}}$ que no està normalitzat.

Com que la norma d'aquest és:

$$\|\mathrm{Z}_j\|^2 = \mathrm{Z}_j^t\mathrm{Z}_j = (\bar{\mathbf{X}}a_j)^t\bar{\mathbf{X}}a_j = a_j^t(\bar{\mathbf{X}}^t\bar{\mathbf{X}}a_j) = a_j^t(n\lambda_j)a_j = n\lambda_j$$

el vector propi normalitzat serà:

$$v_j = \frac{1}{\sqrt{n\lambda_j}}\mathrm{Z}_j.$$

Generalitzant, la matriu de vectors propis de $\bar{\mathbf{X}}\bar{\mathbf{X}}^{\mathbf{t}}$ normalitzats serà:

$$V = [\mathrm{v}_1, \ldots, \mathrm{v}_p] = [\frac{1}{\sqrt{n\lambda_1}}\mathrm{Z}_1, \ldots, \frac{1}{\sqrt{n\lambda_p}}\mathrm{Z}_p] = Z\frac{1}{\sqrt{n}}D^{-1/2}$$

i amb aquesta normalització

$$V^tV = \frac{1}{\sqrt{n}}D^{-1/2}(\mathrm{Z}^t\mathrm{Z})D^{-1/2}\frac{1}{\sqrt{n}} = \frac{1}{n}D^{-1/2}(nD)D^{1/2} = \mathbf{I}.$$

Per tant, si representem els punts per V_2 tindrem les projeccions estandarditzades de les obsevacions sobre les dues primeres components.

(ii) Anem a comprovar ara la segona propietat. Una observació es representa per les components principals per x_iT, i si estandarditzem les components a variància unitat, es representa per $\mathrm{x}_i^tTD^{-1/2}$. Les distàncies euclidianes al quadrat entre dues observacions en termes de les seves coordenades expressades en les components principals seran:

$$\|\mathrm{x}_i^tTD^{-1/2} - \mathrm{x}_j^tTD^{-1/2}\|^2 = (\mathrm{x}_i - \mathrm{x}_j)^tTD^{-1}T^t(\mathrm{x}_i - \mathrm{x}_j)$$

i com $\mathbf{S} = TDT^t$ per ser aquesta la seva descomposició espectral, aleshores $\mathbf{S}^{-1} = TD^{-1}T^t$ i obtenim la distància de Mahalanobis entre les observacions originals. Com que al biplot no s'agafen les p components sinó només les dues més importants, la relació deixa de ser exacta i és aproximada.

(iii) Per últim, veiem que si representem les variables com vectors amb coordenades $D_2^{1/2}A_2^t = C$, els angles entre aquests representen aproximadament la correlació entre les variables. Per fer-ho, notem que al ser $a_1, \ldots, a_p$ vectors propis tant de $\mathbf{S}$ com de $\bar{\mathbf{X}}^{\mathbf{t}}\bar{\mathbf{X}}$ (tal com hem vist en *(i)*), les matrius T i A coincideixen per la seva definició i podem considerar:

$$S \simeq A_2D_2A_2^t = CC^t = \begin{bmatrix} c_1^t \\ \vdots \\ c_p^t \end{bmatrix} \begin{bmatrix} c_1 & \ldots & c_p \end{bmatrix}$$

on c_i és un vector 2×1 corresponent a la columna i-èssima de la matriu C. D'aquesta expressió s'obté que $c_i^tc_i \simeq s_i^2$ i $c_i^tc_j \simeq s_{ij}$ i finalment

$$r_{ij} = \frac{s_{ij}}{s_is_j} \simeq \frac{c_i^tc_j}{\|c_i\|\|c_j\|} = \cos(\hat{c_ic_j})$$

Per tant, aproximadament l'angle entre aquests vectors és el coeficient de correlació de les variables. □

Observació 2.4.3. La precisió de la representació del biplot depèn de la importància dels dos primers valors propis respecte al total. Com més proper a 1 sigui $(\lambda_1 + \lambda_2)/\mathrm{tr}(\mathbf{S})$, és a dir, més proper a 100 sigui P_2, més fiable serà la representació de les dades.

3 Anàlisi Discriminant Lineal

El problema de discriminació o classificació és un dels problemes més recurrents en l'anàlisi de dades. Iniciat al 1936 per Ronald Aylmer Fisher (1890 - 1962), aborda el problema següent: donat un conjunt d'elements que poden venir de dues o més poblacions diferents i dels quals s'ha observat una variable aleatòria p-dimensional, es vol classificar un nou element amb valors de les variables conegudes, en una de les poblacions.

L'anàlisi discriminant lineal, també conegut com l'anàlisi discriminant clàssic degut a Fisher (LDA, *Linear Discriminant Analysis*), és un dels mètodes de classificació supervisada (és a dir, que parteix d'una mostra d'elements amb una classificació coneguda per canviar la representació i entrenar el model de classificació de les observacions) que dóna solució a aquest problema. Per fer-ho, requereix que les variables considerades siguin contínues, independents i que n'hi hagin menys que el nombre total d'individus menys dos. A més, alhora de procedir assumeix normalitat multivariant, que tot i que no sempre es té, com les variables són contínues és possible transformar-les per tal que ho siguin.

El plantejament d'aquest mètode és divers en funció de la informació que es té de les dades i del nombre de classes o poblacions entre les quals s'ha de discriminar. Per entendre bé el procediment que hi ha darrera d'aquest, estudiarem primer el cas bàsic de dues poblacions amb distribucions normals multivariants conegudes, i en veure'm després la seva generalització.

3.1 Plantejament del problema

Siguin Ω_1, Ω_2 dues poblacions on estan definides les variables contínues $\chi_1, \ldots, \chi_p$ i on es coneixen: les funcions de densitat d'ambdues poblacions f_1, f_2, les probabilitats de que un individu a l'atzar provingui de cada una de les poblacions π_1, π_2 i els costos $c(i|j)$ associats a classificar en Ω_i un individu que pertany realment a Ω_j. Es vol estudiar el problema de classificar en una d'aquestes poblacions un nou individu ω amb vector d'observacions $\mathbf{x} = (x_1, \ldots, x_p)$.

Definició 3.1.1. *Una **regla discriminant** és un criteri que permet assignar ω a una de les poblacions considerades Ω_1, Ω_2 en funció de les observacions $(x_1, \ldots, x_p)$ conegudes. Aquest es sol plantejar mitjançant una funció discriminant $D(x_1, \ldots, x_p)$ amb la regla següent:*

1. *Si $D(x_1, \ldots, x_p) \geq 0 \Rightarrow$ assignem ω a Ω_1.*
2. *en cas contrari, assignem ω a Ω_2.*

Així, una regla discriminant equival a fer una partició de l'espai mostral $\mathbb{R}^p$ en dues regions: $R_1 = \{\mathbf{x}|D(\mathbf{x}) > 0\}$, $R_2 = \{\mathbf{x}|D(\mathbf{x}) < 0\}$. I per tant el problema de classificació pot plantejar-se com:

1. Si $\mathbf{x} \in R_1 \Rightarrow$ assignem ω a Ω_1.

2. Si $\mathbf{x} \in R_2 \Rightarrow$ assignem ω a Ω_2.

Observació 3.1.2. La probabilitat de que l'individu amb observació $\mathbf{x}$ pertanyi a cada una de les poblacions vindrà donada per

$$P(\Omega_j|\mathbf{x}) = \frac{P(\mathbf{x}|\Omega_j)\pi_j}{P(x|\Omega_1)\pi_1 + P(x|\Omega_2)\pi_2}, \text{ que és equivalent a}$$

$$P(\Omega_j|\mathbf{x}) = \frac{f_j(\mathbf{x})\pi_j}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})\pi_2}.$$

Si diem d_j a la decisió de classificar ω en Ω_j, el valor esperat de cada una d'aquestes serà:

$$E(d_1) = 0P(\Omega_1|\mathbf{x}) + c(1|2)P(\Omega_2|\mathbf{x}) = c(1|2)P(\Omega_2|\mathbf{x}).$$

$$E(d_2) = c(2|1)P(\Omega_1|\mathbf{x}) + 0P(\Omega_2|\mathbf{x}) = c(2|1)P(\Omega_1|\mathbf{x}).$$

I per tant assignarem ω a aquella població que tingui associat un valor esperat menor. Si substituïm les expressions anteriors, la regla de decisió correspondrà finalment a:

1. Si $c(1|2)f_2(\mathbf{x})\pi_2 < c(2|1)f_1(\mathbf{x})\pi_1 \Rightarrow$ assignem ω a Ω_1.
2. Si $c(1|2)f_2(\mathbf{x})\pi_2 > c(2|1)f_1(\mathbf{x})\pi_1 \Rightarrow$ assignem ω a Ω_2.

3.2 Funció lineal discriminant

Suposem ara que les distribucions f_1, f_2 són normals multivariants amb diferents vectors de mitjanes, μ_1, μ_2 respectivament, però mateixa matriu de covariàncies, Σ. Aleshores, les funcions de densitat seran:

$$f_j(\mathbf{x}) = \frac{|\Sigma|^{-1/2}}{(\sqrt{2\pi})^p} e^{-\frac{1}{2}(x-\mu_j)^t\Sigma^{-1}(x-\mu_j)},$$

i si les substituïm en la regla de decisió d'abans, $\frac{f_2(\mathbf{x})\pi_2}{c(2|1)} < \frac{f_1(\mathbf{x})\pi_1}{c(1|2)}$, aplicant logaritmes (en tant que ambdós termes són positius) i simplificant les constants que tenen en comú per ser Σ la mateixa, obtenim:

$$\log(\tfrac{\pi_2}{c(2|1)}) - \tfrac{1}{2}(\mathbf{x}-\mu_2)^t\Sigma^{-1}(\mathbf{x}-\mu_2) < \log(\tfrac{\pi_1}{c(1|2)}) - \tfrac{1}{2}(\mathbf{x}-\mu_1)^t\Sigma^{-1}(\mathbf{x}-\mu_1)$$

és a dir, operant

$$\tfrac{1}{2}(\mathbf{x}-\mu_1)^t\Sigma^{-1}(\mathbf{x}-\mu_1) < \tfrac{1}{2}(\mathbf{x}-\mu_2)^t\Sigma^{-1}(\mathbf{x}-\mu_2) - \log\left(\tfrac{c(1|2)\pi_2}{c(2|1)\pi_1}\right).$$

Si diem ara D_j a les distàncies de Mahalanobis (definició 2.4.1) entre $\mathbf{x}$ i la mitjana de la població Ω_j, $D_j = \sqrt{(\mathbf{x}-\mu_j)^t\Sigma^{-1}(\mathbf{x}-\mu_j)}$, la regla resultant de classificació és:

1. Classificar ω en Ω_1 si: $\frac{1}{2}D_1^2 < \frac{1}{2}D_2^2 - \log\left(\frac{c(1|2)\pi_2}{c(2|1)\pi_1}\right)$.

2. Classificar ω en Ω_2 si: $\frac{1}{2}D_1^2 > \frac{1}{2}D_2^2 - \log\left(\frac{c(1|2)\pi_2}{c(2|1)\pi_1}\right)$.

Observació 3.2.1. Si suposem iguals tants els costos $c(i|j)$ com les probabilitats π_i, la regla resultant consisteix en classificar l'observació en la població que tingui la mitjana més propera usant la distància de Mahalanobis. En cas que les variables tinguéssin matriu de covàriances $\Sigma = \mathbf{I}\sigma^2$, la regla seria equivalent a utilitzar la distància euclidiana.

3.3 Interpretació de la regla de classificació

La regla anterior pot escriure's d'una manera equivalent que permet interpretar geomètricament el mètode de classificació utilitzat. Si simplifiquem els termes comuns en ambdues distàncies

$$D_i^2 = (\mathbf{x} - \mu_i)^t \Sigma^t - 1(\mathbf{x} - \mu_i) = \mathbf{x}^t\Sigma^{-1}\mathbf{x} - 2\mu_i^t\Sigma^{-1}\mathbf{x} + \mu_i^t\Sigma^{-1}\mu_i$$

la regla obtinguda divideix el conjunt de possibles valors d'$\mathbf{x}$ en dues regions que tenen per frontera:

$$-\mu_1^t\Sigma^{-1}\mathbf{x} + \tfrac{1}{2}\mu_1^t\Sigma^{-1}\mu_1 = -\mu_2^t\Sigma^{-1}\mathbf{x} + \tfrac{1}{2}\mu_2^t\Sigma^{-1}\mu_2 - \log\left(\frac{c(1|2)\pi_2}{c(2|1)\pi_1}\right)$$

i que com a funció lineal en $\mathbf{x}$ és equivalent a:

$$(\mu_2 - \mu_1)^t\Sigma^{-1}\mathbf{x} = (\mu_2 - \mu_1)^t\Sigma^{-1}\left(\tfrac{\mu_2+\mu_1}{2}\right) - \log\left(\frac{c(1|2)\pi_2}{c(2|1)\pi_1}\right).$$

Si diem $\mathbf{w} = \Sigma^{-1}(\mu_2 - \mu_1)$, aleshores, la frontera entre les regions de classificació per Ω_1, Ω_2 pot escriure's com:

$$\mathbf{w}^t\mathbf{x} = \mathbf{w}^t\left(\tfrac{\mu_2+\mu_1}{2}\right) - \log\left(\frac{c(1|2)\pi_2}{c(2|1)\pi_1}\right)$$

que és l'equació d'un hiperplà.

Observació 3.3.1. Pel cas particular en el qual $c(1|2)\pi_2 = c(2|1)\pi_1$, la regla de decisió serà estudiar la inequació

$$\mathbf{w}^t\mathbf{x} > \mathbf{w}^t\left(\tfrac{\mu_2+\mu_1}{2}\right), \text{ o, } \mathbf{w}^t\mathbf{x} - \mathbf{x}^t\mu_1 > \mathbf{w}^t\mu_2 - \mathbf{w}^t\mathbf{x}$$

que indica que el procediment per classificar $\mathbf{x}$ pot resumir-se com:

i) Calcular el vector $\mathbf{w} = \Sigma^{-1}(\mu_2 - \mu_1)$

ii) Construir la funció (o variable indicadora) discriminant: $g(z) = \mathbf{w}^t z$

iii) Evaluar la funció discriminant amb l'observació $\mathbf{x}$, $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x}$, i les mitjanes de les poblacions, $m_i = g(\mu_i) = \mathbf{w}^t\mu_i$.

iv) Classificar $\mathbf{x}$ en aquella població on la distància $|g(\mathbf{x}) - m_i|$ sigui mínima.

És a dir que, en cas que s'estandaritzi la variable discriminant $g(z)$, la regla equival a projectar el punt $\mathbf{x}$ que es vol classificar i les mitjanes d'ambues poblacions sobre una recta, per després assignar a $\mathbf{x}$ la població que tingui la mitjana més propera en la projecció.

El cas general en el qual no es té la igualtat esmentada, es pot veure que la interpretació és la mateixa.

3.4 Generalització del problema. Regla estimada per la classificació

Considerem ara G poblacions entre les quals volem classificar un nou individu ω amb observació $\mathbf{x}$. La matriu de dades multivariants $\mathbf{X}$ de dimensió $n \times p$ podem particionar-la en G matrius corresponents a les subpoblacions. Així, denotarem per x_{ijg} als elements d'aquestes submatrius on i representa l'individu, j la variable i g el grup o submatriu al qual pertany.

El nombre d'individus pertanyents a cada grup l'indicarem per n_g, i $x_{ig}^t = (x_{i1g}, \ldots, x_{ipg})$ correspondrà al vector fila que conté els p valors de les variables per l'individu i en el grup g.

El vector de mitjanes dins de cada classe o subpoblació serà un vector columna amb les p mitjanes de les observacions de la classe g, és a dir:

$$\bar{x}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} x_{ig}$$

Per tal de tenir estimacions centrades de les variàncies i les covariàncies dels elements de cada classe, es defineix la matriu de covariàncies de la classe g com:

$$\hat{\mathbf{S}}_g = \frac{1}{n_g - 1}\sum_{i=1}^{n_g}(x_{ig} - \bar{x}_g)(x_{ig} - \bar{x}_g)^t.$$

I es pren com a millor estimació centrada de la matriu de covariàncies amb totes les dades, la combinació lineal de les estimacions centrades de cada població amb un pes proporcional a la seva precisió, és a dir:

$$\hat{\mathbf{S}}_\omega = \sum_{g=1}^{G}\frac{n_g - 1}{n - G}\hat{\mathbf{S}}_g.$$

Per obtenir les funcions discriminants s'utilitzarà llavors $\bar{x}_g$ com a estimació de μ_g i $\hat{\mathbf{S}}_\omega$ com a estimació de Σ. En particular, suposant iguals les probabilitats a

priori i els costos de classificació, el nou element es classificarà en el grup que tingui un valor mínim de la distància de Mahalanobis entre l'observació $\mathbf{x}$ i la mitjana de cada grup.

3.5 Sistemes d'avaluació

Per tal d'avaluar la eficàcia en la classificació i predicció de l'Anàlisi Discriminant Lineal, farem servir el mètode de la validació creuada en k iteracions i les mesures habituals que s'empren en la classificació.

El mètode de la **validació creuada de K iteracions** (o *K-fold cross-validation*) té com a objectiu garantitzar que els resultats obtinguts en l'anàlisi estadístic siguin independents de la partició del conjunt de dades realitzada, en els subconjunts d'entrenament i de prova. Per fer-ho, es divideixen les dades en K subconjunts i es realitzen K iteracions considerant, per cada una d'aquestes i de manera seqüencial, només un dels subconjunts com a *test* i la resta com a *train*.
Quan es pren $K =$ nombre de dades total, considerant llavors una única observació com a *test* per cada una de les iteracions, aquest mètode rep el nom de ***leave-one-out cross-validation*** *(LOOCV).*

Siguin Ω_1, Ω_2 les dues poblacions en les quals es vol dur a terme la classificació. Si considerem Ω_1 la classe positiva i Ω_2 la negativa, en diem:

- $tp =$ *true positive*: a classificar un individu en la classe positiva correctament,
- $tn =$ *true negative*: a classificar l'individu correctament en la classe negativa,
- $fp =$ *false positive*: a classificar erròniament l'individu en la classe positiva i,
- $fn =$ *false negative*: a classificar erròniament l'individu en la classe negativa.

Aleshores, les mesures estàndards que es prenen en una classificació binària són les següents:

$$\textbf{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn},$$

$$\textbf{Precision} = \frac{tp}{tp + fp},$$

$$\textbf{Recall} = \frac{tp}{tp + fn},$$

$$\textbf{F1} = \frac{2tp}{2tp + fp + fn}.$$

Com *accuracy* és tan sols la proporció d'encerts respecte del nombre total d'individus a classificar i aquesta pot no ser una bona mesura si les dades d'entrenament estan descompensades (ja que un classificador que adjudiqués a tots els individus el grup de major nombre d'integrants, obtindria un bon resultat si la diferència de mida d'ambdós grups és prou significativa); es consideren també les mesures *precision*, *recall* i *F1*. *Precision* i *recall* contesten respectivament a les preguntes de

quants dels que s'han classificat en la classe positiva són realment d'aquesta classe i quants dels que pertanyen a la classe positiva s'han classificat correctament. *F1* és simplement la mitjana harmònica d'aquestes dues mesures

$$F1 = 2\,\frac{precision \times recall}{precision + recall}.$$

i serveix per tenir una estimació més general dels resultats obtinguts. En funció de com es defineixin les classes positiva i negativa i dels interessos que tingui l'estudi, es dóna prioritat a una mesura o una altra.

Per tal d'acabar de validar els resultats que s'obtinguin en la classificació i poder concloure si aquests són estadísticament significatius (és a dir, que no són deguts a l'atzar), farem servir el **test de les permutacions**. Aquest consisteix a realitzar M iteracions del mètode que es vol avaluar modificant, en cada una d'elles i de manera aleatòria, alguns dels marcadors de les dades. Així, comparant quants dels resultats que s'obtenen amb aquestes mostres artificals són iguals o superiors als obtinguts amb les dades originals, es construeix el *p-valor* i es decideix en funció del valor de significació α triat, si existeixen evidències estadístiques que hi ha diferència o no.

4 Experiments i resultats

Un cop introduït els conceptes necessaris per a la realització de l'anàlisi de dades, procedim a l'aplicació dels mètodes estudiats i al posterior estudi dels resultats obtinguts.

4.1 Obtenció i tractament de les dades

Com ja hem comentat anteriorment, les dades amb les quals treballarem provenen de la Universitat de Barcelona. Aquestes han estat sotmeses prèviament tant a un procés d'anonimització com a un de neteja.

El procés d'anonimització de les dades ha vingut fet pel departament de planificació i gestió acadèmica. Tot i que no han seguit el protocol establert per la Universitat de Barcelona, ja que les dades no seran publicades i han estat utilitzades només per un ús intern, sí que han generat un identificador de manera aleatòria per cadascun dels alumnes que no guarda relació amb cap identificador personal com ara el DNI o el NIUB. Així, l'única manera de desanonimitzar un alumne seria conèixer dades personals d'ell i creuar-les amb la base de dades. Com que no és el cas ni es pretén fer-ho, aquest procés d'anonimització és tot el que necessitem.

El procés de neteja de les dades, en el qual figuren procesos com ara la unificació dels Graus o l'eliminació d'aquelles dades atípiques que distorsionaven la mostra, el van realitzar l'equip del projecte d'innovació docent *Sistema intel·ligent de suport per al tutor d'estudis* [5].
Tot i que aquest procés de neteja ha estat impecable i per aquest motiu l'hem volgut aprofitar, encara ens hem trobat amb algunes anomalies en la base de dades degudes exclusivament a la procedència d'aquestes. Fem referència a dos casos: assignatures convalidades o reconegudes on estranyament no s'ha adjudicat cap nota i per tant consten amb 0.0; i alumnes que per provenir de Llicenciatura o d'una altra facultat no s'han omplert totes les seves dades i directament no tenim resultats seus per a certes assignatures. La manera amb la qual hem tractat els casos del primer tipus, ha estat considerar en el seu lloc la mitjana de les notes de l'assignatura en qüestió. Pels del segon tipus, no hem pogut fer cap altra cosa que ignorar-los.

Les dades a les quals aplicarem els mètodes estudiats en aquest treball són les següents:

a) Dos fitxers *qualifications_maths_info.csv* i *qualifications_law.csv* que contenen respectivament les qualificacions dels alumnes dels Graus de Matemàtiques i d'Enginyeria Informàtica, i del Grau de Dret, per cada una de les assignatures cursades entre el 2009 i mitjans del 2014. Més concretament, aquests fitxers consten d'una columna per cada un d'aquests ítems:

   - identificador de l'alumne
   - identificador de l'assignatura

- any de matriculació de l'assignatura per part de l'alumne
- identificador del grau al qual pertany l'assignatura
- nota final obtinguda de l'alumne en l'assignatura

En total en aquests fitxers tenim 516 alumnes del Grau de Matemàtiques, 455 del d'Enginyeria Informàtica i 3463 de Dret.

b) Dos fitxers *assigs_maths_info.csv* i *assigs_law.csv* que contenen respectivament una descripció general de totes les assignatures dels Graus de Matemàtiques, Informàtica i Dret. Com també els fitxers anteriors, aquests recullen les característiques de la informació a representar mitjançant columnes amb les etiquetes següents:

- identificador de l'assignatura
- nom que rep l'assignatura
- curs en el qual s'imparteix l'assignatura
- semestre associat a l'assignatura
- identificador del grau al qual pertany l'assignatura
- nombre de crèdits associats a l'assignatura

Així, dels primers fitxers extreurem les corresponents matrius de dades multivariants $\mathbf{X}$ on les variables, situades com a columnes, seran les assignatures, els individus a estudiar, en les files, seran els alumnes i les observacions seran les notes obtingudes per cada alumne en l'assignatura pertinent. Amb l'objectiu de tenir la màxima variabilitat possible en les dades, hem decidit considerar com a observació x_{ij} la mitjana de les notes finals obtingudes per l'alumne i en l'assignatura j.
Els dos últims fitxers, juntament amb el Pla d'Estudis penjat en la web de la facultat, ens serviran per tal d'identificar les assignatures ja sigui per curs o ensenyament.

Com el nostre objectiu és estudiar i preveure l'abandonament, els grups d'alumnes que considerarem seran el dels alumnes que han acabat la carrera, el dels que l'han abandonada i la unió d'ambdós grups. El dels alumnes que han abandonat la carrera vindran donats seguint la normativa oficial de la UB, en el qual un alumne es considera que ha abandonat la carrera si fa dos anys o més que no matricula cap assignatura. Per identificar als alumnes del primer grup, com que les dades corresponen a un període curt de temps, vam observar que si agafàvem només com a filtre que tinguéssin nota en l'assignatura Treball Final de Grau (sigui aprovada o suspesa) ens reduíem a treballar amb només 29 alumnes pel Grau de Matemàtiques, 61 pel d'Enginyeria Informàtica i 903 per Dret. Per aquest motiu i basant-nos en la taxa d'abandonament per curs (que queda reflectida en la taula 1), vam considerar prendre també com a indicador que un alumne hagués finalitzat, el fet que tingués nota en 30 o més assignatures (descartant pertinentment els casos que es trobéssin ja en el segon grup), és a dir, que hagués cursat l'equivalent a tres anys de carrera en els Graus de Matemàtiques o Informàtica.

Pel que fa a les variables, ens hem reduït a considerar només les 10 assignatures que formen el primer any de carrera. D'aquesta manera, minimitzem l'exigència respecte a les dades tant a l'hora de filtrar-les, evitant així el màxim possible els forats de la base de dades, com a l'hora d'observar i avaluar un nou individu.

	Matemàtiques		**Informàtica**		**Dret**	
	1er curs	*2on curs*	*1er curs*	*2on curs*	*1er curs*	*2on curs*
2009	42.17	-	17.91	-	17.20	11.76
2010	42.17	19.64	36.76	11.67	16.59	9.30
2011	38.30	16.00	31.48	5.66	15.69	7.76
2012	30.69	11.67	25.93	15.25	11.01	6.73
Total	37.95	15.66	27.98	10.86	15.20	7.88

Taula 1: % d'abandonament.

Nota 1. *El càlcul del % d'aquesta taula s'ha realitzat sobre el nombre total de matriculats en cada un dels cursos. S'ha considerat que un alumne cursava segon quan es van prendre les dades si tenia alguna assignatura de segon matriculada aquell any.*

4.2 Anàlisi dels resultats

Abans de comentar els resultats obtinguts, notem que en el nostre cas les poblacions a considerar per l'Anàlisi Discriminant Lineal (vist en el capítol 3) seran: $\Omega_1 = \{$els alumnes que han acabat$\}$ i $\Omega_2 = \{$els alumnes que han abandonat$\}$, on per l'avaluació considerarem Ω_2 com a classe positiva i Ω_1 com a classe negativa. Els costos de l'error en la classificació i les probabilitats de pertinença a cada un dels grups es consideraran iguals. Així, el mètode ens vindrà definit per:

$$L(x) = [x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)]^t \hat{\mathbf{S}}_\omega^{-1}(\bar{x}_1 - \bar{x}_2)$$

1. Si $L(x) > 0$, classifiquem l'observació x en Ω_1.

2. En cas contrari, classifiquem l'observació x en Ω_2.

Pel que fa a l'Anàlisi de Components Principals (vist en el capítol 2), la seva aplicació a les nostres dades no requereix cap simplificació, així que s'ha realitzat el mètode tal com s'ha explicat.
Finalment, respecte a la visualització, aclarir que per tal de facilitar la interpretació dels gràfics s'ha optat per: acolorir en tots aquells gràfics on es dibuixen els identificadors dels alumnes, de color vermell els que pertanyen als que han acabat i de color blau als que han abandonat; i representar la classificació 1-dimensional obtinguda amb el mètode LDA en un gràfic de dues dimensions afegint com a eix d'ordenades l'ordre en el qual s'han classificat.

En els apartats següents, procedim en aplicar els mètodes comentats en aquest treball (dels quals podeu consultar la implementació en la documentació complementària) per cadascun dels ensenyaments considerats i al posterior anàlisi dels resultats. Com que el que volem és interpretar i caracteritzar les dades, aplicarem primer per separat el mètode PCA a cada un dels grups d'alumnes considerats (alumnes que han acabat, alumnes que han abandonat i unió d'ambdós grups), i després aplicarem el LDA al conjunt total d'alumnes.

4.2.1 Grau de Matemàtiques

El nombre total d'alumnes considerats per l'execució d'ambós mètodes és 181, d'on 53 són alumnes que hem considerat pertanyents al grup dels que havien finalitzat la carrera i 128 són dels que l'han abandonada. No ens ha d'estranyar la gran diferència entre el nombre d'integrants de cada grup si tenim en compte l'alt percentatge d'abandonament que té el Grau de Matemàtiques i el fet que l'interval de temps al qual pertanyen les dades comprèn només dues promocions senceres d'alumnes: la del $2009-2013$ i la del $2010-2014$.

Les assignatures i l'ordre en el qual han estat considerades dins la matriu de dades, són les següents:

Nom de l'assignatura	**Àlies utilitzat**
Matrius i Vectors	*MaVe*
Algebra Lineal	*AlLi*
Introducció al Càlcul Diferencial	*IaCD*
Introducció al Càlcul Integral	*IaCI*
Llenguatge i Raonament Matemàtic	*LiRM*
Aritmètica	*Arit*
Elements de Programació	*ElPrg*
Programació Científica	*PrgCien*
Anàlisi de Dades i Introducció a la Probabilitat	*ADiP*
Física	*Fisc*

Taula 2: Assignatures considerades pel Grau de Matemàtiques.

PCA:

Alumnes que han acabat:

Els dos vectors propis normalitzats associats als dos valors propis més grans de la matriu de covàriancies, $\lambda_1 = 14.317$ i $\lambda_2 = 2.029$, són:

$a_1 = (+0.286, +0.393, +0.260, +0.276, +0.286, +0.391, +0.295, +0.365, +0.242, +0.326)$

$a_2 = (+0.082, -0.141, -0.311, -0.014, -0.089, +0.276, -0.245, -0.023, -0.479, +0.709)$

Observem, que la primera component principal $Z_1 = \bar{\mathbf{X}} a_1$ queda formada com una mena de mitjana ponderada (ja que els coeficients que formen a_1 no són exactament els mateixos) dels resultats obtinguts per cada un dels alumnes en totes les variables. Aquest fet que pot semblar puntual, veurem més endavant que es dóna

en cada una de les execucions del mètode. I es que si ho pensem detingudament, té molt de sentit que sigui mitjançant una mitjana ponderada la manera de captar més variabilitat possible en una sola variable.

Respecte al vector que forma la segona component principal podem observar que l'assignatura amb un valor clarament diferent de la resta és la de Física (l'última variable), fet que queda reflectit en el gràfic biplot de la figura 1 on totes les assignatures apunten més o menys en la mateixa direcció menys aquesta. Això pot ser degut al fet que Física és l'assignatura que guarda menys relació explícita amb la resta de matèries de la carrera. L'única amb la qual té una connexió directa és amb Modelització i aquesta és de tercer curs.

La representació dels alumnes sobre les dues components principals ens dóna un núvol força dispers que no tendeix cap a cap assignatura en concret. Això ens diu que els alumnes considerats, els que han acabat la carrera en aquest cas, no estan ben caracteritzats pels resultats que van obtenir en les assignatures de primer, per això la variabilitat explicada per les dues primeres components principals es redueix a un 71.44% i el criteri de Kaiser ens aconsella treballar amb com a mínim tres components.

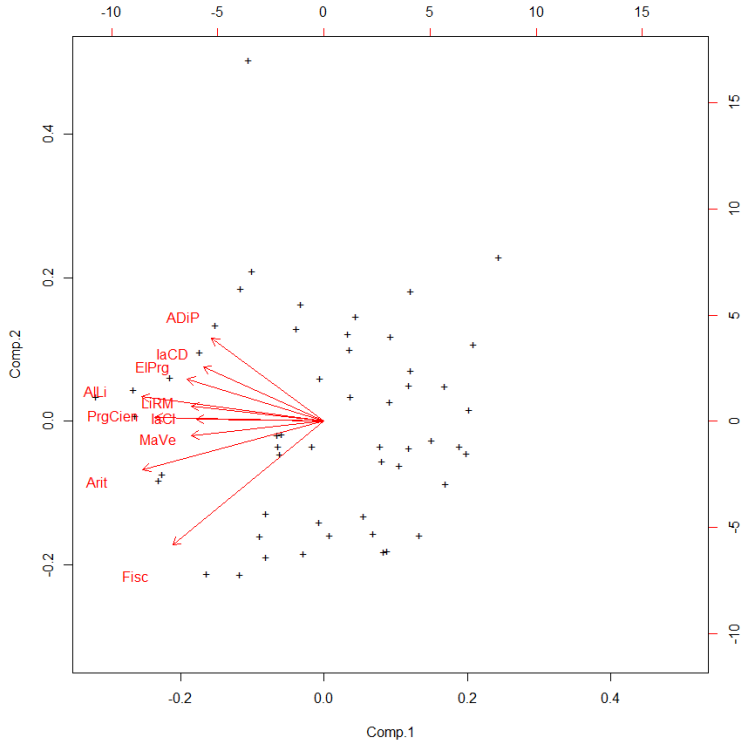

Figura 1: Biplot per Matemàtiques dels alumnes que han acabat.

Alumnes que han abandonat:

Els vectors propis que definiran les dues components principals en aquest cas són els següents:

$a_1 = (-0.286, -0.217, -0.352, -0.389, -0.369, -0.221, -0.349, -0.261, -0.369, -0.289)$

$a_2 = (-0.108, -0.306, +0.254, -0.323, +0.281, -0.362, +0.363, -0.095, +0.389, -0.472)$

Novament, els coeficients del que formarà la primera component principal són més o menys semblants per tal que el resultat de multiplicar-lo per la matriu de dades centrades, sigui una mitja ponderada de les notes obtingudes per cada alumne en les 10 assignatures de primer.
Si ens fixem en els de la segona component i els contrastem amb el biplot obtingut (la figura 2), veiem que aquí el signe és un factor determinant que separa les assignatures en dos grans grups: les que van presentar menys dificultats pels alumnes considerats, com *ADiP*, *LiRM*, *ElPrg* i *IaCD* amb mitjanes (per aquests alumnes) que van del 2.2 al 3.9, i *PrgCien*, *MaVe*, *IaCI*, *AlLi*, *Arit* i *Fisc* amb mitjanes que oscil·len del 0.8 al 2.2, que van presentar més dificultat.

Pel que fa a la representació dels alumnes en funció de les dues primeres components principals, s'observa que tot i la diversitat un gran nombre d'alumnes es concentra al voltant de l'origen de coordenades, tenint així notes molt baixes en totes les assignatures. Aquells alumnes que es troben en la part superior del gràfic corresponen als que només van aconseguir bons resultats en les assignatures del primer grup, i els de la part inferior esquerra, en canvi, van obtenir notes altes en totes les assignatures.
La variabilitat explicada per les dues components principals, i en conseqüència la precisió de la representació del gràfic biplot, és tan sols d'un 70.76%, fet que ens informa que el comportament dels alumnes que abandonen el Grau de Matemàtiques és força similar i per tant, en cas que aquest sigui diferent al dels que finalitzen, es podrien identificar bé.

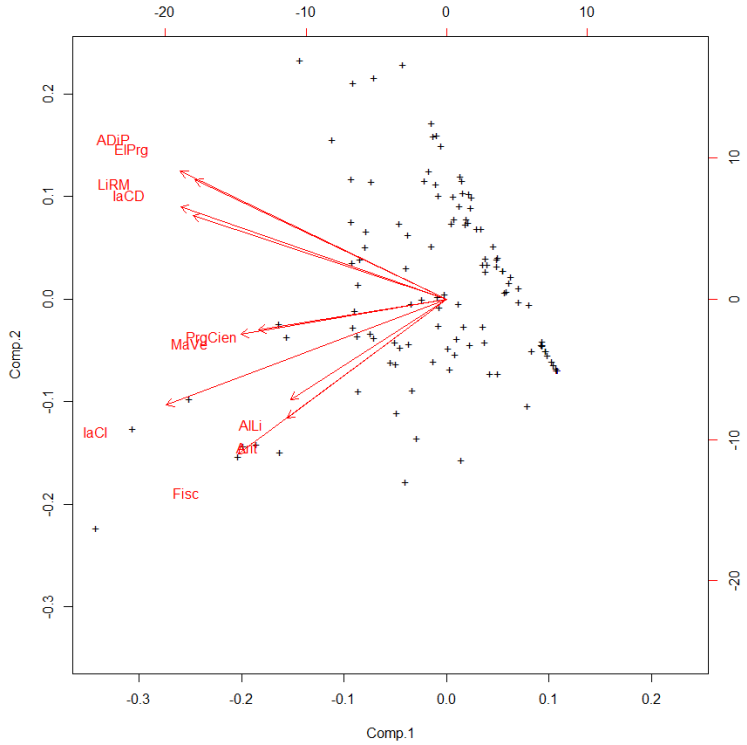

Figura 2: Biplot per Matemàtiques dels alumnes que han abandonat.

Unió dels alumnes que han acabat i dels que han abandonat:

Veiem finalment quina seria la representació i la interpretació de les components principals si consideréssim la unió dels alumnes que han abandonat i dels que han acabat.

$a_1 = (-0.306, -0.341, -0.280, -0.371, -0.259, -0.337, -0.304, -0.335, -0.271, -0.339)$

$a_2 = (-0.070, -0.316, +0.359, -0.056, +0.413, -0.358, +0.301, -0.140, +0.490, -0.337)$

Els coeficients del vector que definirà la segona component principal tornen a separar, mitjançant el signe, les assignatures en els dos grups vistos abans tot i que ara la separació entre aquests no és tan pronunciada (tal com podem veure al gràfic 3), degut al fet que pels alumnes que han acabat aquesta diferència és gairebé imperceptible.

Si bé en el biplot dibuixat no s'arriba a percebre quant de discriminant poden arribar a ser les dues components principals, es fa palès si observem la variabilitat que aquestes recullen, ja que com es podia esperar a l'haver considerat dos grups d'alumnes tan diferents entre si, ha augmentat fins a un $84,87\%$. Per tal de visualitzar aquest fet, hem representat novament els alumnes en funció de les dues primeres components acolorint aquest cop els alumnes que han acabat de color vermell i els que han abandonat de color blau, obtenint com a resultat el gràfic de la figura 4.

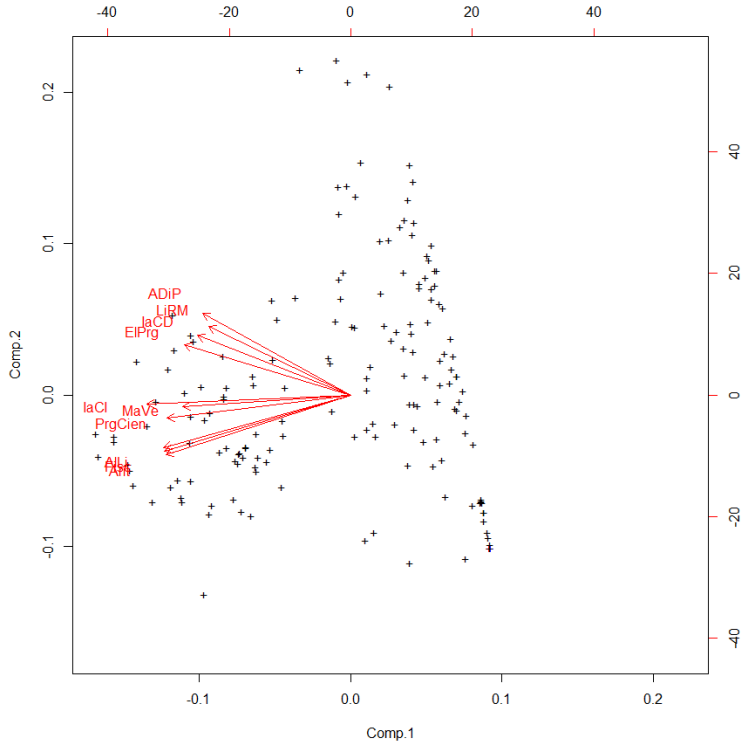

Figura 3: Biplot per Matemàtiques dels alumnes totals considerats.

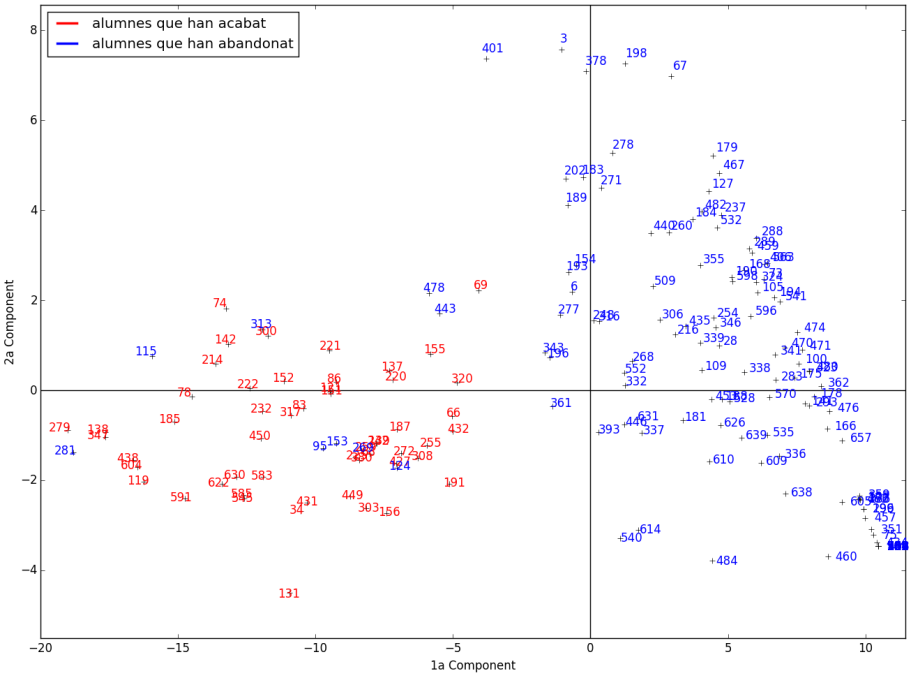

Figura 4: Representació dels alumnes de Matemàtiques.

En aquest, podem observar una primera aproximació a un sistema de classificació, en tant que la primera component principal ha aconseguit discriminar (considerant el llindar al 0) amb només 22 errors el conjunt total d'alumnes considerats (i per tant, una *accuracy* del 87.84%). Tanmateix el marge de millora és alt, ja que tot i tenir un 100% de *precision* (cap alumne que ha acabat s'ha classificat dins dels que han abandonat), el *recall* és d'un 82.81%. Veurem a continuació com el mètode de l'Anàlisi Discriminant Lineal aconsegueix millors resultats.

LDA:

L'execució i l'avaluació del mètode LDA les hem realitzades, gràcies al fet de treballar amb pocs individus, mitjançant el *leave-one-out cross-validation* (comentat en l'apartat 3.5), obtenint els resultats següents:

- *Accuracy*: 0.95
- *Precision*: 0.99
- *Recall*: 0.94
- *F1*: 0.96

En tant que el nostre objectiu pel classificador és preveure l'abandonament dels alumnes per tal que el tutor d'estudis en pugui prendre mesures que els hi donin suport, el que volem és minimitzar els fn (alumnes que es classifiquen en el grup dels que acaben, Ω_1, quan en realitat són dels que abandonen, Ω_2) i això correspon, tal com s'han definit les mesures considerades, a maximitzar el *recall*. Observem que aquest pren el valor més baix de les mesures considerades, ja que com podem comprovar en la figura 5, el classificador s'ha equivocat en 9 dels 181 alumnes considerats i 8 d'ells són *false-negative*. Tot i això, hem obtingut un coeficient força proper a 1 així que podem concloure afirmant que aquest seria un bon model de predicció de l'abandonament.

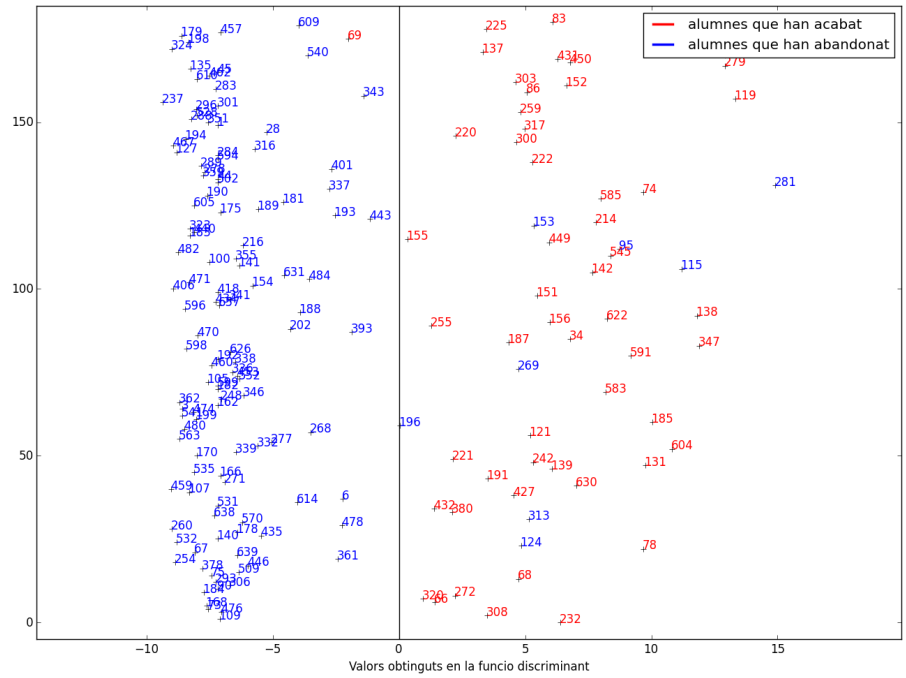

Figura 5: Classificació obtinguda per LDA en Matemàtiques.

4.2.2 Grau d'Enginyeria Informàtica

Aquest ensenyament és el que més s'ha vist afectat pels forats que hem comentat que té la base de dades. Suposem que és degut al fet que a tots aquells alumnes que provenen d'algun cicle relacionat amb la computació, se'ls convaliden o reconeixen assignatures bàsiques i que aquestes no han estat ben incorporades en els arxius proporcionats. És per això que hem hagut de treballar amb només 87 alumnes dins del grup dels que han finalitzat la carrera i 67 dels que l'han abandonada, en total, 154 alumnes.

Les assignatures amb les quals s'ha construït la matriu de dades multivariants per aquest ensenyament són les següents:

Nom de l'assignatura	Àlies utilitzat
Programació I	*PrgI*
Disseny Digital Bàsic	*DDB*
Introducció als Ordinadors	*IO*
Àlgebra	*Alg*
Càlcul	*Calc*
Matemàtica Discreta	*MD*
Física	*Fisc*
Algorísmica	*Algo*
Programació II	*PrgII*
Estructura de Dades	*ED*

Taula 3: Assignatures considerades pel Grau d'Enginyeria Informàtica.

PCA:

En aplicar el mètode PCA en aquest ensenyament hem observat que, així com el vector que defineix la primera component principal forma una mitjana ponderada ja comentada anteriorment, els coeficients del que defineix la segona discriminen mitjançant el signe, en tots tres grups d'alumnes considerats, les assignatures de primer en dos grups ben diferenciats: el format per *PrgI*, *PrgII* i *ED*; i el que està compost per *MD*, *Calc*, *Alg*, *Fisc*, *IO* i *DDB*. L'assignatura *Algo* en funció del grup d'alumnes considerats es posiciona a favor d'un grup o un altre. Veiem-ho:

Alumnes que han acabat:

En aquest cas la separació de les variables considerades adjudica l'assignatura en discòrdia, Algorísmica, en el segon grup tal com podem veure al gràfic 6 i en els signes dels coeficients d'a_2:

$a_1 = (-0.299, -0.393 - 0.298, -0.431, -0.277, -0.284, -0.226, -0.223, -0.384, -0.274)$,

$a_2 = (+0.322, -0.119, -0.236, -0.228, -0.221, -0.256, -0.289, -0.141, +0.566, +0.485)$.

Aquesta classificació sembla donar-se a les dificultats amb les quals es van trobar els alumnes alhora de superar-les. Així, les assignatures que conformen el primer grup, que a priori només tenen en comú el fet de ser les úniques de primer que s'imparteixen fent servir el llenguatge de programació *Java*, són també les que tenen una mitjana més baixa pels alumnes considerats.

El comportament dels alumnes al primer curs d'aquest ensenyament per part dels que l'han finalitzat és tan similar que el criteri de Kaiser demanava de quatre components principals, ja que la variabilitat que recullen les dues primeres és només d'un 59.10%. A més, gran part d'aquesta variabilitat ve donada per aquells alumnes (situats en la part inferior dreta del biplot) als quals hem hagut de substituir els 0.0 de les seves qualificacions finals per les mitjanes de les assignatures pertinents.

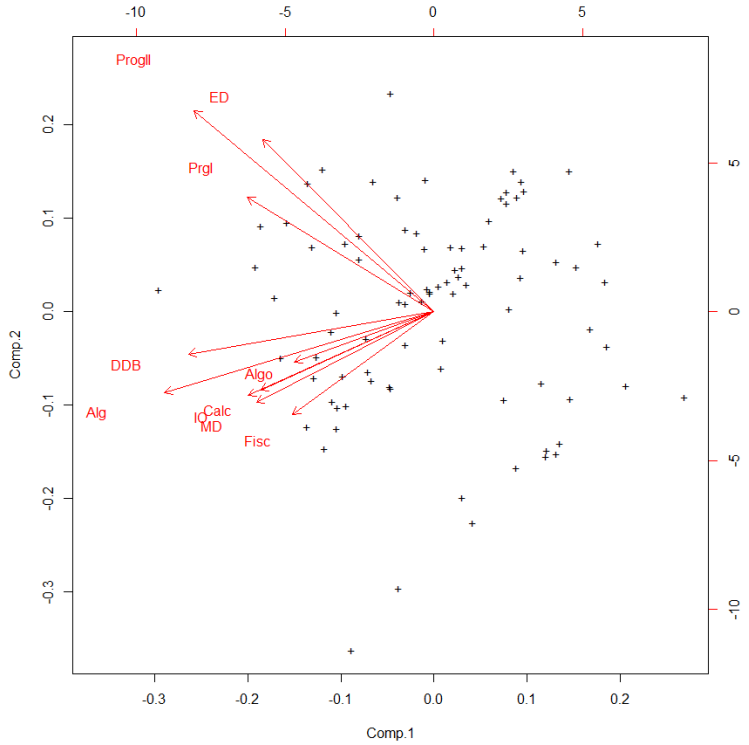

Figura 6: Biplot per Informàtica dels alumnes que han acabat.

Alumnes que han abandonat:

Pel que fa a la classificació de les assignatures per aquest grup d'alumnes, tot i la distància que s'observa al gràfic 7, podem veure en els coeficients del a_2 obtingut

$$a_1 = (+0.367, +0.278, +0.242, +0.308, +0.392, +0.263, +0.386, +0.348, +0.265, +0.267),$$

$$a_2 = (+0.381, -0.074, -0.110, -0.269, -0.308, -0.264, -0.329, +0.144, +0.480, +0.490)$$

que *Algo* comparteix el signe amb les del primer grup. I és que aquí el PCA ha optat per discriminar en funció del tipus d'assignatura que és, essent les del primer grup i *Algo* les que són més de programació, i la resta les que menys. Podem observar a més, que la distribució dels alumnes al gràfic s'ajusta a la destresa que tenen per unes o altres assignatures, tenint així els més programadors envoltant les assignatures posicionades en la part superior del gràfic i els que menys en les de la part inferior.

Aquí, a causa de l'existència d'aquests dos tipus d'alumnes d'entre els considerats, la variabilitat explicada ha pujat a un 68.75%, reduint el nombre de components demanades pel criteri de Kaiser a tres.

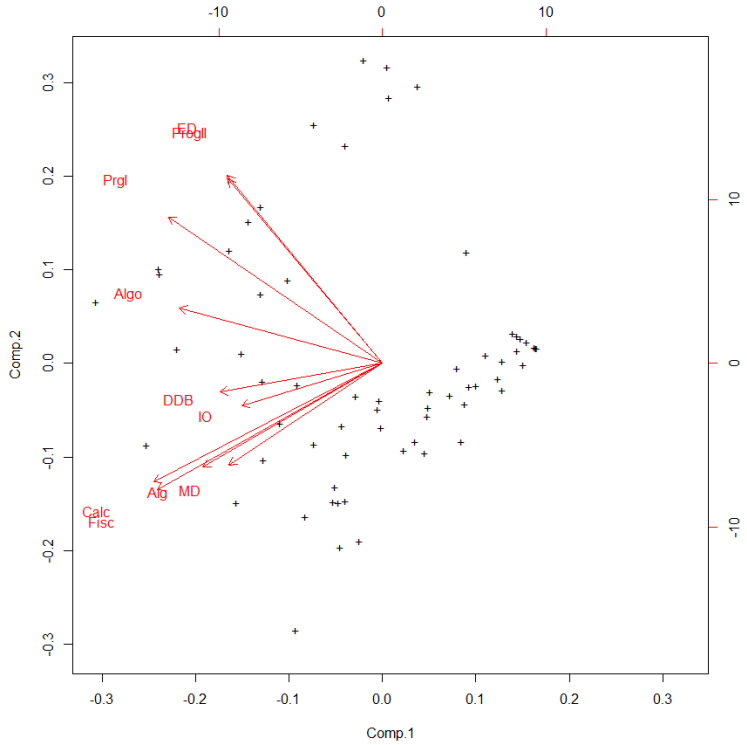

Figura 7: Biplot per Informàtica dels alumnes que han abandonat.

Unió dels alumnes que han acabat i dels que han abandonat:

$a_1 = (+0.281, +0.366, +0.342, +0.314, +0.287, +0.339, +0.315, +0.269, +0.332, +0.302)$,

$a_2 = (-0.362, +0.074, +0.119, +0.272, +0.305, +0.269, +0.338, -0.090, -0.496, -0.491)$.

Finalment, si considerem l'unió dels alumnes que han acabat i dels que han abandonat l'ensenyament, obtenim com és d'esperar un augment considerable en quant a la variabilitat explicada per les dues components principals, essent aquesta d'un 82.88% i satisfent les demandes del criteri de Kaiser. L'assignatura Algorísmica que ha estat anteriorment classficada en cada un dels grups d'assignatures discrimats, torna a compartir signe amb les del primer tot i que ara ha quedat clarament posicionada en mig dels dos grups com es pot observar a la figura 8.

Pel que fa a la representació dels alumnes en funció de les dues components principals, observem al gràfic de la figura 9, com la primera estableix una classificació (considerant el llindar al 0) amb només sis errors i per tant una *accuracy* del 96.10%. El *recall* és d'un 92.54% i es té una *precision* de 98.41%, sens dubtes uns resultats res menyspreables de cara a establir un model de predicció de l'abandonament. Anem a veure ara els resultats obtinguts amb un mètode de classificació supervisada.

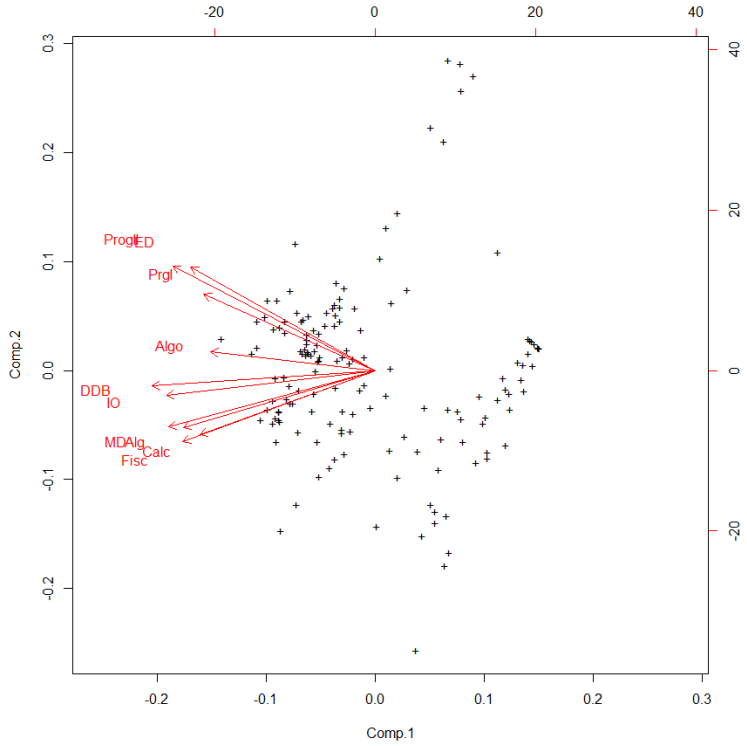

Figura 8: Biplot per Informàtica dels alumnes totals considerats.

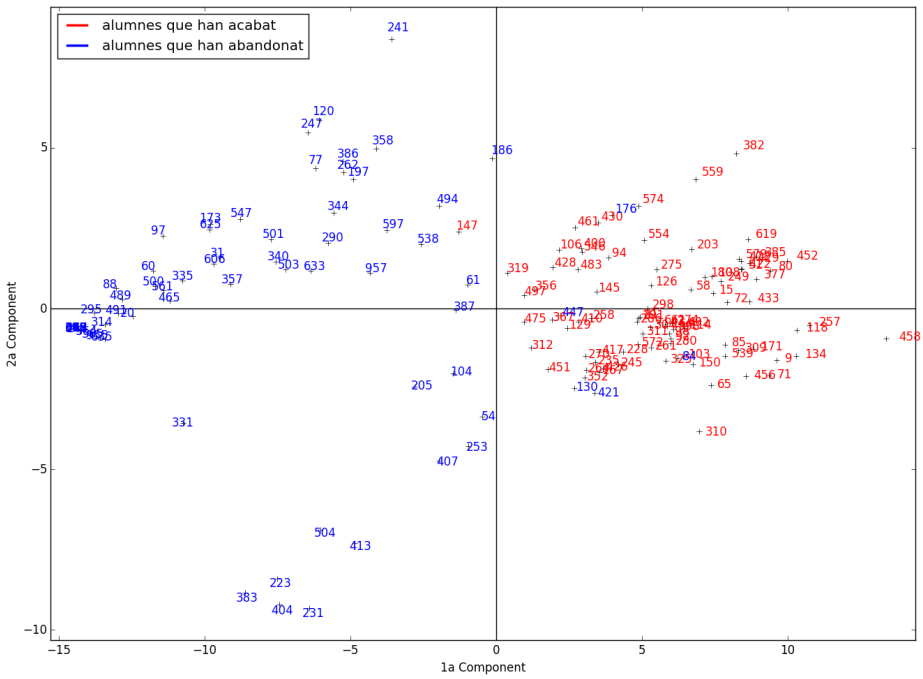

Figura 9: Representació dels alumnes d'Informàtica.

LDA:

Com també ens ha passat amb el Grau de Matemàtiques, al treballar amb només 154 alumnes hem pogut dur a terme l'execució i avaluació del mètode LDA amb el LOOCV, obtenint així els següents resultats:

- *Accuracy*: 0.95
- *Precision*: 0.98
- *Recall*: 0.89
- *F1*: 0.94

La mesura *recall*, que és la més interessant per nosaltres tal com hem explicat anteriorment, torna a tenir el valor més baix de totes les mesures considerades. A la figura 10 podem observar com 8 dels alumnes totals considerats s'han classificat erròniament essent 7 alumnes els que no hauríem aconseguit preveure el seu abandonament. Tot i això, el mètode sembla robust i suficientment efectiu com per servir d'eina de suport per als tutors d'estudis. Observem a més, que tot i que no és la nostra prioritat minimitzar els fp (alumnes que es classifiquen en el grup dels que abandonaran quan en realitat no ho faran), aquest mètode ha demostrat ser, tant en el Grau de Matemàtiques com en aquest, òptim pel que fa al temps de dedicació a l'hora de donar suport a un alumne, ja que pràcticament no s'equivoca en alertar que un alumne acabarà abandonant i prova d'això és l'alt coeficient de *precision* obtingut en ambdós ensenyaments.

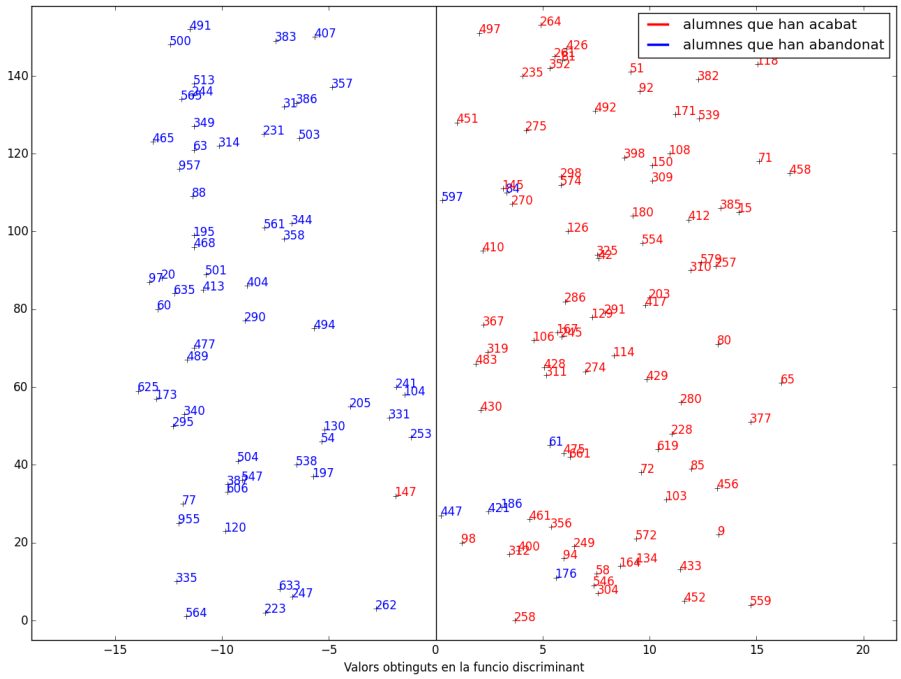

Figura 10: Classificació obtinguda per LDA en Informàtica.

4.2.3 Grau de Dret

Finalment, l'ensenyament amb més matriculats dels que disposem ens ha proporcionat un nombre final de 1468 alumnes per treballar, d'entre els quals 1100 s'han considerat com a alumnes que han acabat la carrera i els altres 368 com els que l'han abandonada.

Les assignatures de primer d'aquest Grau i l'ordre en el qual les hem considerades a l'hora de construir la matriu de dades multivariants són:

Nom de l'assignatura	**Àlies utilitzat**
Tècniques de Treball i Comunicació	*TTC*
Ciència Política	*CP*
Economia	*Econ*
Fonaments del Dret	*FD*
Principis i Institucions Constitucionals	*PIC*
Dret Civil de la Persona	*DCP*
Organització Territorial de l'Estat	*OTE*
Fonaments del Dret Penal i Teoria del Delicte	*FDPTD*
Història del Dret	*HD*
Dret Romà	*DR*

Taula 4: Assignatures considerades pel Grau de Dret.

Les interpretacions dels resultats obtinguts per aquest ensenyament s'han vist limitades pels coneixements que en tenim d'aquest Grau, ja que tant les relacions que mantenen les assignatures entre si com la tendència de comportament que segueixen els seus alumnes, ens són a priori desconegudes. Veiem com tot i així, les dades han parlat per si soles.

PCA:

Alumnes que han acabat:

La representació biplot dels alumnes que van acabar l'ensenyament de Dret en el període del 2009-2014 no deixa lloc al dubte, el comportament d'aquests és homogeni. Tant, que tot i el gran nombre d'alumnes amb els quals s'ha treballat, les dues primeres components només aconsegueixen explicar una variabilitat del 52.49% i en el gràfic de la figura 11 s'observa el núvol d'alumnes (d'un mateix grup) més compacte que s'ha vist fins ara.

$$a_1 = (-0.199, -0.262, -0.283, -0.309, -0.362, -0.356, -0.395, -0.297, -0.349, -0.302)$$

$$a_2 = (+0.078, +0.251, -0.021, -0.106, +0.201, +0.065, +0.327, +0.005, -0.869, +0.114)$$

Si ens fixem en els coeficients del vector propi que defineix la segona component principal, notem que les assignatures *Econ*, *FD* i *HD* són les úniques amb coeficient negatiu. D'aquestes, és només l'última, Història del Dret, la que ocupa una posició radicalment diferent de la resta en el biplot de la figura 12 obtingut. Sembla que ens trobem en un cas semblant al del Grau de Matemàtiques i l'assignatura Física (figura 1), on com en aquest cas *HD* no manté una connexió directa amb la resta

d'assignatures, la seva nota no sembla caracteritzar els alumnes considerats i per això es troba aïllada.

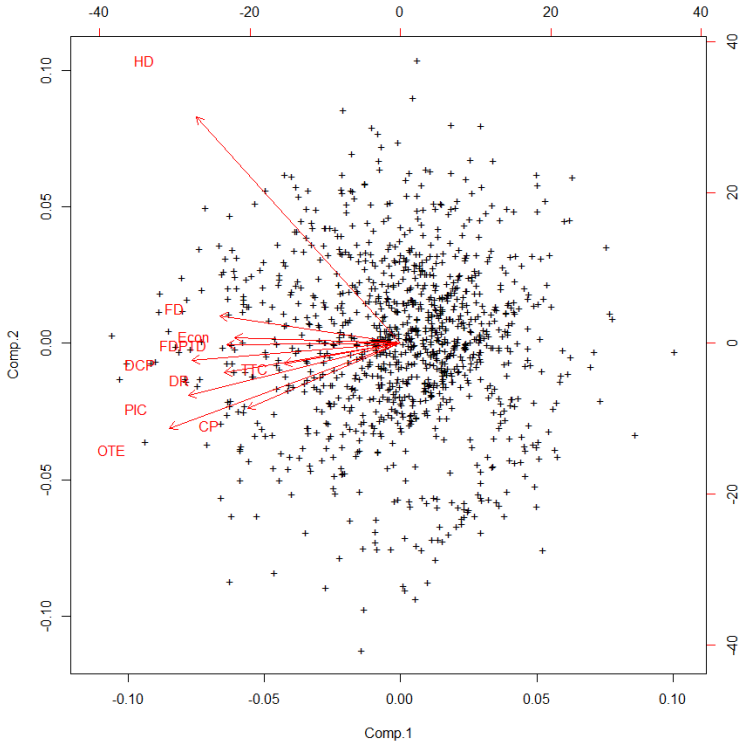

Figura 11: Biplot per Dret dels alumnes que han acabat.

Alumnes que han abandonat:

El vector propi a_2 que forma la segona component principal, en aquest cas ens aporta molta informació sobre el grup d'alumnes considerat i prova d'això és que la variabilitat explicada per les dues primeres components ha augmentat en aquest cas a un 69.72%. Vegem-ne per què:

$$a_1 = (+0.265, +0.324, +0.356, +0.346, +0.322, +0.291, +0.297, +0.338, +0.292, +0.318),$$

$$a_2 = (-0.449, -0.509, +0.262, -0.332, -0.094, +0.306, +0.315, +0.214, +0.331, -0.050).$$

El primer que s'observa dels seus coeficients és que n'hi han dos que en valor absolut són molt més petits que la resta: -0.094 i -0.050 que corresponen respectivament a *PIC* i *DR*. Per algun motiu que desconeixem, aquestes dues assignatures no aporten gaire informació sobre els alumnes que abandonen aquest ensenyament, i per això, en aplicar el PCA s'han obtingut uns coeficients tan baixos. L'altra informació que en podem extreure dels coeficients i del biplot obtingut (figura 12), és com el signe d'aquests discrimina de manera representativa dos grups d'assignatures: el format per *CP*, *TTC* i *FD*; i el que conté *FDPTD*, *Econ*, *OTE*, *HD* i *DCP*. Enmig d'aquests dos i quasi paral·lelament a l'eix $y = 0$, es troben les altres dues assignatures abans comentades. Indagant una mica en la base dades trobem que la distinció d'aquests dos grups d'assignatures es deu a la dificultat que van presentar pels alumnes considerats, essent així les del primer grup les que van presentar menys

dificultat, ja que són les que tenen per aquests alumnes les mitjanes més altes (amb valors entre el 3.68 i el 4.44), i les del segon grup les que més (les seves mitjanes oscil·len entre l'1.84 i el 3.02).

Aquesta distinció en les assignatures també repercuteix en la distribució dels alumnes i així es percep al gràfic obtingut: com més a prop es troben de les assignatures del segon grup, més bona nota van tenir en aquestes i en la resta en general, i el mateix s'obté en cas contrari. La franja quasi vertical d'alumnes de la part dreta del gràfic, correpon a alumnes que sembla que van abandonar l'ensenyament abans de completar les deu assignatures de primer, en tant que són aquells que més zeros tenen en la matriu de dades multivariants.

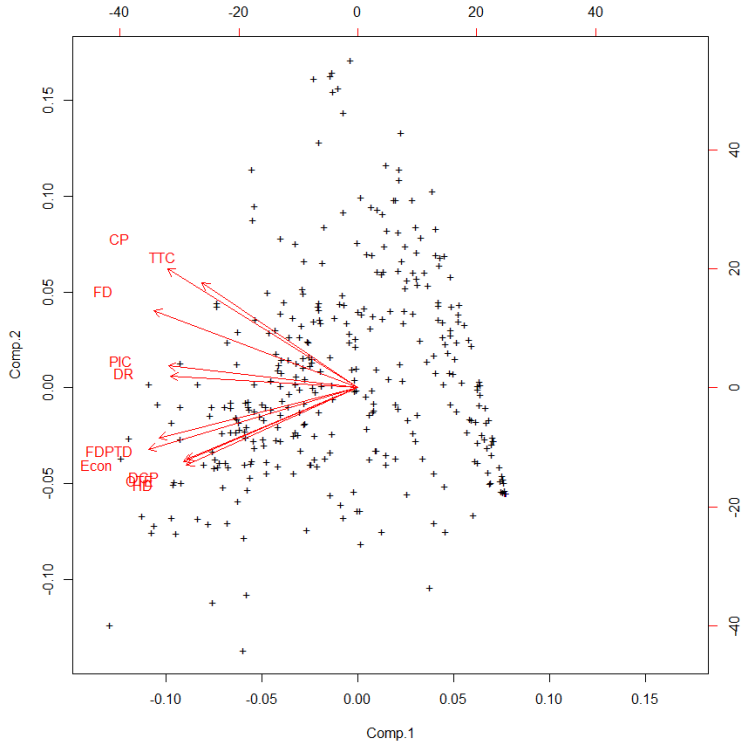

Figura 12: Biplot per Dret dels alumnes que han abandonat.

Unió dels alumnes que han acabat i dels que han abandonat:

Pel que fa als resultats obtinguts en aplicar el mètode PCA a l'unió dels alumnes totals considerats, es té que el vector que defineix la segona component principal ve donat per:

$$a_1 = (-0.218, -0.260, -0.321, -0.289, -0.317, -0.341, -0.376, -0.350, -0.352, -0.306),$$

$$a_2 = (+0.459, +0.567, -0.147, +0.360, +0.101, -0.246, -0.257, -0.132, -0.391, +0.092).$$

Novament tornem a tenir la mateixa partició d'assignatures per signe que en el cas dels alumnes que han abandonat. De fet, la represetnació de totes elles al biplot obtingut (figura 13) és molt semblant a l'anterior, tot i que ara, a l'haver considerat també els alumnes que han finalitzat l'ensenyament (pels quals ja hem vist que

aquesta distinció no era perceptible), la diferència entre aquestes assignatures és menor i per això estan més properes al gràfic.

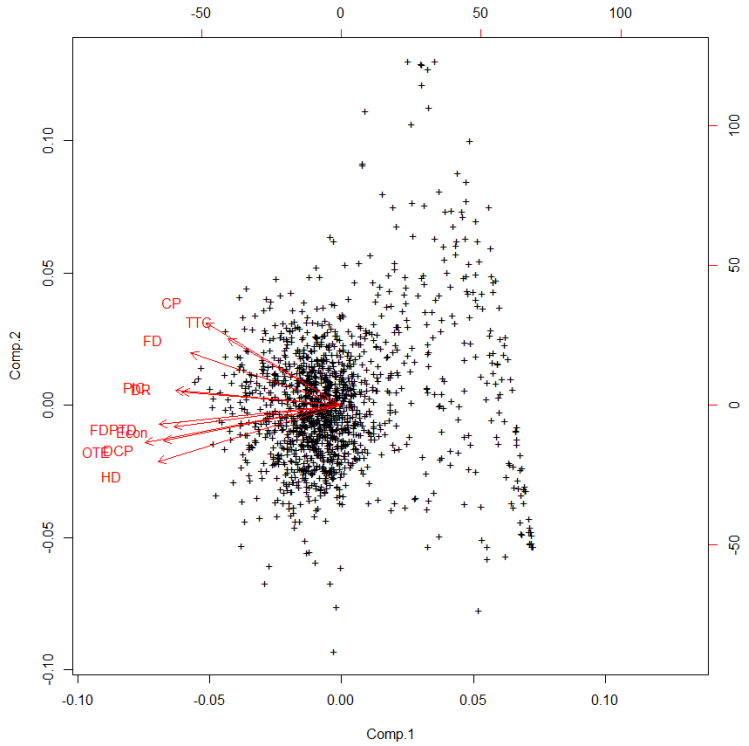

Figura 13: Biplot per Dret dels alumnes totals considerats.

Malgrat haver considerat dos grups d'alumnes amb comportaments tan diferents envers les assignatures considerades, la variabilitat explicada és només d'un 76.08% i com podem observar al gràfic de la figura 14, aquest cop sembla complicat escollir un llindar amb el qual la primera component principal aconsegueixi discriminar bé el grup dels alumnes que abandonen dels que acaben.

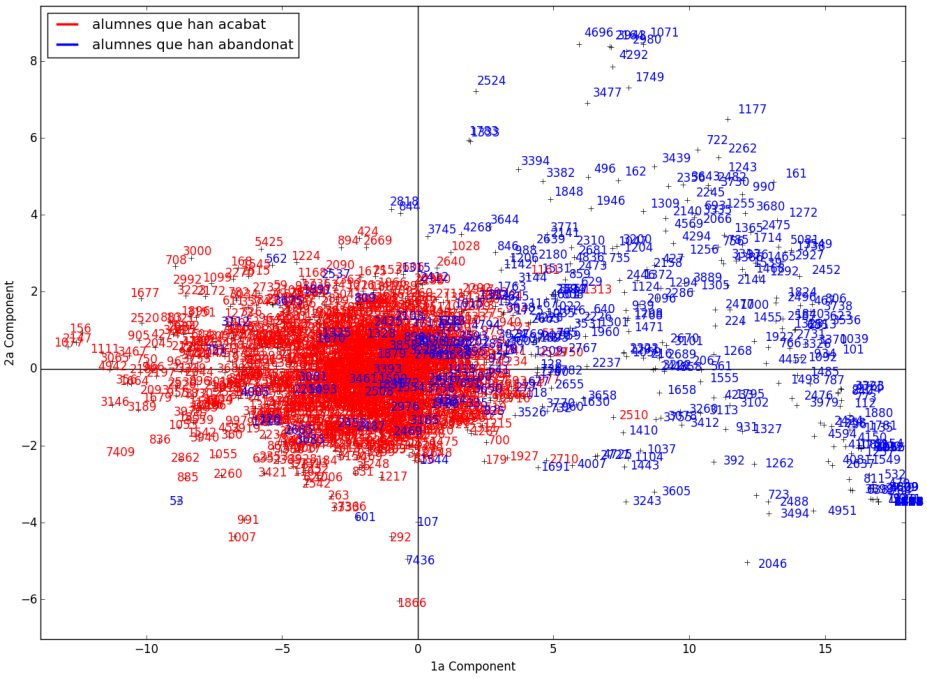

Figura 14: Representació dels alumnes de Dret.

LDA:

En tant que per Dret hem treballat amb un nombre considerable d'alumnes, 1468, l'avaluació del mètode LDA fent servir *leave-one-out* requeria massa temps d'execució, així que hem optat per utilitzar la validació creuada de K iteracions amb l'estàndard $K = 10$. Si bé és cert que aquest mètode sí que és sensible a l'elecció dels K conjunts d'entrenaments escollits, aquests s'han agafat de manera aleatòria desordenant el conjunt de dades abans de començar. Com que això fa que en cada execució s'obtinguin valors lleugerament diferents, hem optat per realitzar-ne 50 execucions i quedar-nos amb les mitjanes dels valors obtinguts. Així, les mesures finals recollides en la classificació d'aquest ensenyament són:

- *Accuracy*: 0.93
- *Precision*: 0.92
- *Recall*: 0.80
- *F1*: 0.85

A priori, no semblen uns resultats dolents si tenim en compte que el mètode PCA no havia aconseguit discriminar molt bé cap dels dos grups amb les dues components principals. És un fet però, que el *recall*, la mesura que més ens interessa estudiar, s'ha allunyat del 90% obtingut en els Graus de Matemàtiques i d'Enginyeria Informàtica i que si ens fixem en el gràfic 15, el nombre d'alumnes que abandonen que no hem aconseguit detectar no és gens menyspreable.

Per entendre bé els resultats obtinguts i poder arribar a un veredicte respecte a la utilitat d'aquest mètode a l'hora de construir un model de predicció de l'abandonament, hem de tenir molt present els de l'Anàlisi de Components Principals. En ells hem vist que tant la variabilitat explicada per les dues primeres components (només d'un 76.08%), com la representació dels alumnes en funció d'aquestes (figura 14), apunten al fet que part de l'abandonament en Dret té un comportament molt similar al dels alumnes que sí que acaben i que per tant, aquest no es veu caracteritzat per les notes que s'obtenen a primer, tot i ser en aquest curs en el qual es produeix l'abandonament més gran en la carrera (tal com s'ha reflectit en la taula 1).

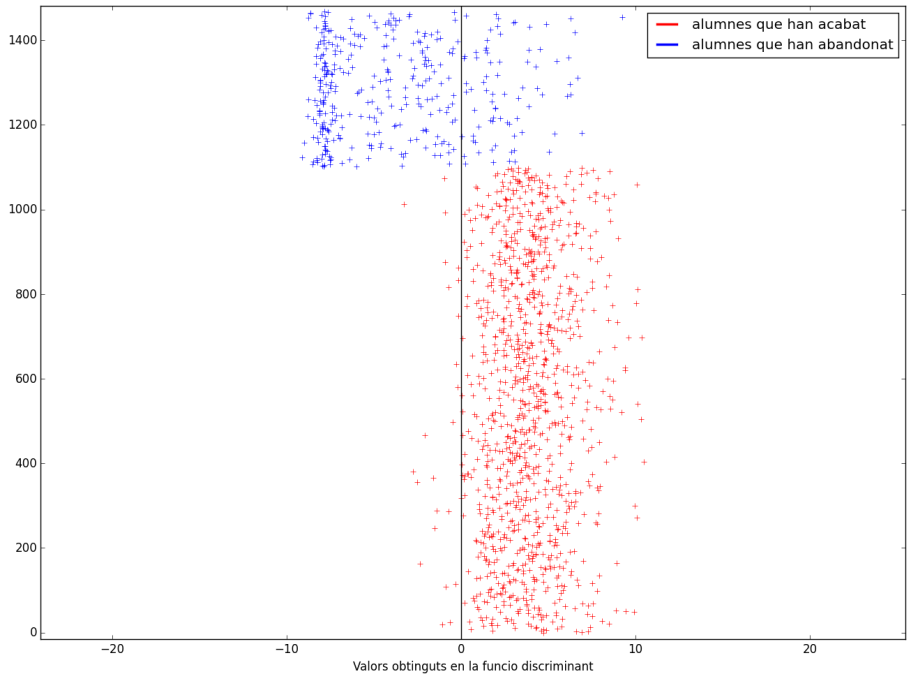

Figura 15: Classificació obtinguda per LDA en Dret.

Així, podem concloure que l'Anàlisi Discriminant Lineal pot ser un bon mètode per a predir l'abandonament també en aquest ensenyament, ja que tot i no tenir l'eficàcia que ha tingut en els altres dos conjunts estudiats, gran part d'aquest que no es detecta sembla degut a factors externs a la trajectòria de l'alumne en la carrera i per tant, s'hauria d'incorporar més informació sobre els estudiants i no limitar-se a les notes, per a poder millorar el sistema de predicció de l'abandonament.

4.3 Validació dels resultats

Com hem comentat en l'apartat 3.5, per tal de concloure si els resultats obtinguts en la predicció de l'abandonament amb el mètode LDA són estadísticament significatius o no, hem fet servir el test de les permutacions.

En ser el LDA un mètode de classificació supervisat, el que s'ha fet amb el test de les permutacions és modificar de manera aleatòria, en cada una de les M iteracions que s'han considerat, les etiquetes dels alumnes que es fan servir per a la construcció del model de predicció. D'aquesta manera s'han obtingut M mesures d'avaluació del LDA amb les quals poder construir el *p-valor*, tot comparant si són iguals o més bones que les que s'han obtingut amb les dades originals.

Com ja s'ha comentat en l'anàlisi dels resultats, la mesura que més informació ens aporta a nosaltres sobre com és de bona la classificació realitzada és el *recall*. És per això que hem decidit que fos aquesta la que es tingués en compte per la construcció del *p-valor*:

Si diem r al recall obtingut en la classificació del LDA amb les dades originals:

$$p\text{-}valor = \frac{\text{n}^\circ \text{ iteracions en les quals s'ha obtingut un recall} \geq r}{M}$$

El valor de significació triat és $\alpha = 0.01$, i el nombre d'iteracions és de $M = 2000$ pel cas de Matemàtiques i Enginyeria Informàtica (on el nombre d'alumnes amb els quals treballem és només de 181 i 154 respectivament) i $M = 800$ pel Grau de Dret (on el volum de dades augmenta fins a 1468 alumnes). Per tal de reduir també el temps d'execució del test, s'ha optat per considerar en tots tres ensenyaments l'avaluació del LDA fent servir la validació creuada amb $K = 5$ iteracions, deixant enrere el *leave-one-out* considerat abans en els Graus de Matemàtiques i Informàtica i el *10-fold cross-validations* pel Grau de Dret.

Així, els resultats obtinguts han estat els següents:

- Pel Grau de Matemàtiques s'ha obtingut un *recall* amb les dades originals de 0.93; i com que només una de les 2000 iteracions ha aconseguit igualar o superar aquesta marca, el *p-valor* calculat és: 0.0005.

- Pel Grau d'Enginyeria Informàtica s'ha obtingut per les dades originals un *recall* de 0.88; i com només 9 de les 2000 iteracions han aconseguit igualar o superar aquest resultat, el *p-valor* calculat és: 0.0045.

- Pel Grau de Dret s'ha obtingut un *recall* de 0.80 amb les dades originals; i en cap de les 800 iteracions que s'han pogut realitzar, s'ha aconseguit igualar o superar aquest resultat. És per això que el *p-valor* calculat és: 0.0.

Com que en tots tres ensenyaments s'ha aconseguit un *p-valor* inferior al valor de significació α considerat, 0.001, podem validar els resultats obtinguts en la predicció de l'abandonament i concloure afirmant que aquests sí que són estadísticament significatius.

5 Conclusions i treballs futurs

Gràcies a l'anàlisi conjunt que s'ha fet sobre l'abandonament amb els mètodes de l'Anàlisi de Components Principals (PCA) i l'Anàlisi Discriminant Lineal (LDA), hem pogut observar com gran part dels factors que el produeixen estan caracteritzats per la trajectòria que segueixen els alumnes al primer curs de carrera, tenint un comportament prou diferent dels que acaben per poder identificar-los i classificar-los correctament.

El mètode PCA ens ha permès, juntament amb les consultes que hem realitzat sobre la base de dades, extreure'n informació i caracteritzar el comportament que tenen els alumnes d'ambdós grups (els que acaben i els que abandonen) envers les assignatures de primer. Així, mentre que per Matemàtiques o Dret s'ha obtingut distincions clares entre grups d'assignatures degudes a la dificultat que suposen pels alumnes, per Informàtica en canvi, s'han discriminat aquelles assignatures més enfocades en la programació de la resta. La representació dels alumnes en funció de les variables discriminants trobades per aquest mètode, les components principals, ens han proporcionat una primera aproximació a un sistema de classificació que a excepció del Grau de Dret, ha donat molts bons resultats a l'hora de discriminar els grups d'alumnes.

La classificació definitiva l'hem obtinguda mitjançant el mètode LDA, que ha continuat amb els bons resultats en la discriminació obtinguda pel PCA en els ensenyaments de Matemàtiques i d'Enginyeria Informàtica, i ens ha sorprès gratament obtenint també unes bones mesures en la classificació per Dret. Com ja hem comentat al llarg de l'anàlisi, l'objectiu que perseguíem amb aquest mètode era maximitzar les deteccions d'alumnes que abandonen, i amb el que ens hem trobat és que el LDA minimitza sobretot els errors en la classificació a aquest grup. És a dir, que un model de predicció basat en aquest mètode, permetrà als tutors d'estudis focalitzar realment la seva atenció en els alumnes que tenen major risc d'abandonament.

Encara que ens hagués agradat treballar amb més mètodes de classificació per poder comparar els resultats obtinguts, com ara el *Support Vector Machines*, un mètode d'aprenentatge supervisat i de classificació lineal com el LDA, les dificultats inicials amb les quals ens hem trobat a l'hora de tractar i processar les dades no ens han permès tenir prou temps per poder estudiar-los amb la mateixa profunditat que el PCA i el LDA, dels quals s'ha aconseguit tenir una idea tant teòrica com pràctica. Així, com a possibles treballs futurs, suggerim tant l'opció d'estudiar amb les mateixes dades algun altre mètode de classificació i comparar els resultats obtinguts amb aquests; com aplicar els mètodes implementats a més dades, ja siguin noves dades d'algun altre ensenyament, per observar quins comportaments es troben en els dos grups d'alumnes a considerar, com simplement ampliar les dades dels estudis que ja es tenen als anys 2015-2017 per tal d'observar si els comportaments identificats es veuen afectats, arribant a poder caracteritzar fins i tot patrons en promocions d'alumnes.

Pel que fa a aquest treball i als resultats obtinguts en ell, creiem que un model de predicció de l'abandonament és perfectament factible i encoratgem a tot l'equip del projecte d'innovació docent a seguir treballant-hi, per tal de proporcionar una eina de suport als tutors d'estudis prou potent perquè els permeti aconsellar i guiar als seus alumnes.

Referències

[1] Cuadres, Carles M.: *Nuevos Métodos de Análisis Multivariante*, CMC Editions, Barcelona, 2014.

[2] Peña, Daniel: *Análisis de datos multivariantes*, 1a edició, McGraw−Hill Interamericana de España, S.L, Madrid, 2013.

[3] Duda, Richard O.; Hart, Peter E.; Stork, David G.: *Pattern Classification*, 2a edició, John Wiley & Sons, Inc., New York, 2001.

[4] Johnson, Dallas E.: *Métodos multivariados aplicados al análisis de datos*, International Thomson Editores, Mèxic, 2000.

[5] Projecte: *Intelligent Support System for Tutor of Studies* - http://pid-ub.github.io/.

[6] *INDOMAIN, Grup de Millora en Innovació Docent Consolidat* - http://mid.ub.edu/webpmid/content/indomain

[7] Rovira, Sergi; Puertas, Eloi; Igual, Laura: *Data-driven system to predict academic grades and dropout*, PLOS ONE, Public Library of Science, 2017.