Descomposició en valors singulars: introducció i aplicacions

Anàlisi de components principals (PCA) i descomposició en valors singulars (SVD)

PID_00262386

Francesc Pozo Montero
Núria Parés Mariné

Universitat Oberta de Catalunya

Francesc Pozo Montero

Llicenciat en Matemàtiques per la Universitat de Barcelona (2000) i doctor en Matemàtica Aplicada per la Universitat Politècnica de Catalunya (2005). Ha estat professor associat a la Universitat Autònoma de Barcelona i professor associat, col·laborador i actualment professor agregat a la Universitat Politècnica de Catalunya. A més, és cofundador del Grup d'Innovació Matemàtica E-learning (GIMEL), responsable de diversos projectes d'innovació docent i autor de diverses publicacions. Com a membre del grup de recerca consolidat CoDAlab, centra la recerca en la teoria de control i les aplicacions en enginyeria mecànica i civil, com també en l'ús de la ciència de dades per al monitoratge de la integritat estructural i per al monitoratge de la condició, sobretot en turbines eòliques.

Núria Parés Mariné

Llicenciada en Matemàtiques per la Universitat Politècnica de Catalunya (1999) i doctora en Matemàtica Aplicada per la Universitat Politècnica de Catalunya (2005). És professora de la Universitat Politècnica de Catalunya des del 2000 —actualment, com a professora agregada—, cofunfadora del Grup d'Innovació Matemàtica E-learning (GIMEL), responsable de diversos projectes d'innovació docent i autora de diverses publicacions i llibres docents. Com a membre del grup de recerca consolidat LaCàN (UPC), centra la investigació en el desenvolupament de tècniques eficients per a la resolució numèrica d'equacions en derivades parcials i en l'estimació de l'error associat a aquestes simulacions numèriques.

La revisió d'aquest recurs d'aprenentatge UOC ha estat coordinada per la professora: Cristina Cano Bastidas

Segona edició: setembre 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Autoria: Francesc Pozo Montero, Núria Parés Mariné
Producció: FUOC

Índex

1. La maledicció de la dimensió

1.1. Exemple introductori: la interpolació polinòmica pura

Un problema d'interpolació clàssic és trobar, per exemple, la paràbola que s'ajusta millor a tres punts donats, com ara:

$$(x_0,y_0),\ (x_1,y_1) \text{ i } (x_2,y_2).$$

L'equació d'una paràbola és

$$y = ax^2 + bx + c,$$

en què a,b i c són nombres reals. Es tracta, doncs, de trobar el valor de tres paràmetres. Per fer-ho, imposem que els tres punts donats estiguin sobre la paràbola. És a dir, els tres punts han de satisfer l'equació de la paràbola. Això equival al sistema d'equacions lineals següent:

$$y_0 = ax_0^2 + bx_0 + c$$

$$y_1 = ax_1^2 + bx_1 + c$$

$$y_2 = ax_2^2 + bx_2 + c$$

que es pot expressar en forma matricial:

$$\underbrace{\begin{bmatrix} x_0^2 & x_0 & 1 \\ x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \end{bmatrix}}_{\mathbf{M}} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}$$

Matriu de Vandermonde

Per la forma de la matriu $\mathbf{M}$, en què a cada fila hi ha els termes d'una progressió geomètrica, direm que la matriu és de Vandermonde. El nom prové del matemàtic francès Alexandre-Théophile Vandermonde.

Per tant, la solució del sistema —suposant que el determinant de la matriu M no és zero— es pot calcular així:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \mathrm{M}^{-1} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}$$

En aquest cas, el càlcul de la inversa de la matriu M —amb tres files i tres columnes— és assequible i el podríem fer fàcilment. Però imagineu què passaria si, en comptes de buscar la paràbola que s'ajusta millor a tres punts donats, busquéssim el polinomi $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ de grau n que s'ajusta millor a $n+1$ punts donats, com ara:

$$(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$$

En aquest cas, el sistema d'equacions que hauríem de resoldre, en forma matricial, seria:

$$\underbrace{\begin{bmatrix} x_0^n & \cdots & x_0 & 1 \\ x_1^n & \cdots & x_1 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_n^n & \cdots & x_n & 1 \end{bmatrix}}_{\mathrm{M}_{n+1}} \begin{bmatrix} a_n \\ a_{n-1} \\ \vdots \\ a_0 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

També en aquest cas, la solució del sistema —suposant que el determinant de la matriu M_{n+1} no és zero— es pot calcular així:

$$\begin{bmatrix} a_n \\ a_{n-1} \\ \vdots \\ a_0 \end{bmatrix} = \mathrm{M}_{n+1}^{-1} \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Ara bé, el càlcul de la inversa de la matriu M_{n+1} ja no resulta senzill si n és gran. Es pot recórrer, per descomptat, a mètodes numèrics, però es pot demostrar que la matriu de Vandermonde, com ara M_{n+1}, està mal condicionada. Per tant, els petits errors numèrics que es puguin produir en el procés de càlcul de la matriu inversa poden afectar molt negativament la solució del sistema. De fet, la solució final proposada pel mètode numèric pot quedar lluny de la solució real.

Condicionament d'una matriu

El nombre de condició d'una matriu indica, per exemple, com el determinant de la matriu pot ésser afectat per petits canvis en els elements que la formen. Si la matriu està mal condicionada, el sistema d'equacions lineals també ho estarà.

En matemàtiques, i més concretament en el camp de la interpolació polinòmica pura, el problema exposat anteriorment es resol, per exemple, amb el mètode de les diferències dividides de Newton o amb els polinomis de Lagrange. Per mitjà d'aquests dos mètodes, els coeficients dels polinomis interpoladors es calculen sense necessitat de resoldre un sistema d'equacions lineals, la qual cosa evita el problema del mal condicionament i, per tant, la maledicció de la dimensionalitat.

1.2. Un altre exemple: l'enquesta de pressupostos familiars

Enquesta de pressupostos familiars

L'Institut Nacional d'Estadística (www.ine.es) publica anualment l'enquesta de pressupostos familiars (EPF).

L'enquesta de pressupostos familiars (EPF) subministra informació anual sobre la naturalesa i el destí de les despeses de consum, a més de diverses característiques relatives a les condicions de vida de les llars. Les despeses de consum es refereixen tant al flux monetari que destina la llar a pagar determinats béns i serveis de consum final, com al valor dels béns percebuts en concepte d'autoconsum, autosubministrament, salari en espècie, etc. La mida de mostra és de 24.000 llars per any, aproximadament.

La informació de l'enquesta es presenta de diverses maneres. Per exemple, pot estar agrupada per comunitats autònomes —incloent-hi les ciutats autònomes de Ceuta i Melilla:

1) Andalusia

2) Aragó

3) Astúries

4) Balears

5) Canàries

6) Cantàbria

7) Castella i Lleó

8) Castella-la Manxa

9) Catalunya

10) Comunitat Valenciana

11) Extremadura

12) Galícia

13) Comunitat de Madrid

14) Múrcia

15) Navarra

16) País Basc

17) La Rioja

18) Ceuta

19) Melilla

Es mesuren un total de dotze variables:

1) Aliments i begudes no alcohòliques

2) Begudes alcohòliques i tabac

3) Vestit i calçat

4) Habitatge, aigua, electricitat, gas i altres combustibles

5) Mobles, articles de la llar i articles per al manteniment corrent de la llar

6) Sanitat

7) Transport

8) Comunicacions

9) Oci i cultura

10) Ensenyament

11) Restaurants i hotels

12) Altres béns i serveis

Per exemple, la taula 1 mostra la despesa mitjana per persona (en euros) en cadascuna de les disset comunitats autònomes i les dues ciutats autònomes espanyoles, en relació amb els aliments i les begudes no alcohòliques. Amb aquesta taula de doble entrada és fàcil extreure alguna conclusió. Per exemple, al País Basc la despesa mitjana per persona —d'aliments i begudes no alcohòliques— és superior; en canvi, a Ceuta és on aquesta despesa mitjana és inferior. Si afegim més columnes a la taula 1, és a dir, si hi incloem la informació de més variables, possiblement les conclusions que es podran treure seran més interessants. Amb tot, també és més difícil veure o inferir alguna conclusió, ja que tindrem una matriu de dades amb $12 \times 19 = 228$ despeses mitjanes.

I si en lloc de tenir la informació agrupada per comunitats autònomes la tinguéssim per províncies? Aleshores disposaríem de $50 \times 12 = 600$ despeses mitjanes. Imagineu la dificultat d'obtenir alguna conclusió només observant aquesta informació.

Províncies

Hem considerat cinquanta províncies espanyoles, és a dir, no hem tingut en compte les dues ciutats autònomes de Ceuta i Melilla.

Com es pot veure, és fàcil que en augmentar el nombre d'informació disponible la informació resultant sigui difícil d'interpretar i de visualitzar i, també, que sigui difícil extreure'n alguna conclusió.

Aquest efecte és el que anomenem *maledicció de la dimensionalitat*. De forma més precisa, si augmentem la dimensió de la informació que tenim, aquesta esdevé més rica. Alhora, però, es fa més difícil d'interpretar. En aquest mòdul veurem dues tècniques per reduir la dimensionalitat, tot mantenint la riquesa de la informació, i aprendrem com podem expressar les nostres dades de manera que sigui més fàcil extreure'n característiques.

Taula 1. Despesa mitjana per persona (en euros)

Comunitat autònoma	Aliments i begudes no alcohòliques
Andalusia	1.533,39
Aragó	1.755,21
Astúries	1.777,14
Balears	1.697,69
Canàries	1.460,25
Cantàbria	1.793,89
Castella i Lleó	1.780,34
Castella-la Manxa	1.432,40
Catalunya	1.833,10
Comunitat Valenciana	1.513,25
Extremadura	1.317,60
Galícia	1.678,53
Comunitat de Madrid	1.639,72
Múrcia	1.662,63
Navarra	1.853,64
País Basc	1.959,03
La Rioja	1.679,08
Ceuta	1.327,42
Melilla	1.473,79

Font: Enquesta de pressupostos familiars 2017 (Institut Nacional d'Estadística)

2. Anàlisi de components principals

L'anàlisi de components principals (PCA, de l'anglès *principal component analysis*) és un mètode simple i no paramètric per extreure informació rellevant de conjunts de dades que poden ser confusos. A més, pot servir per donar arguments o indicacions sobre com reduir la dimensió d'un conjunt complex de dades i així revelar possibles estructures amagades o característiques interessants. Les aplicacions són diverses.

Considerem un primer exemple senzill per il·lustrar com funciona aquesta tècnica. Es tracta d'un famós conjunt de dades, l'anomenat *Iris*, que proporciona les mesures en centímetres de les variables de longitud i amplada del sèpal i de longitud i amplada del pètal, respectivament, per a cinquanta flors de cadascuna de les espècies *Iris setosa*, *Iris versicolor* i *Iris virginica*. D'entrada, per simplificar l'estudi de la tècnica, veurem una mostra de quinze flors, cinc de cada tipus, com mostra la taula 2.

Iris

Les dades van ser recollides per Edgar Anderson l'any 1935 i publicades en l'article "The irises of the Gaspe Peninsula", *Bulletin of the American Iris Society*, 59, p. 2-5.

Taula 2. Longitud i amplada del sèpal i longitud i amplada del pètal (en centímetres)

flor	long. sèpal	ampl. sèpal	long. pètal	ampl. pètal	tipus
1	5.1	3.5	1.4	0.2	*setosa*
2	4.9	3.0	1.4	0.2	*setosa*
3	4.7	3.2	1.3	0.2	*setosa*
4	4.6	3.1	1.5	0.2	*setosa*
5	5.0	3.6	1.4	0.2	*setosa*
6	7.0	3.2	4.7	1.4	*versicolor*
7	6.4	3.2	4.5	1.5	*versicolor*
8	6.9	3.1	4.9	1.5	*versicolor*
9	5.5	2.3	4.0	1.3	*versicolor*
10	6.5	2.8	4.6	1.5	*versicolor*
11	6.3	3.3	6.0	2.5	*virginica*
12	5.8	2.7	5.1	1.9	*virginica*
13	7.1	3.0	5.9	2.1	*virginica*
14	6.3	2.9	5.6	1.8	*virginica*
15	6.5	3.0	5.8	2.2	*virginica*

Font: Edgar Anderson (1935). "The irises of the Gaspe Peninsula"

En aquest cas, hem considerat una mostra de quinze flors, de les quals hem mesurat quatre variables:

1) longitud del sèpal

2) amplada del sèpal

3) longitud del pètal

4) amplada del pètal

En un cas general, podem considerar que cal mesurar m variables d'un total de n elements o experiments i organitzar tota aquesta informació en una matriu $\mathbf{X}$ de n files i m columnes:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix} \in \mathcal{M}_{n\times m}(\mathbb{R})$$

Fixeu-vos que $\mathcal{M}_{n\times m}(\mathbb{R})$ representa l'espai vectorial de les matrius de dimensió $n \times m$ amb coeficients reals.

De la matriu anterior, l'i-èssim vector fila

$$x_i^T = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \end{bmatrix}$$

representa els valors de totes les variables per a un dels elements de la mostra, mentre que el j-èssim vector columna

$$v_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{bmatrix}$$

representa el valor de la j-èssima variable per a tots els elements de la mostra.

En el cas del nostre exemple,

$$x_2^T = \begin{bmatrix} 4.9 & 3.0 & 1.4 & 0.2 \end{bmatrix}$$

mentre que

$$v_3 = \begin{bmatrix} 1.4 \\ 1.4 \\ 1.3 \\ 1.5 \\ 1.4 \\ 4.7 \\ 4.5 \\ 4.9 \\ 4.0 \\ 4.6 \\ 6.0 \\ 5.1 \\ 5.9 \\ 5.6 \\ 5.8 \end{bmatrix}$$

2.1. Preprocessament: l'escalat de les dades

Atès que les variables poden tenir diferents escales i magnituds, cal aplicar a la matriu de dades un preprocés per escalar-ne les variables, de manera que totes tinguin de mitjana 0 i de desviació tipus 1.

Anomenem μ_j la mitjana aritmètica de la variable j-èssima, que es defineix així:

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \frac{x_{1j} + x_{2j} + \cdots + x_{nj}}{n}$$

De la mateixa manera, anomenem σ_j^2 la variància de la variable j-èssima, que es defineix així:

$$\sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2 = \frac{(x_{1j} - \mu_j)^2 + (x_{2j} - \mu_j)^2 + \cdots + (x_{nj} - \mu_j)^2}{n-1}$$

Finalment, cada element de la matriu $\mathbf{X}$ es normalitza de la manera següent:

$$\check{x}_{ij} := \frac{x_{ij} - \mu_j}{\sigma_j},\ i = 1, \ldots, n,\ j = 1, \ldots, m$$

Una matriu on les columnes tenen mitjana zero s'anomena matriu centrada.

Desviació tipus

La desviació tipus es defineix com l'arrel quadrada de la variància.

En el cas del nostre exemple, la mitjana i la desviació tipus de la primera variable són:

$$\mu_1 = \frac{5.1 + 4.9 + \cdots + 6.3 + 6.5}{15} = \frac{88.6}{15} = 5.906667$$

$$\sigma_1 = \sqrt{\frac{(5.1 - \mu_1)^2 + \cdots + (6.5 - \mu_1)^2}{14}} = 0.8737985$$

La resta de les mitjanes i desviacions tipus són:

$$\mu_2 = 3.06$$

$$\sigma_2 = 0.3180296$$

$$\mu_3 = 3.873333$$

$$\sigma_3 = 1.891887$$

$$\mu_4 = 1.246667$$

$$\sigma_4 = 0.8296873$$

Per tant, la taula 2, en normalitzar les seves dades, esdevé la taula 3.

Taula 3. Longitud i amplada del sèpal i longitud i amplada del pètal (dades normalitzades)

flor	long. sèpal	ampl. sèpal	long. pètal	ampl. pètal	tipus
1	-0.923172415	1.383518809	-1.307336408	-1.261519508	*setosa*
2	-1.152058138	-0.188661656	-1.307336408	-1.261519508	*setosa*
3	-1.38094386	0.44021053	-1.360193675	-1.261519508	*setosa*
4	-1.495386722	0.125774437	-1.254479141	-1.261519508	*setosa*
5	-1.037615276	1.697954901	-1.307336408	-1.261519508	*setosa*
6	1.251241951	0.44021053	0.436953409	0.18480859	*versicolor*
7	0.564584783	0.44021053	0.331238874	0.305335932	*versicolor*
8	1.13679909	0.125774437	0.542667943	0.305335932	*versicolor*
9	-0.46540097	-2.389714306	0.066952538	0.064281249	*versicolor*
10	0.679027644	-0.817533841	0.384096141	0.305335932	*versicolor*
11	0.450141921	0.754646623	1.124097882	1.510609348	*virginica*
12	-0.122072385	-1.131969934	0.648382477	0.787445298	*virginica*
13	1.365684812	-0.188661656	1.071240615	1.028499981	*virginica*
14	0.450141921	-0.503097749	0.912668813	0.666917957	*virginica*
15	0.679027644	-0.188661656	1.018383347	1.149027323	*virginica*

Font: Edgar Anderson (1935). "The irises of the Gaspe Peninsula"

Les dades de la taula 3 també es poden expressar en forma matricial:

$$\check{\mathbf{X}} = \begin{bmatrix} -0.923172415 & 1.383518809 & -1.307336408 & -1.261519508 \\ -1.152058138 & -0.188661656 & -1.307336408 & -1.261519508 \\ -1.38094386 & 0.44021053 & -1.360193675 & -1.261519508 \\ -1.495386722 & 0.125774437 & -1.254479141 & -1.261519508 \\ -1.037615276 & 1.697954901 & -1.307336408 & -1.261519508 \\ 1.251241951 & 0.44021053 & 0.436953409 & 0.18480859 \\ 0.564584783 & 0.44021053 & 0.331238874 & 0.305335932 \\ 1.13679909 & 0.125774437 & 0.542667943 & 0.305335932 \\ -0.46540097 & -2.389714306 & 0.066952538 & 0.064281249 \\ 0.679027644 & -0.817533841 & 0.384096141 & 0.305335932 \\ 0.450141921 & 0.754646623 & 1.124097882 & 1.510609348 \\ -0.122072385 & -1.131969934 & 0.648382477 & 0.787445298 \\ 1.365684812 & -0.188661656 & 1.071240615 & 1.028499981 \\ 0.450141921 & -0.503097749 & 0.912668813 & 0.666917957 \\ 0.679027644 & -0.188661656 & 1.018383347 & 1.149027323 \end{bmatrix}$$

Per simplificar la notació, tot i que $\check{\mathbf{X}}$ representa la matriu de dades normalitzada, continuarem parlant de la matriu $\mathbf{X}$ i entendrem, en la resta del mòdul, que està normalitzada.

Les dades de la taula 3 es poden representar gràficament de forma senzilla si generem diagrames de dispersió bidimensionals, per cada parell de variables. En aquest cas, això voldria dir que podem generar un total de

$$\binom{4}{2} = \frac{4 \cdot 3}{2} = 6$$

diagrames de dispersió. Aquesta quantitat sembla raonable. Però què passaria amb una mostra en què hem mesurat deu variables? En aquest cas, hauríem de representar gràficament $\binom{10}{2} = 45$ diagrames de dispersió, i ningú no ens podria garantir que algun d'aquests quaranta-cinc diagrames marqués alguna tendència o mostrés alguna particularitat.

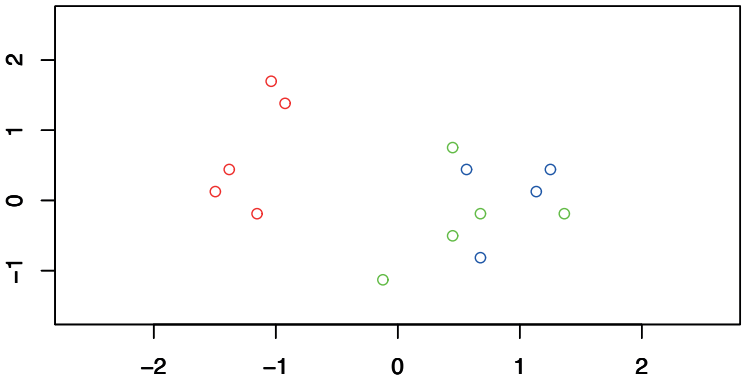
Diagrama de dispersió

Un diagrama de dispersió (en anglès, *scatter plot*) mostra gràficament la relació entre dues variables quantitatives.

Com a mostra d'aquests diagrames de dispersió, la figura 1 recull les variables de longitud del sèpal i amplitud del sèpal, per a les quinze flors. En el cas de la

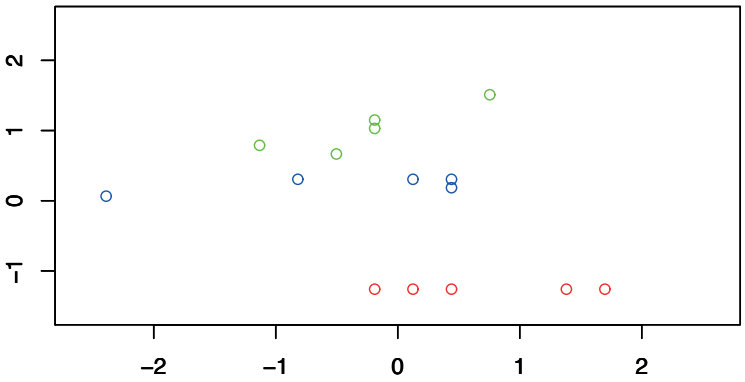
figura 2, es mostren les variables d'amplitud del sèpal i amplitud del pètal. Els colors representen el tipus o espècie de flor: *setosa* (vermell), *versicolor* (blau) i *virginica* (verd). En totes dues figures es pot veure com els punts blaus es confonen amb els punts verds, al mateix temps que els punts vermells queden agrupats de manera ben diferenciada.

Figura 1. Diagrama de dispersió de les variables relatives a la longitud del sèpal (eix horitzontal) i l'amplitud del sèpal (eix vertical).



Els colors representen el tipus o espècie de flor: *setosa* (vermell), *versicolor* (blau) i *virginica* (verd). Font: elaboració pròpia

Figura 2. Diagrama de dispersió de les variables relatives a l'amplitud del sèpal (eix horitzontal) i l'amplitud del pètal (eix vertical).



Els colors representen el tipus o espècie de flor: *setosa* (vermell), *versicolor* (blau) i *virginica* (verd). Font: elaboració pròpia

2.2. Matriu de covariàncies

Donada la matriu (normalitzada)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix} \in \mathcal{M}_{n\times m}(\mathbb{R})$$

$$= \begin{bmatrix} v_1 & v_2 & \cdots & v_j & \cdots & v_m \end{bmatrix}$$

la matriu de covariàncies es defineix així:

$$\mathbf{C_X} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = \frac{1}{n-1}\begin{bmatrix} v_1^T v_1 & v_1^T v_2 & \cdots & v_1^T v_j & \cdots & v_1^T v_m \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ v_j^T v_1 & v_j^T v_2 & \cdots & v_j^T v_j & \cdots & v_j^T v_m \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ v_m^T v_1 & v_m^T v_2 & \cdots & v_m^T v_j & \cdots & v_m^T v_m \end{bmatrix} \in \mathcal{M}_{m\times m}(\mathbb{R})$$

Fixeu-vos que la matriu de covariàncies és una matriu quadrada de tantes files i columnes com columnes té la matriu $\mathbf{X}$.

La matriu de covariàncies $\mathbf{C_X}$

$$\mathbf{C_X} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = \frac{1}{n-1}\begin{pmatrix} \boxed{v_1^T v_1} & v_1^T v_2 & \cdots & v_1^T v_j & \cdots & v_1^T v_m \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ v_j^T v_1 & v_j^T v_2 & \cdots & \boxed{v_j^T v_j} & \cdots & v_j^T v_m \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ v_m^T v_1 & v_m^T v_2 & \cdots & v_m^T v_j & \cdots & \boxed{v_m^T v_m} \end{pmatrix}$$

mesura el grau de relació lineal del conjunt de dades entre cada un dels parells de variables. Els termes de la diagonal principal corresponen a la variància de cadascuna de les variables:

$$\sigma_j^2 = \frac{1}{n-1}v_j^T v_j = \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}^2$$

Com que les nostres dades estan normalitzades, tots els termes de la diagonal principal són 1. El termes que no són de la diagonal principal representen la covariància entre cada parell de variables:

$$\sigma_{jk}^2 = \frac{1}{n-1}v_j^T v_k = \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}x_{ik}$$

En el nostre exemple, la matriu de covariàncies és

$$\mathbf{C_X} = \begin{bmatrix} 1.0000000 & -0.1609042 & 0.8854496 & 0.8251793 \\ -0.1609042 & 1.0000000 & -0.3817905 & -0.3578668 \\ 0.8854496 & -0.3817905 & 1.0000000 & 0.9860398 \\ 0.8251793 & -0.3578668 & 0.9860398 & 1.0000000 \end{bmatrix}$$

Es pot observar:

1) Els elements de la diagonal principal són tots igual a 1. Això és així perquè les nostres dades han estat normalitzades i, per tant, la variància de totes és 1.

2) La matriu de covariàncies és una matriu simètrica. Això succeeix perquè la covariància és simètrica, és a dir,

$$\sigma_{jk}^2 = \sigma_{kj}^2.$$

3) Les variables 3 i 4, corresponents a la longitud i amplitud del pètal, estan altament relacionades, ja que la seva covariància és 0.9860398 (molt propera a 1). Les variables 1 i 3 —longitud del sèpal i del pètal, respectivament— també estan significativament relacionades, tot i que en menor proporció, ja que la seva covariància és 0.8854496.

4) Contràriament, les variables 1 i 2 —longitud i amplitud del sèpal, respectivament— no estan gaire relacionades. En efecte, la seva covariància és –0.1609042.

2.3. Diagonalització de la matriu de covariàncies

L'objectiu de l'anàlisi de components principals és trobar una transformació (aplicació) lineal:

$$\mathbf{P} \in \mathcal{M}_{m\times m}(\mathbb{R})$$

tal que les dades originals recollides a $\mathbf{X}$ es transformin o es projectin en un nou espai mitjançant el producte:

$$\mathbf{T} = \mathbf{XP} \in \mathcal{M}_{n\times m}(\mathbb{R})$$

de manera que la matriu de covariàncies $\mathbf{C_T}$ de les noves dades $\mathbf{T}$ sigui diagonal.

Com que $\mathbf{C_X}$ és una matriu quadrada i simètrica de dimensió $m \times m$, sabem pel mòdul "Aplicacions lineals, diagonalització i vectors propis" que existeixen m valors propis λ_i reals i m vectors propis (ortonormals) p_i que formen una base a l'espai vectorial euclidià $\mathbb{R}^n$ tal que

Vectors ortonormals

Diem que dos vectors p_i i p_j són ortonormals si $p_i^T p_j = 0$ i, a més, $p_i^T p_i = 1$ i $p_j^T p_j = 1$.

$$\mathbf{C_X} = \mathbf{PDP}^T,$$

en què

$$\mathbf{P} = \left[\begin{array}{c|c|c|c} p_1 & p_2 & \cdots & p_m \end{array} \right]$$

$$\mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_m)$$

Sobre la matriu $\mathbf{P}$

Es pot demostrar fàcilment que la transposada de la matriu $\mathbf{P}$ és, alhora, la seva inversa. És a dir, $\mathbf{P}^T = \mathbf{P}^{-1}$.

Donada la matriu $\mathbf{X}$ que conté les dades originals (normalitzades), les dades de la nova matriu transformada $\mathbf{T}$ es calculen així:

$$\mathbf{T} = \mathbf{XP} \in \mathcal{M}_{n \times m}(\mathbb{R}),$$

en què $\mathbf{P}$ és la matriu on les columnes són els vectors propis de la matriu de covariàncies $\mathbf{C_X}$.

2.4. La matriu de covariàncies de les dades transformades

Quina és la matriu de covariàncies de les dades transformades? Calculem-la:

$$\mathbf{C_T} = \frac{1}{n-1}\mathbf{T}^T\mathbf{T} = \frac{1}{n-1}\mathbf{P}^T\mathbf{X}^T\mathbf{XP} = \mathbf{P}^T\mathbf{C_X}\mathbf{P}$$

$$= \mathbf{P}^T\mathbf{PDP}^T\mathbf{P} = \mathbf{D} = \text{diag}(\lambda_1, \ldots, \lambda_m)$$

Això significa que la matriu de covariàncies de les dades transformades és diagonal. Per tant, les noves variables estan incorrelacionades. És habitual ordenar els vectors propis en funció del valor propi associat, de major a menor. És a dir, si els valors propis són:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m,$$

els vectors propis els ubicarem a la matriu $\mathbf{P}$ en aquest ordre:

$$p_1, p_2, \ldots, p_m.$$

Tornant a l'exemple del conjunt de dades *Iris*, els valors propis i els vectors propis de la matriu de covariàncies $\mathbf{C_X}$ són:

$$\lambda_1 = 2.941490992, \quad p_1^T = \begin{bmatrix} 0.5260194 & -0.2616562 & 0.5786532 & 0.5656856 \end{bmatrix}$$

$$\lambda_2 = 0.891699528 \quad p_2^T = \begin{bmatrix} -0.31992346 & -0.94247968 & -0.06660176 & -0.07032238 \end{bmatrix}$$

$$\lambda_3 = 0.162361649 \quad p_3^T = \begin{bmatrix} 0.7635779 & -0.2005818 & -0.2275748 & -0.5700223 \end{bmatrix}$$

$$\lambda_4 = 0.004447831 \quad p_4^T = \begin{bmatrix} 0.1946827 & -0.0550912 & -0.7803425 & 0.5917171 \end{bmatrix}$$

El vector propi associat al valor propi més gran s'anomena *primera component principal*. El vector propi associat al segon valor propi més gran s'anomena *segona component principal*. I així successivament.

2.4.1. Com hem d'interpretar, per exemple, la primera component principal?

Denotem les quatre variables que hem considerat en l'exemple de les flors com a u_1, u_2, u_3 i u_4, en què

1) u_1 és la longitud del sèpal;

2) u_2 és l'amplada del sèpal;

3) u_3 és la longitud del pètal; i

4) u_4 és l'amplada del pètal.

Les components de la primera component principal p_1 indiquen que

$$p_1 = 0.5260194u_1 - 0.2616562u_2 + 0.5786532u_3 + 0.5656856u_4.$$

En altres paraules, p_1 representa una nova variable, que és combinació lineal de les quatre variables originals. Fixeu-vos que en la definició d'aquesta nova

variable, no totes les variables originals tenen el mateix pes. En efecte, la variable que té més pes és u_3 (longitud del pètal), seguida de la variable u_4 (amplada del pètal). Clarament, la que té menys influència en la primera component és la variable u_2 (amplada del sèpal).

Quin és el valor de la nova variable p_1 per al cas de la primera flor de la taula 3? Recordeu que les variables normalitzades en el cas de la primera flor són:

$$u_1 = -0.923772415$$

$$u_2 = 1.383518809$$

$$u_3 = -1.307336408$$

$$u_4 = -1.261519508$$

Per tant, per a la primera flor

$$p_1 = 0.5260194u_1 - 0.2616562u_2 + 0.5786532u_3 + 0.5656856u_4 = -2.3177306$$

Si fem el mateix amb les altres components principals, obtenim

$$p_2 = -0.8328099$$

$$p_3 = 0.03418822$$

$$p_4 = 0.017762021$$

En el cas general, les mesures en les noves variables s'obtenen de multiplicar les dades originals de la matriu $\mathbf{X}$ per la matriu $\mathbf{P}$ de les components principals:

$$\mathbf{T} = \mathbf{XP} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1j} & \cdots & t_{1m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ t_{i1} & t_{i2} & \cdots & t_{ij} & \cdots & t_{im} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nj} & \cdots & t_{nm} \end{bmatrix} \in \mathcal{M}_{n\times m}(\mathbb{R})$$

Fixeu-vos que les matrius $\mathbf{X}$ i $\mathbf{T}$ tenen la mateixa dimensió.

La matriu T per a l'exemple de les flors seria:

$$\mathbf{T} = \begin{bmatrix}
-2.31773055645763 & -0.832809881637843 & 0.0341882231504618 & 0.0177620212434633 \\
-2.02675817141067 & 0.722164171113012 & 0.174766954611649 & 0.0598152395437477 \\
-2.34229080476915 & 0.206211215432147 & -0.114116476558486 & 0.021856594229762 \\
-2.25904377241743 & 0.53213302544293 & -0.162490314930362 & -0.0655943343449898 \\
-2.46020386002701 & -1.09254655308811 & -0.116267982265906 & -0.0218406874991054 \\
0.900381739958164 & -0.857289184578846 & 0.662337916991262 & -0.0122754574068039 \\
0.546195294147313 & -0.639046440726892 & 0.0933763504900945 & 0.00785590242460743 \\
1.05180934990985 & -0.539843144155038 & 0.545260806182149 & -0.0284082915673973 \\
0.455578827059034 & 2.39217029459659 & 0.0720848822574889 & 0.0268371163600159 \\
0.966077049158865 & 0.506218727354301 & 0.4210140657984 & 0.0581798257612053 \\
1.54431815987357 & -1.03634661634201 & -0.924547274237649 & 0.0627325552563454 \\
1.05260949959829 & 1.00735403635774 & -0.462576054699356 & -0.00141932059541922 \\
1.96942554310489 & -0.402777918923283 & 0.250593622580613 & 0.0489153084361122 \\
1.27380657875815 & 0.222463831446303 & -0.143227795787201 & -0.202216367937914 \\
1.64582512351377 & -0.188055562291003 & -0.330396923583155 & 0.0277998960963719
\end{bmatrix}$$

2.4.2. Quin és el pes de la primera component principal?

Per les característiques de la matriu de covariàncies $\mathbf{C_X}$ —simètrica i definida positiva—, tots els valors propis són positius. A més, podem observar que en l'exemple de les flors

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 4,$$

és a dir,

$$\sum_{i=1}^{m} \lambda_i = \mathrm{tr}(\mathbf{C_X})$$

Traça d'una matriu

Recordeu que la traça d'una matriu quadrada, que denotem $\mathrm{tr}(A)$, és la suma dels elements de la diagonal principal.

Ja hem vist que els elements de la diagonal principal de la matriu de covariàncies $\mathbf{C_X}$ representen el valor de la variància de cadascuna de les variables originals. En el cas de les noves variables, els elements de la diagonal principal de la matriu de covariàncies $\mathbf{C_T}$ —que són els valors propis λ_i— també representen aquesta variància. Si el total de la variància, és a dir, la suma dels valors propis, és m, l'aportació de la primera component és

$$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}\times 100\%$$

Per al nostre exemple, la primera component principal és capaç de retenir un percentatge de variabilitat igual a:

$$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}\times 100\% = \frac{2.941490992}{4}\times 100\% = 73.53727480\%$$

De la mateixa manera, la resta de les components principals són capaces de retenir el percentatge de variabilitat següent:

$$\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}\times 100\% = \frac{0.891699528}{4}\times 100\% = 22.29248820\%$$

$$\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}\times 100\% = \frac{0.162361649}{4}\times 100\% = 4.059041225\%$$

$$\frac{\lambda_4}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}\times 100\% = \frac{0.004447831}{4}\times 100\% = 0.1111957750\%$$

Cadascuna de les quatre variables originals u_1,u_2,u_3 i u_4 reté un 25% de la variabilitat. En canvi, les noves variables p_1,p_2,p_3 i p_4 —les quatre components principals— retenen un 73.5%,22.3%,4.1% i 0.1%, respectivament. És a dir, calen tres variables originals per obtenir la mateixa quantitat d'informació que s'aconseguiria amb una única variable nova, la primera component principal.

Variabilitat

Hem d'entendre la variabilitat com la quantitat d'informació. Com més variabilitat hi ha, més informació tenim.

Taula 4. Variabilitat acumulada per les quatre variables originals i les quatre variables noves (components principals).

variables	variabilitat acumulada (variables originals)	variabilitat acumulada (components principals)
1	25%	73.5%
2	50%	95.8%
3	75%	99.9%
4	100%	100%

Font: elaboració pròpia

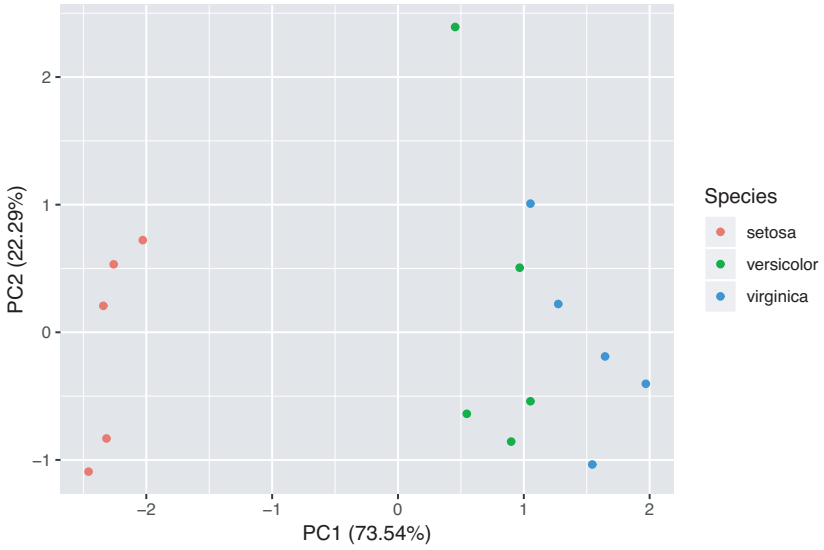
Observem ara les figures 3 i 4, que contenen informació interessant:

1) Totes dues figures contenen la projecció de les quinze flors sobre les dues primeres components principals. En aquest cas, el codi de colors és: *setosa* (vermell), *versicolor* (verd) i *virginica* (blau). El grup *setosa* (vermell) continua clarament diferenciat. Al mateix temps, la separació entre el grup de flors *versicolor* (verd) i *virginica* (blau) ara és més clara.

2) La figura 4 conté, a més, la contribució de cadascuna de les quatre variables originals a les dues primeres components principals. Si mirem, per exemple, la primera component principal (la direcció horitzontal), les variables que més intervenen són la longitud del sèpal (*Sepal.Length*) i la longitud i amplitud del pètal (*Petal.Length*, *Petal.Width*). En el cas de la segona component principal (la direcció vertical), la variable que té més pes és l'amplitud del sèpal (*Sepal.Width*).

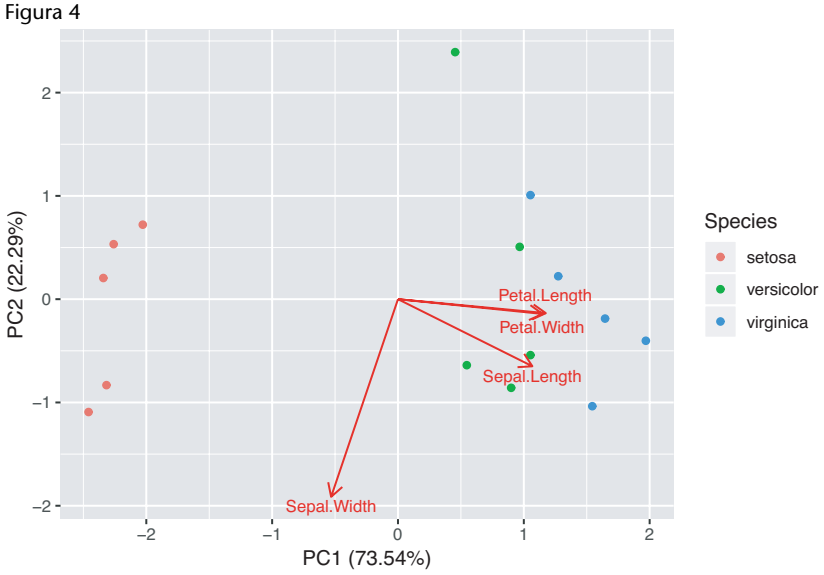
3) A la figura 4 també es pot veure com les fletxes que indiquen les direccions de les variables longitud i amplitud del pètal (*Petal.Length*, *Petal.Width*) estan pràcticament superposades. Recordem que, en aquest cas, la covariància entre aquestes dues variables és 0.9860398, que representa un valor molt proper a 1. És a dir, ja havíem dit que les variables 3 i 4 estan altament relacionades.

Figura 3



Font: elaboració pròpia

Figura 4



Font: elaboració pròpia

2.5. Reducció de la dimensió

A la taula 4 hem vist que, en l'exemple de les flors, amb dues components principals podem retenir el 95.8% de la variabilitat o la informació. Això significa que podem passar d'una mostra de quinze flors en què hem mesurat quatre variables diferents a una mostra de quinze flors en què només mesurem dues variables. Vegem-ho amb més detall en el cas general.

Si considerem totes les components principals, l'ortonormalitat de les components principals implica que

$$\mathbf{P}\mathbf{P}^T = \mathbf{I}_m,$$

en què $\mathbf{I}_m$ és la matriu identitat de dimensió m. Aleshores, la projecció:

$$\mathbf{T} = \mathbf{X}\mathbf{P}$$

es pot invertir per recuperar les dades originals, a partir de les dades projectades:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T,$$

ja que

$$\mathbf{T} = \mathbf{X}\mathbf{P} \Leftrightarrow \mathbf{T}\mathbf{P}^T = \mathbf{X}\mathbf{P}\mathbf{P}^T \Leftrightarrow \mathbf{T}\mathbf{P}^T = \mathbf{X}\mathbf{I}_m \Leftrightarrow \mathbf{T}\mathbf{P}^T = \mathbf{X}.$$

No obstant això, un dels objectius de l'anàlisi de components principals és la reducció de la dimensió. Per això, considerem ara un nombre inferior de components principals, $\ell < m$, és a dir, només considerem els vectors propis associats als ℓ valors propis més grans. Aleshores, si definim la matriu reduïda:

$$\hat{\mathbf{P}} = (p_1|p_2|\cdots|p_\ell) \in \mathcal{M}_{m\times\ell}(\mathbb{R})$$

la matriu de les projeccions es defineix així:

$$\hat{\mathbf{T}} = \mathbf{X}\hat{\mathbf{P}} \in \mathcal{M}_{n\times\ell}(\mathbb{R})$$

Sobre la dimensió de $\hat{\mathbf{T}}$

$\hat{\mathbf{T}}$ és una matriu que continua tenint tantes files com la matriu $\mathbf{X}$ original. De tota manera, el nombre de columnes de la matriu $\hat{\mathbf{T}}$ passa de tenir m columnes a tenir ℓ columnes.

2.5.1. L'error residual

Una de les conseqüències d'haver reduït la dimensionalitat és que la matriu $\hat{\mathbf{P}}$ ja no és invertible. Per tant, les dades originals contingudes a $\mathbf{X}$ no es poden recuperar completament mitjançant la matriu $\hat{\mathbf{T}}$. No obstant això, es pot invertir la projecció de la manera següent:

Nota

La matriu $\hat{\mathbf{P}}$ ja no és invertible. De fet, la matriu $\hat{\mathbf{P}}$ no és, ni tan sols, quadrada.

$$\hat{\mathbf{X}} = \hat{\mathbf{T}}\hat{\mathbf{P}}^T \in \mathcal{M}_{n\times m}(\mathbb{R})$$

per obtenir les dades originals amb pèrdua d'informació. La diferència entre les dades originals recollides a la matriu $\mathbf{X}$ i les dades originals amb pèrdua d'informació de la matriu $\hat{\mathbf{X}}$ s'anomena *error residual* i es representa amb la matriu $\mathbf{E}$. En efecte:

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} \in \mathcal{M}_{n\times m}(\mathbb{R})$$

2.6. Exemple d'aplicació: enquesta de pressupostos familiars

Continuem amb l'exemple del subapartat 1.2., en el qual es presenten les dades de l'enquesta de pressupostos familiars (EPF) per comunitats autònomes i les variables que es mesuren. Com que no hi ha informació de la variable *Ensenyament* per a Ceuta i Melilla, eliminem aquestes dues ciutats autònomes de l'anàlisi. Volem veure si, gràcies a l'anàlisi de components principals, podem extreure alguna conclusió o destacar algun patró que hagi quedat ocult en la quantitat de dades.

Dades completes

Les dades completes de l'enquesta de pressupostos familiars 2017 es poden obtenir en aquest enllaç de l'Institut Nacional d'Estadística: https://www.ine.es/jaxiT3/Tabla.htm?t=25143&L=0.

Amb l'ajuda del llenguatge de programació `R`, les dades han estat emmagatzemades a la matriu `INE`, que conté disset files —una per comunitat autònoma— i dotze columnes —una per variable. Procedim de la manera següent, tal com es pot veure a la figura 5:

R

`R` és un entorn de programació lliure especialitzat en estadística i representació gràfica.

1) Amb la instrucció `prcomp`, calculem l'anàlisi de components principals, que emmagatzemem a la variable `ine.pca`. És important afegir les opcions `centre = TRUE` i `scale = TRUE` per garantir que les dades han estat escalades i centrades.

2) La primera component principal s'obté amb `ine.pca$rotation[,1]`. Es pot observar que la primera component és una combinació lineal de les dotze variables originals. En particular:

$$p_1 = 0.30u_1 + 0.12u_2 + 0.28u_3 + 0.33u_4 + 0.33u_5 + 0.31u_6$$

$$+ 0.18u_7 + 0.25u_8 + 0.31u_9 + 0.30u_{10} + 0.33u_{11} + 0.33u_{12}$$

Figura 5. Instruccions de R per al càlcul de les dues primeres components principals, així com els valors de les variables originals projectats sobre aquestes dues primeres components.

```
> ine.pca <- prcomp(INE[,2:13],center = TRUE, scale=TRUE)
> ine.pca$rotation[,1]
       V1        V2        V3        V4        V5        V6        V7
0.3048175 0.1177153 0.2797855 0.3336785 0.3299060 0.3053362 0.1788030
       V8        V9       V10       V11       V12
0.2512488 0.3112978 0.3010705 0.3349005 0.3267037
> ine.pca$rotation[,2]
         V1          V2          V3          V4          V5
-0.01383321 -0.71634090 -0.36251483  0.18264741 -0.13857533
         V6          V7          V8          V9         V10
-0.07815930  0.08018175  0.44706142 -0.12810849  0.22574457
        V11         V12
-0.01588222  0.15052448
> ine.pca$x[,1]
 [1] -1.5284356  0.5230432  0.1936867  1.8834382 -4.5577019  1.4096488
 [7] -0.6333325 -4.0799862  2.4422616 -0.8963849 -4.9655611 -0.8830605
[13]  3.7092028 -0.8849228  3.6507548  4.3414574  0.2758919
> ine.pca$x[,2]
 [1] -1.38763977 -0.42984445 -0.20352852 -0.01118605  2.85476875
 [6] -0.01202456 -0.13105989 -0.46300999  1.14629871 -0.07641837
[11]  0.41923006 -1.32859163  0.47571073 -1.71830933 -0.45566480
[16]  1.29743154  0.02383759
```

Font: elaboració pròpia

La primera component principal representa una nova variable on totes les variables originals *sumen* en més o menys proporció. En particular, les variables que tenen més pes són: la u_4 (habitatge, aigua, electricitat, gas i altres combustibles), la u_5 (mobles, articles de la llar i articles per al manteniment de la llar), la u_{11} (restaurants i hotels) i la u_{12} (altres béns i serveis). És especialment rellevant el poc pes que tenen les variables u_2 (begudes alcohòliques i tabac) i u_7 (transport). La variable u_2 serà, en canvi, la més important en la segona component principal, com es veurà a continuació. La variable u_7 continuarà tenint poca influència. Com es pot veure a la figura 6, la informació o variabilitat explicada per la primera component principal és del 65.1%.

3) Si projectem les dades originals de les disset comunitats autònomes sobre la primera component principal, que podem fer amb l'ordre `ine.pca$x[,1]`, obtindrem els resultats que es poden veure a la figura 5. Si ordenem de grans a petits aquests valors, obtindrem una primera ordenació interessant, que es pot veure a la taula 5. Podríem dir que la primera component principal ha estat capaç d'ordenar les comunitats autònomes en funció de la renda.

4) La segona component principal s'obté amb `ine.pca$rotation[,2]`. Es pot observar que la segona component és també una combinació lineal de les dotze variables originals. En particular:

$$p_2 = -0.01u_1 - 0.72u_2 - 0.36u_3 + 0.18u_4 - 0.14u_5 - 0.08u_6$$

$$+ 0.08u_7 + 0.45u_8 - 0.13u_9 + 0.23u_{10} - 0.02u_{11} + 0.15u_{12}$$

Taula 5. Projecció de les dades originals sobre la primera component principal (de gran a petita)

Comunitat autònoma	PC1
País Basc	4.3414574
Comunitat de Madrid	3.7092028
Navarra	3.6507548
Catalunya	2.4422616
Balears	1.8834382
Cantàbria	1.4096488
Aragó	0.5230432
La Rioja	0.2758919
Astúries	0.1936867
Castella i Lleó	–0.6333325
Galícia	–0.8830605
Múrcia	–0.8849228
Comunitat Valenciana	–0.8963849
Andalusia	–1.5284356
Castella-la Manxa	–4.0799862
Canàries	–4.5577019
Extremadura	–4.9655611

Font: Enquesta de pressupostos familiars 2017 (Institut Nacional d'Estadística)

En aquest cas, algunes de les variables originals sumen, mentre que d'altres resten, en la seva contribució a la nova variable que representa la segona component principal. Les variables amb més pes (en valor absolut) són: u_2 (begudes alcohòliques i tabac), u_8 (comunicacions) i u_3 (vestit i calçat). També en aquesta segona component principal, el pes de la variable u_7 (transport) és molt petit. Això significa que la variable que mesura les despeses en transport no marca una diferència entre les comunitats autònomes (podríem obtenir un resultat diferent si estudiéssim l'enquesta de pressupostos familiars per províncies i no per comunitats autònomes). Com es pot veure a la figura 6, la informació o variabilitat explicada per la segona component principal és del 9.83%. Per tant, la variabilitat explicada per les dues primeres components principals és del 74.93%.

5) Si projectem les dades originals de les disset comunitats autònomes sobre la segona component principal, que podem fer amb l'ordre `ine.pca$x[,2]`, aconseguirem els resultats que es poden veure a la figura 5. Si ordenem de grans a petits aquests valors, obtindrem una segona ordenació interessant, que es pot veure a la taula 6. En aquest cas, però, és més difícil explicar de manera qualitativa quina és la interpretació d'aquesta segona variable.

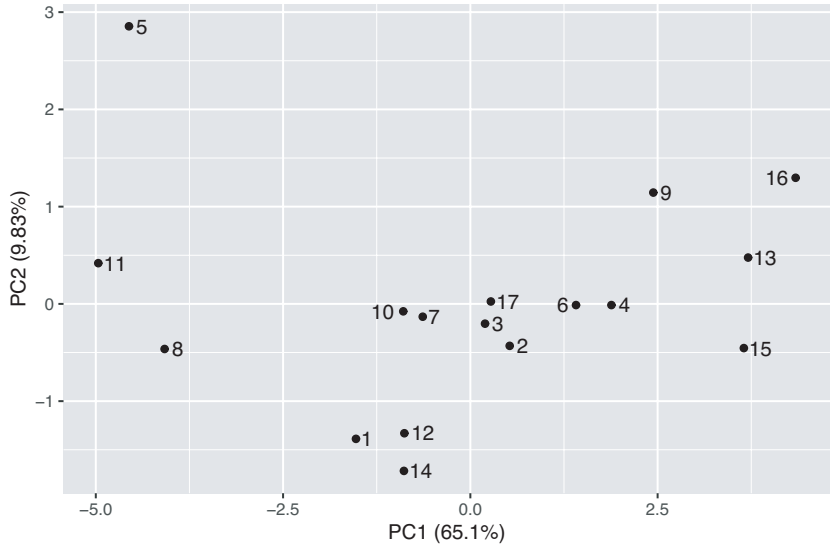
6) A la figura 6 es pot veure la projecció de les dades originals de les disset comunitats autònomes sobre les dues primeres components principals, que són capaces de retenir quasi el 75% de la informació.

Taula 6. Projecció de les dades originals sobre la segona component principal (de gran a petita)

Comunitat autònoma	PC2
Canàries	2.85476875
País Basc	1.29743154
Catalunya	1.14629871
Comunitat de Madrid	0.47571073
Extremadura	0.41923006
La Rioja	0.02383759
Balears	–0.01118605
Cantàbria	–0.01202456
Comunitat Valenciana	–0.07641837
Castella i Lleó	–0.13105989
Astúries	–0.20352852
Aragó	–0.42984445
Navarra	–0.45566480
Castella-la Manxa	–0.46300999
Galícia	–1.32859163
Andalusia	–1.38763977
Múrcia	–1.71830933

Font: Enquesta de pressupostos familiars 2017 (Institut Nacional d'Estadística)

Figura 6. Projecció de les dades originals de les disset comunitats autònomes sobre les dues primeres components principals, que són capaces de retenir quasi el 75% de la informació.



Font: elaboració pròpia

3. Descomposició en valors singulars

3.1. Exemple introductori

Píxel

Segons el *Diccionari de la llengua catalana*, és la unitat mínima d'informació en què resulta dividida una imatge en sotmetre-la a un escombratge electrònic.

Les matrius, com a element bàsic de l'àlgebra lineal, poden ser utilitzades en un ampli espectre d'aplicacions, des de les més senzilles (sistema d'equacions lineals) fins a les més complexes (sistemes dinàmics discrets, transició d'estats). Una d'aquestes aplicacions és emprar una matriu per emmagatzemar els píxels d'una imatge. Cada element de la matriu representa un píxel. Si la imatge és en escala de grisos, els valors que pot adoptar cada píxel estarien en el rang [0,1].

Considerem, per exemple, una imatge de 6×6 píxels emmagatzemada en forma de matriu:

$$\mathbf{A} = \begin{bmatrix} 0.1 & 0.1 & 0.5 & 0.5 & 0.9 & 0.9 \\ 0.1 & 0.1 & 0.5 & 0.5 & 0.9 & 0.9 \\ 0.1 & 0.1 & 0.5 & 0.5 & 0.9 & 0.9 \\ 0.1 & 0.1 & 0.5 & 0.5 & 0.9 & 0.9 \\ 0.1 & 0.1 & 0.5 & 0.5 & 0.9 & 0.9 \\ 0.1 & 0.1 & 0.5 & 0.5 & 0.9 & 0.9 \end{bmatrix}$$

A partir d'aquesta matriu podem fer les observacions següents:

1) La matriu $\mathbf{A}$ té rang 1, ja que totes les files són iguals.

2) La matriu $\mathbf{A}$ està formada per 36 elements, que és el resultat de multiplicar el nombre de files (6) pel nombre de columnes (també 6).

Es pot veure fàcilment que la matriu $\mathbf{A}$ es pot expressar com el producte d'un vector columna ($\mathbf{u}$) per un vector fila ($\mathbf{v}^T$) si considerem que:

$$\mathbf{A} = \mathbf{u}\mathbf{v}^T = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0.1 & 0.1 & 0.5 & 0.5 & 0.9 & 0.9 \end{bmatrix}$$

Quina és la importància de la descomposició de la matriu $\mathbf{A}$ actual com a producte d'aquests dos vectors? D'entrada, que per transmetre la informació continguda en aquests dos vectors només necessitem:

$$6 \times 2 = 12$$

nombres, en lloc dels 36 necessaris per transmetre tota la matriu.

3.2. Descomposició en valors singulars reduïda

A partir d'una matriu $\mathbf{A} \in \mathcal{M}_{m\times n}(\mathbb{R})$:

$\mathcal{M}_{m\times n}(\mathbb{R})$

Recordeu que denotem mitjançant $\mathcal{M}_{m\times n}(\mathbb{R})$ l'espai vectorial de les matrius de m files i n columnes a coeficients reals.

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

la descomposició en valors singulars (SVD) **reduïda** de $\mathbf{A}$ és una factorització (o descomposició)

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

en què

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} u_{11} & u_{21} & \cdots & u_{n1} \\ u_{12} & u_{22} & \cdots & u_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1m} & u_{2m} & \cdots & u_{nm} \end{bmatrix} \in \mathcal{M}_{m\times n}(\mathbb{R}) \text{ és una matriu ortogonal,}$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix} = \begin{bmatrix} v_{11} & v_{21} & \cdots & v_{n1} \\ v_{12} & v_{22} & \cdots & v_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1n} & v_{2n} & \cdots & v_{nn} \end{bmatrix} \in \mathcal{M}_{n\times n}(\mathbb{R}) \text{ és una matriu ortogonal i}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix} \in \mathcal{M}_{n\times n}(\mathbb{R}) \text{ és una matriu diagonal.}$$

Els valors $\sigma_i \geq 0$, $i = 1, \ldots, n$ s'anomenen **valors singulars** i en general s'enumeren de manera descendent:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

Els vectors $\mathbf{u}_i$, $i = 1, \ldots, n$ s'anomenen **vectors singulars per l'esquerra** i, de manera similar, els vectors $\mathbf{v}_i$, $i = 1, \ldots, n$ s'anomenen **vectors singulars per la dreta**.

El fet que tant la matriu $\mathbf{U}$ com la matriu $\mathbf{V}$ siguin ortogonals significa que

$$\mathbf{u}_i^T \mathbf{u}_j = 0, \; i \neq j,$$

$$\mathbf{u}_i^T \mathbf{u}_i = 1,$$

$$\mathbf{v}_i^T \mathbf{v}_j = 0, \; i \neq j,$$

$$\mathbf{v}_i^T \mathbf{v}_i = 1.$$

3.2.1. Càlcul dels valors singulars i dels vectors singulars

Suposem que la matriu $\mathbf{A}$ té rang r. Per calcular-ne els valors singulars i els vectors singulars, farem el següent:

1) Els valors singulars no nuls σ_i, amb $i \leq r$, són l'arrel quadrada dels valors propis no nuls de les matrius $\mathbf{A}^T\mathbf{A}$ i $\mathbf{A}\mathbf{A}^T$.

2) Els vectors singulars per la dreta de la matriu $\mathbf{V}$ són els vectors propis de la matriu $\mathbf{A}^T\mathbf{A}$ (base ortonormal).

3) Els vectors singulars per l'esquerra de la matriu $\mathbf{U}$ són els vectors propis de la matriu $\mathbf{A}\mathbf{A}^T$ (base ortonormal).

4) Fent servir aquestes dues bases, $\mathbf{A}$ esdevé una matriu diagonal $\mathbf{\Sigma}$ i s'obté:

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i,$$

en què σ_i són els valors singulars.

Els vectors singulars són únics, llevat de signe. Per tant, es proposa la seqüència següent per tal de trobar-ne els vectors propis per la dreta i per l'esquerra:

1) Càlcul dels valors singulars, que són les arrels quadrades dels valors propis de la matriu $\mathbf{A}^T\mathbf{A}$ —i $\mathbf{A}\mathbf{A}^T$.

2) Càlcul dels vectors singulars per la dreta, $\mathbf{v}_i$, que són els vectors propis de la matriu $\mathbf{A}^T\mathbf{A}$.

3) Càlcul dels vectors singulars per l'esquerra, $\mathbf{u}_i$, tenint en compte la relació

$$\mathbf{u}_i = \mathbf{A}\mathbf{v}_i/\sigma_i,$$

de manera que assegurem que els signes estan escollits correctament.

Exemple

Considereu la matriu

$$\mathbf{A} = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}.$$

Les matrius $\mathbf{A}^T\mathbf{A}$ i $\mathbf{A}\mathbf{A}^T$ són:

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$$

$$\mathbf{A}\mathbf{A}^T = \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}$$

Totes dues matrius tenen els mateixos valors propis. Com que $\mathbf{A}\mathbf{A}^T$ ja és una matriu diagonal, els valors propis són $\lambda_1 = 8$ i $\lambda_2 = 2$. Per tant, els valors singulars són:

$$\sigma_1 = \sqrt{\lambda_1} = \sqrt{8} = 2\sqrt{2}$$

$$\sigma_2 = \sqrt{\lambda_2} = \sqrt{2}$$

Per trobar els vectors propis de la matriu $\mathbf{A}\mathbf{A}^T$, tenint en compte que els valors propis són $\lambda_1 = 8$ i $\lambda_2 = 2$, cal resoldre aquestes equacions:

$$\mathbf{A}\mathbf{A}^T\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

$$\mathbf{A}\mathbf{A}^T\mathbf{u}_2 = \lambda_2\mathbf{u}_2$$

En el cas de la primera, tenim el següent:

$$\mathbf{A}\mathbf{A}^T\mathbf{u}_1 = \lambda_1\mathbf{u}_1 \Leftrightarrow \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} = 8\begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} \Leftrightarrow \begin{cases} u_{11} = \kappa, \\ -6u_{12} = 0 \end{cases}$$

El sistema anterior és un sistema compatible indeterminat, en què κ és un paràmetre i $u_{12} = 0$. Atès que volem imposar que el vector $\mathbf{u}_1$ tingui norma 1, proposem, per exemple, $u_{11} = 1$. Per tant:

$$\mathbf{u}_1 = \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Procedirem d'una manera similar en el cas de la segona equació associada al valor propi $\lambda_2 = 2$; el vector propi $\mathbf{u}_2$ que obtenim és:

$$\mathbf{u}_2 = \begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Per trobar els vectors propis de la matriu $\mathbf{A}^T\mathbf{A}$, tenint en compte que els valors propis són $\lambda_1 = 8$ i $\lambda_2 = 2$, cal resoldre aquestes equacions:

$$\mathbf{A}^T\mathbf{A}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$$

$$\mathbf{A}^T\mathbf{A}\mathbf{v}_2 = \lambda_2\mathbf{v}_2$$

En el cas de la primera, tenim el següent:

$$\mathbf{A}^T\mathbf{A}\mathbf{v}_1 = \lambda_1\mathbf{v}_1 \Leftrightarrow \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 8\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} \Leftrightarrow \begin{cases} -3v_{11} + 3v_{12} = 0, \\ 3v_{11} - 3v_{12} = 0 \end{cases}$$

El sistema anterior és un sistema compatible indeterminat, amb solució paramètrica $v_{11} = v_{12}$. Atès que volem imposar que el vector $\mathbf{v}_1$ tingui norma 1, proposem, per exemple, $v_{11} = v_{12} = 1/\sqrt{2}$. Per tant:

$$\mathbf{v}_1 = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Procedirem d'una manera similar en el cas de la segona equació associada al valor propi $\lambda_2 = 2$; el vector propi $\mathbf{v}_2$ que obtenim és:

$$\mathbf{v}_2 = \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Finalment, fixeu-vos que:

$$\mathbf{A} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T$$

$$= \begin{bmatrix} 2 & 2 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -1 & 1 \end{bmatrix}$$

Es pot observar, a més, que, a partir d'una matriu $\mathbf{A} \in \mathcal{M}_{m\times n}(\mathbb{R})$ de rang r:

1) Els vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ formen una base ortonormal del subespai generat per les **files** de la matriu $\mathbf{A}$.

2) Els vectors $\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \ldots, \mathbf{v}_n$ formen una base ortonormal del subespai generat pel **nucli** de la matriu $\mathbf{A}$.

3) Els vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ formen una base ortonormal del subespai generat per les **columnes** de la matriu $\mathbf{A}$.

4) Els vectors $\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_n$ formen una base ortonormal del subespai generat pel **nucli** de la matriu $\mathbf{A}^T$.

El comentari anterior s'entendrà millor si considerem l'exemple següent:

Exemple il·lustratiu

Aquest exemple il·lustra una manera de calcular els vectors singulars per l'esquerra i per la dreta sense haver de calcular vectors propis.

Exemple

Volem calcular la descomposició en valors singulars de la matriu:

$$\mathbf{A} = \begin{bmatrix} 4 & 3 \\ 8 & 6 \end{bmatrix}.$$

La primera observació que podem fer és que el rang de la matriu és $r = 1$, ja que el determinant de la matriu és zero. Per tant, la dimensió del subespai generat per les files de la matriu $\mathbf{A}$ és igual que $r = 1$. Així doncs, una base d'aquest subespai estarà formada per un únic vector, que serà el primer vector singular per la dreta $\mathbf{v}_1^T$ i serà igual que el vector de la primera fila de la matriu, si el normalitzem. En efecte:

$$\mathbf{v}_1^T = [4,3]/\sqrt{4^2 + 3^2} = [4,3]/5 = [4/5, 3/5]$$

Atès que el nombre de columnes és $n = 2$, la dimensió del subespai generat pel nucli és $n - r = 2 - 1 = 1$. Així doncs, qualsevol vector ortogonal a $\mathbf{v}_1$ serà la base d'aquest subespai. Així doncs, n'hi ha prou de considerar

$$\mathbf{v}_2^T = [-u_{12}, u_{11}] = [-3/5, 4/5],$$

ja que

$$\mathbf{v}_1^T \mathbf{v}_2 = 0.$$

A més, sabem que

$$\mathbf{A}\mathbf{v}_1 = \sigma_1 \mathbf{u}_1.$$

Per tant:

$$\mathbf{A}\mathbf{v}_1 = \begin{bmatrix} 4 & 3 \\ 8 & 6 \end{bmatrix} \begin{bmatrix} 4/5 \\ 3/5 \end{bmatrix} = \begin{bmatrix} 5 \\ 10 \end{bmatrix} = \sqrt{125} \begin{bmatrix} 5/\sqrt{125} \\ 10/\sqrt{125} \end{bmatrix} = 5\sqrt{5} \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$$

Així doncs, $\sigma_1 = 5\sqrt{5}$ i $\mathbf{u}_1^T = [1/\sqrt{5}, 2/\sqrt{5}]$. Addicionalment,

$$\mathbf{u}_2^T = [-2/\sqrt{5}, 1/\sqrt{5}].$$

En aquest cas, també resulta fàcil comprovar que

$$\mathbf{A} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T$$

És fàcil comprovar que $\mathbf{U}$ i $\mathbf{V}$ són els vectors propis de valor propi σ_1^2 i $\sigma_2^2 = 0$ de les matrius $\mathbf{A}\mathbf{A}^T$ i $\mathbf{A}^T\mathbf{A}$, respectivament.

3.3. Descomposició en valors singulars completa

A partir d'una matriu $\mathbf{A} \in \mathcal{M}_{m\times n}(\mathbb{R})$:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

SVD completa

La diferència entre l'SVD reduïda i l'SVD completa és que la base ortonormal $\mathbf{U}$ formada per n elements en el cas de l'SVD reduïda s'amplia a una base ortonormal completa de $\mathbb{R}^n$. En fer-ho, la matriu $\mathbf{\Sigma}$ s'amplia a una matriu rectangular $m \times n$ afegint zeros.

la descomposició en valors singulars (SVD) **completa** de $\mathbf{A}$ és una factorització (o descomposició)

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

en què

$$\mathbf{U} = \left[\begin{array}{cccc} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_m \end{array}\right]$$

$$= \left[\begin{array}{cccc|ccc} u_{11} & u_{21} & \cdots & u_{n1} & u_{n+1,1} & \cdots & u_{m1} \\ u_{12} & u_{22} & \cdots & u_{n2} & u_{n+1,2} & \cdots & u_{m2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ u_{1m} & u_{2m} & \cdots & u_{nm} & u_{n+1,m} & \cdots & u_{mm} \end{array}\right] \in \mathcal{M}_{m\times m}(\mathbb{R})$$ és una matriu ortogonal quadrada d'ordre m,

$$\mathbf{V} = \left[\begin{array}{cccc} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{array}\right]$$

$$= \left[\begin{array}{cccc} v_{11} & v_{21} & \cdots & v_{n1} \\ v_{12} & v_{22} & \cdots & v_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1n} & v_{2n} & \cdots & v_{nn} \end{array}\right] \in \mathcal{M}_{n\times n}(\mathbb{R})$$ és una matriu ortogonal quadrada d'ordre n i

$$\mathbf{\Sigma} = \left[\begin{array}{cccc} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \\ \hline 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{array}\right] \in \mathcal{M}_{m\times n}(\mathbb{R})$$ és una matriu diagonal.

Els valors $\sigma_i \geq 0$, $i = 1,\ldots,n$ s'anomenen **valors singulars** i en general s'enumeren de manera descendent:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

Els vectors $\mathbf{u}_i$, $i = 1,\ldots,m$, com abans, s'anomenen **vectors singulars per l'esquerra** i, de manera similar, els vectors $\mathbf{v}_i$, $i = 1,\ldots,n$ s'anomenen **vectors singulars per la dreta**.

Considerem ara un exemple en què la matriu $\mathbf{A}$ ja no serà quadrada, com en els dos exemples anteriors.

Exemple

Volem calcular la descomposició en valors singulars (SVD) completa de la matriu

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 2 \\ 2 & 1 \end{bmatrix}$$

Fixeu-vos que el nombre de files de la matriu és $m = 3$ i que el nombre de columnes és $n = 2$. El rang màxim de la matriu és, per tant, 2, ja que és el mínim entre aquestes dues quantitats:

$$\text{mín}(m,n) = \text{mín}(3,2) = 2.$$

El rang de la matriu és $r = 2$, ja que la submatriu quadrada 2×2

$$\begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$$

té determinant no nul. Per al càlcul dels vectors singulars per la dreta considerem la matriu $\mathbf{A}^T A$. En efecte:

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 9 & 8 \\ 8 & 9 \end{bmatrix}$$

Els valors propis de la matriu $\mathbf{A}^T\mathbf{A}$ són $\lambda_1 = 17$ i $\lambda_2 = 1$, de manera que els valors singulars són:

$$\sigma_1 = \sqrt{\lambda_1} = \sqrt{17}$$

$$\sigma_2 = \sqrt{\lambda_2} = 1$$

Els vectors propis de la matriu $\mathbf{A}^T\mathbf{A}$ han de satisfer:

$$\mathbf{A}^T\mathbf{A}\mathbf{v}_1 = 17\mathbf{v}_1$$

$$\mathbf{A}^T\mathbf{A}\mathbf{v}_2 = \mathbf{v}_2$$

Si resolem les equacions corresponents, obtindrem:

$$\mathbf{v}_1^T = [1/\sqrt{2}, 1/\sqrt{2}]$$

$$\mathbf{v}_2^T = [1/\sqrt{2}, -1/\sqrt{2}]$$

Podem trobar els vectors singulars per la dreta tenint en compte que:

$$\mathbf{A}\mathbf{v}_1 = \sigma_1\mathbf{u}_1 = \sqrt{17}\mathbf{u}_1$$

$$\mathbf{A}\mathbf{v}_2 = \sigma_2\mathbf{u}_2 = \mathbf{u}_2$$

En el primer cas,

$$\mathbf{A}\mathbf{v}_1 = \begin{bmatrix} 3/\sqrt{2} \\ 4/\sqrt{2} \\ 3/\sqrt{2} \end{bmatrix} = \sqrt{17}\begin{bmatrix} 3/\sqrt{34} \\ 4/\sqrt{34} \\ 3/\sqrt{34} \end{bmatrix}$$

i n'obtenim

$$\mathbf{u}_1^T = [3/\sqrt{34},4/\sqrt{34},3/\sqrt{34}]$$

El vector $\mathbf{u}_2$ l'obtenim de manera similar i en resulta:

$$\mathbf{u}_2^T = [-1/\sqrt{2},0,1/\sqrt{2}]$$

Addicionalment, per a la descomposició en valors singulars completa, necessitem un tercer vector,

$$\mathbf{u}_3^T = [u_{31},u_{32},u_{33}],$$

que sigui ortogonal a $\mathbf{u}_1$ i $\mathbf{u}_2$. Aleshores, resulta un sistema d'equacions compatible indeterminat:

$$3u_{31} + 4u_{32} + 3u_{33} = 0$$

$$-u_{31} + u_{33} = 0$$

que té solució paramètrica

$$u_{31} = \kappa$$

$$u_{32} = -3\kappa/2$$

$$u_{33} = \kappa$$

Per exemple, podem considerar

$$\mathbf{u}_3^T = [2/\sqrt{17}, -3/\sqrt{17},2/\sqrt{17}].$$

A continuació podem mostrar tant la descomposició en valors singulars completa com la reduïda de la matriu $\mathbf{A}$:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

$$= \begin{bmatrix} 3/\sqrt{34} & -1/\sqrt{2} \\ 4/\sqrt{34} & 0 \\ 3/\sqrt{34} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \sqrt{17} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}^T$$

$$= \left[\begin{array}{cc|c} 3/\sqrt{34} & -1/\sqrt{2} & 2/\sqrt{17} \\ 4/\sqrt{34} & 0 & -3/\sqrt{17} \\ 3/\sqrt{34} & 1/\sqrt{2} & 2/\sqrt{17} \end{array}\right] \left[\begin{array}{cc} \sqrt{17} & 0 \\ 0 & 1 \\ \hline 0 & 0 \end{array}\right] \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}^T$$

3.3.1. Propietats interessants de la descomposició en valors singulars

1) Els valors singulars σ_i són únics.

2) El rang de la matriu $\mathbf{A}$ és el nombre de valors singulars no nuls.

3) $\|\mathbf{A}\|_2 = \sigma_1$ i $\|\mathbf{A}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_n^2}$, en què

$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}$$

4) Si $\mathbf{A}$ és una matriu simètrica $\mathbf{A} = \mathbf{A}^T$, aleshores els valors singulars de la matriu $\mathbf{A}$ són el valor absolut dels valors propis de la matriu $\mathbf{A}$. En particular, si $\mathbf{A}$ és una matriu simètrica i definida positiva, els valors propis i els valors singulars coincideixen.

5) Si $\mathbf{A}$ és una matriu quadrada, aleshores el valor absolut del determinant de la matriu coincideix amb el producte dels valors singulars, és a dir, $|\det(A)| = \prod_{i=1}^{n} \sigma_i$.

6) Si $\mathbf{A}$ és una matriu quadrada i tots els valors singulars σ_i són diferents, aleshores els vectors singulars per la dreta i per l'esquerra $\mathbf{u}_i$ i $\mathbf{v}_i$ són únics, llevat de signe.

3.4. Aplicació de la descomposició en valors singulars: compressió d'imatges

La compressió d'imatges que permet fer la descomposició en valors singulars es basa en el resultat següent:

Tota matriu $\mathbf{A}$ de dimensió $m \times n$ i rang r ($r < \min\{m,n\}$) es pot descompondre com a suma de r matrius de rang 1:

$$\mathbf{A} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T,$$

en què σ_j són els valors singulars, $\mathbf{u}_j$ són els vectors singulars per l'esquerra i $\mathbf{v}_j$ són els vectors singulars per la dreta, $j = 1,\ldots,r$, en la descomposició en valors singulars (SVD).

La matriu $\mathbf{A}$ representarà una imatge en escala de grisos rectangular de $m \times n$ píxels. En general, la imatge contindrà píxels que estaran relacionats entre si, és a dir, els píxels propers tenen alta correlació. En un cas extrem, per exemple, si tenim la fotografia d'un camp de futbol, molts píxels correspondran a la zona de la gespa amb tonalitats de verd molt similars. Per tant, en general, el rang de la matriu r pot ser significativament inferior al mínim del nombre de files o columnes de la matriu.

Ara bé, la compressió en la imatge prové del resultat següent:

Norma 2 d'una matriu

La norma 2 d'una matriu correspon a $\|\mathbf{A}\|_2 = \sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$ i es defineix, en certa manera, com es defineix la norma 2 d'un vector.

$\mathbf{A} = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T$ és una matriu de dimensió $m \times n$ i rang r ($r < \min\{m,n\}$). Considerem la v-èssima suma parcial:

$$\mathbf{A}_v = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_v \mathbf{u}_v \mathbf{v}_v^T = \sum_{j=1}^{v} \sigma_j \mathbf{u}_j \mathbf{v}_j^T$$

Aleshores, la diferència en norma 2 entre aquestes dues matrius és igual a:

$$\|\mathbf{A} - \mathbf{A}_v\|_2 = \sigma_{v+1}.$$

Sabem que en la descomposició en valors singulars, els valors singulars:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$$

estan ordenats de manera decreixent. Per tant, si observem que $\sigma_5 = 10^{-2}$, caldrà considerar la quarta suma parcial:

$$\mathbf{A}_4 = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T + \sigma_3\mathbf{u}_3\mathbf{v}_3^T + \sigma_4\mathbf{u}_4\mathbf{v}_4^T$$

per garantir que la matriu original i la quarta suma parcial tenen una diferència, en norma 2, igual a $\sigma_5 = 10^{-2}$, que és un error petit.

Per tant, la compressió d'imatges es basarà en aquesta estratègia:

Considerem que $\mathbf{A}$ és la matriu que representa una imatge rectangular, en escala de grisos, de $m \times n$ píxels i que $\sum_{j=1}^{r} \sigma_j\mathbf{u}_j\mathbf{v}_j^T$ és la descomposició de la imatge emmagatzemada a la matriu $\mathbf{A}$ com a suma de r matrius de rang 1. Aleshores, la v-èssima suma parcial

$$\mathbf{A}_v = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T + \cdots + \sigma_v\mathbf{u}_v\mathbf{v}_v^T = \sum_{j=1}^{v} \sigma_j\mathbf{u}_j\mathbf{v}_j^T$$

emmagatzema la imatge comprimida amb un error en norma 2 igual a σ_{v+1}.

Exemple

En aquest exemple tindrem en compte una fotografia del Parc Nacional de Yosemite (Califòrnia, Estats Units d'Amèrica), tal com es pot veure a la figura 7 en escala RGB i de mida 750×1.000 píxels. La imatge es pot representar com una matriu de 750 files i 1.000 columnes per a cada color: vermell (R, *red*), verd (G, *green*) i blau (B, *blue*). Com que la fotografia és en color, aplicarem la descomposició en valors singulars a cadascun dels colors per tornar a formar la imatge.

Figura 7. Imatge del Parc Nacional de Yosemite (en escala RGB i 750×1.000 píxels)

Font: elaboració pròpia

Farem servir `R` per calcular la descomposició en valors singulars i les imatges comprimides. Primer carregarem la llibreria `jpeg` i, després, llegirem la imatge `YOSEMITE.jpg`. La variable `yosemite` és una matriu tridimensional, en què per a cada color tenim 750 files i 1.000 columnes. Podem comprovar la dimensió de les matrius amb les instruccions `nrow(yosemite)` i `ncol(yosemite)`.

```
library(jpeg)
yosemite<-readJPEG('YOSEMITEC.jpg')
nrow(yosemite)
ncol(yosemite)
```

A continuació generem les tres matrius a què aplicarem la descomposició en valors singulars per separat. La matriu `r` contindrà el color vermell; la matriu `g`, el verd, i la matriu `b`, el blau. Fixeu-vos que hem accedit a la dimensió del color mitjançant la instrucció `yosemite[„1]` (per al color vermell), `yosemite[„2]` (per al color verd) o `yosemite[„3]` per al color blau. Com que `yosemite` era una matriu de dimensió $750 \times 1.000 \times 3$, hem accedit a la tercera component de la matriu esmentada.

```
r <- yosemite[,,1]
g <- yosemite[,,2]
b <- yosemite[,,3]
```

Després calculem la descomposició en valors singulars mitjançant la instrucció `svd`. Emmagatzemem els resultats en les següents variables: `yosemite.r.svd`, `yosemite.g.svd` i `yosemite.b.svd`. Ajuntem totes tres descomposicions en una llista mitjançant la instrucció `list`.

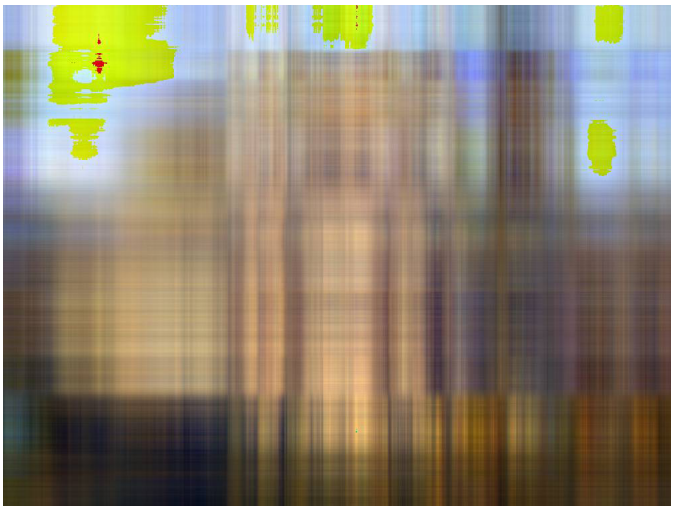
```
yosemite.r.svd <- svd(r)
yosemite.g.svd <- svd(g)
yosemite.b.svd <- svd(b)
rgb.svds <- list(yosemite.r.svd, yosemite.g.svd, yosemite.b.svd)
```

Cada variable (`yosemite.r.svd`, `yosemite.g.svd` i `yosemite.b.svd`) és una estructura de dades que conté tres camps: `u` (matriu que conté els vectors singulars per l'esquerra), `d` (vector amb els valors singulars) i `v` (matriu que conté els vectors singulars per la dreta). Accedim a cadascun d'aquests camps escrivint, per exemple, `yosemite.r.svd$u` o `yosemite.g.svd$d`.

Per calcular d'una manera ràpida i compacta la ν-èssima suma parcial fem servir la instrucció `sapply`, que permet aplicar una funció a una llista. En un primer exemple, si fixem $\nu = 3$, la tercera suma parcial es pot emmagatzemar en la variable `a`. Finalment, es pot desar la imatge comprimida resultant fent servir la instrucció `writeJPEG`. La imatge que en resulta es pot veure a la figura 8.

```
nu <- 3
a <- sapply(rgb.svds, function(i) {
     yosemite.compress <- i$u[,1:nu]
     %*% diag(i$d[1:nu]) %*% t(i$v[,1:nu])},
     simplify = 'array')
writeJPEG(a,'yose003.jpg')
```

Figura 8. Imatge del Parc Nacional de Yosemite (en escala RGB, 750×1.000 píxels i comprimida a tres valors singulars)

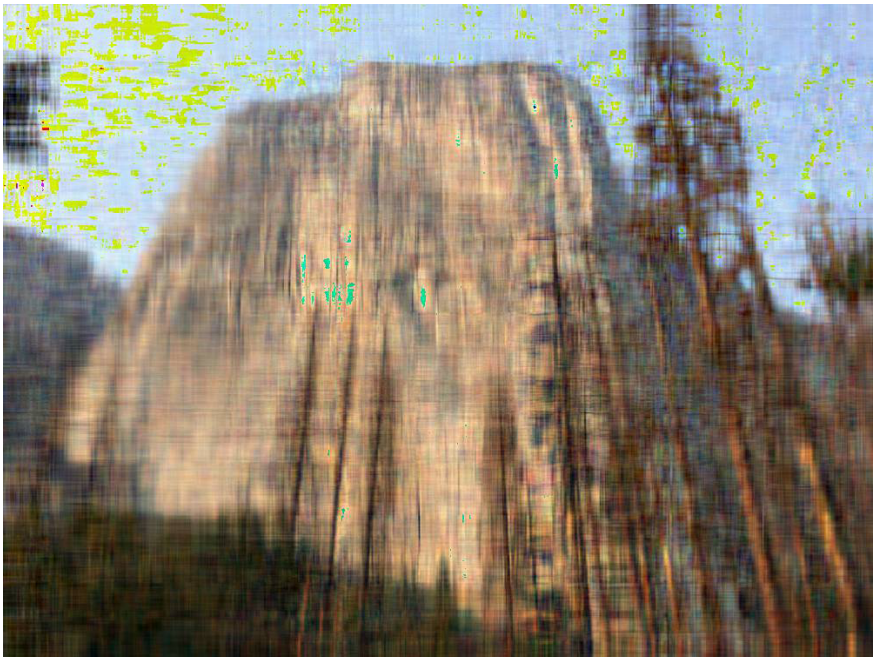


Font: elaboració pròpia

De la mateixa manera podem calcular la imatge comprimida que resulta de la ν-èssima suma parcial si $\nu = 20$, que hem representat a la figura 9.

```
nu <- 20
a <- sapply(rgb.svds, function(i) {
     yosemite.compress <- i$u[,1:nu]
     %*% diag(i$d[1:nu]) %*% t(i$v[,1:nu])},
     simplify = 'array')
writeJPEG(a,'yose020.jpg')
```

Figura 9. Imatge del Parc Nacional de Yosemite (en escala RGB, 750×1.000 píxels i comprimida a vint valors singulars)



Font: elaboració pròpia

Finalment, la imatge comprimida que resulta de la ν-èssima suma parcial si $\nu = 200$ es pot veure a la figura 10.

```
nu <- 200
a <- sapply(rgb.svds, function(i) {
     yosemite.compress <- i$u[,1:nu]
     %*% diag(i$d[1:nu]) %*% t(i$v[,1:nu])},
     simplify = 'array')
writeJPEG(a,'yose200.jpg')
```

A tall il·lustratiu i pel que fa a la imatge de la figura 10 comprimida a dos-cents valors singulars, els valors singulars σ^r_{201},σ^g_{201} i σ^b_{201}, corresponents a les matrius de colors vermell, verd i blau, respectivament, són:

$$\sigma^r_{201} = 4.195227,$$

$$\sigma^g_{201} = 4.19045,$$

$$\sigma^b_{201} = 4.033507.$$

Hem obtingut aquests valors amb `yosemite.r.svd$d[201]`, `yosemite.g.svd$d[201]` i `yosemite.g.svd$d[201]`, respectivament.

Figura 10. Imatge del Parc Nacional de Yosemite (en escala RGB, 750×1.000 píxels i comprimida a 200 valors singulars)

Font: elaboració pròpia

Com hem d'interpretar aquestes quantitats? Recordeu que la nostra imatge inicial està formada per tres matrius: una per al color vermell, una per al color verd i una altra per al color blau. És a dir, tenim tres matrius, $\mathbf{A}^r$, $\mathbf{A}^g$ i $\mathbf{A}^b$, que hem aproximat (en la figura 10, per una imatge comprimida formada per sumes parcials) per a cada un dels colors: $\mathbf{A}^r_{200}$, $\mathbf{A}^g_{200}$ i $\mathbf{A}^b_{200}$. Aleshores, l'error per a cadascun dels colors, en norma 2, és:

$$\|\mathbf{A}^r - \mathbf{A}^r_{200}\| = \sigma^r_{201} = 4.195227$$

$$\|\mathbf{A}^g - \mathbf{A}^g_{200}\| = \sigma^g_{201} = 4.19045$$

$$\|\mathbf{A}^b - \mathbf{A}^b_{200}\| = \sigma^b_{201} = 4.033507$$

L'error en la compressió de la imatge és petit, tot i que encara és observable. Això és així perquè la imatge presenta un nombre important de contrastos i els valors singulars decreixen a poc a poc. En termes d'espai necessari per emmagatzemar la imatge original calen

$$750 \times 1.000 = 750.000$$

elements. Si fem servir la suma parcial amb $v = 200$, cal emmagatzemar 200 matrius de rang 1 definides per $\sigma_i \mathbf{u}_i \mathbf{v}_i^T$, en què $\mathbf{u}_i$ té dimensió 750 i $\mathbf{v}_i$ té dimensió 1.000. Això correspon a:

$$v \cdot (m + n) + v = 200 \cdot (750 + 1.000) + 200 = 350.200$$

elements. En aquest cas, això representa un:

$$\frac{350.200}{750.000} \times 100\% = 46.70\%$$

de la mida original.

Exemple

Aquest exemple és igual que l'anterior, en el sentit que considerem una imatge en escala de colors RGB i de dimensió 750×1.000 píxels. La imatge és, però, una mica més senzilla, com es pot veure a la figura 11.

Figura 11. Imatge d'una flor d'hibisc sobre un llençol (en escala RGB i 750×1.000 píxels)

Font: elaboració pròpia

A la figura 12 hem representat la imatge comprimida amb els dos-cents primers valors singulars. Els valors singulars σ^r_{201},σ^g_{201} i σ^b_{201}, corresponents a les matrius de colors vermell, verd i blau, respectivament, són, en aquest cas:

$$\begin{aligned}\sigma^r_{201} &= 3.1221,\\ \sigma^g_{201} &= 3.483641,\\ \sigma^b_{201} &= 3.200724\end{aligned}$$

Figura 12. Imatge d'una flor d'hibisc sobre un llençol (en escala RGB, 750×1.000 píxels i comprimida a 200 valors singulars)

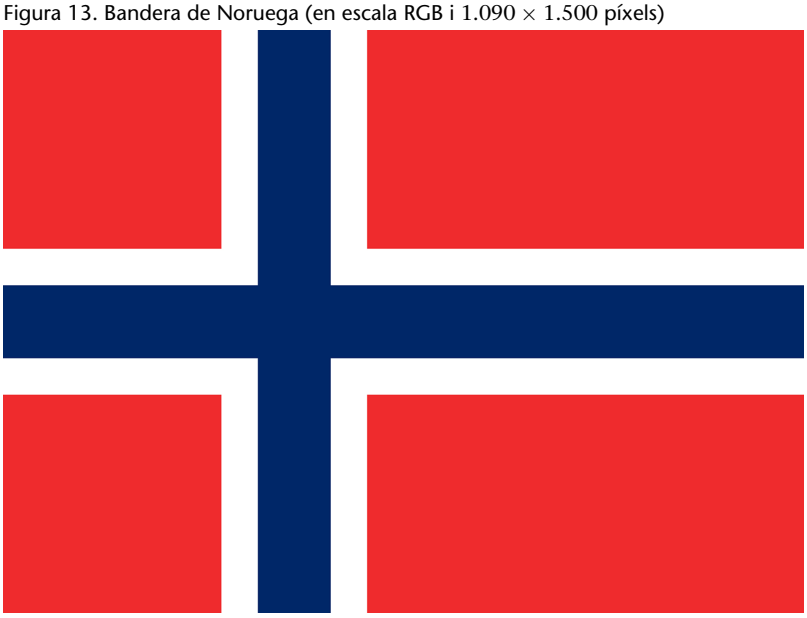
Font: elaboració pròpia

Aquests valors són inferiors als valors singulars de l'exemple anterior. Això significa que amb el mateix nombre de valors singulars (200) obtenim una compressió més bona. De fet, aquesta millora és perfectament observable.

Exemple

Finalment, veurem un exemple de compressió extrema a partir de la imatge de la bandera de Noruega. Es tracta d'una imatge en escala de colors RGB, de dimensió 1.090×1.500 píxels, tal com mostra la figura 13.

Figura 13. Bandera de Noruega (en escala RGB i 1.090×1.500 píxels)

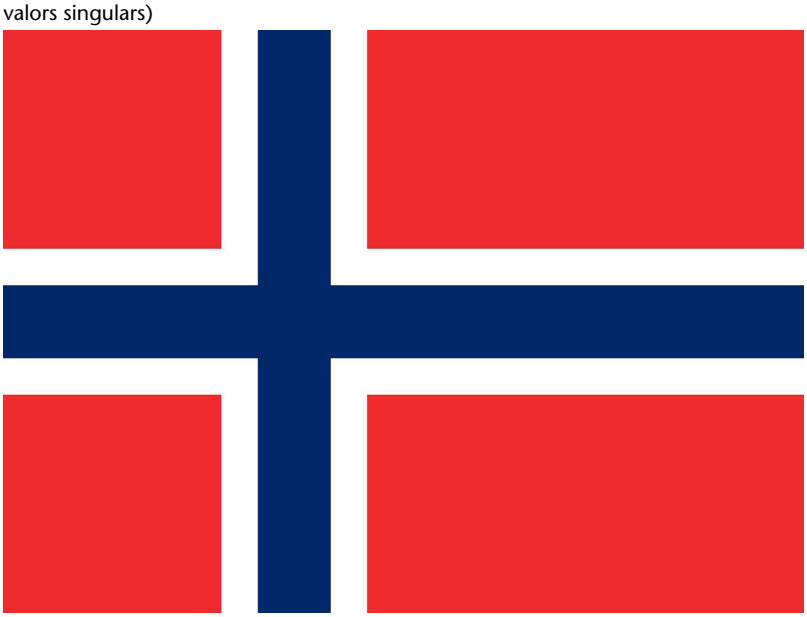


Font: https://www.countryflags.com/en/norway-flag-image.html

A la figura 14 hem representat la imatge comprimida amb els tres primers valors singulars. Els valors singulars σ_4^r,σ_4^g i σ_4^b, corresponents a les matrius de colors vermell, verd i blau, respectivament, són, en aquest cas:

$$\sigma_4^r = 5.199605 \cdot 10^{-2},$$
$$\sigma_4^g = 3.652678 \cdot 10^{-2},$$
$$\sigma_4^b = 5.095715 \cdot 10^{-2}.$$

Figura 14. Bandera de Noruega (en escala RGB, 1.090×1.500 píxels i comprimida a tres valors singulars)



Font: elaboració pròpia

L'error en la compressió de la imatge és inapreciable. En termes d'espai necessari per emmagatzemar la imatge original calen

$$1.090 \times 1.500 = 1.635.000$$

elements. Si fem servir la suma parcial amb $v = 3$, cal emmagatzemar tres matrius de rang 1 definides per $\sigma_i \mathbf{u}_i \mathbf{v}_i^T$, en què $\mathbf{u}_i$ té dimensió 1.090 i $\mathbf{v}_i$ té dimensió 1.500. Això correspon a

$$v \cdot (m + n) + v = 3 \cdot (1.090 + 1.500) + 3 = 6.423$$

elements. En aquest cas, això representa un:

$$\frac{6.423}{1.635.000} \times 100\% = 0.39\%$$

de la mida original.

Resum

En aquest mòdul hem vist dues tècniques, l'anàlisi de components principals (PCA) i la descomposició en valors singulars (SVD), en què la base matemàtica en tots dos casos és el càlcul de valors i vectors propis.

Pel que fa a l'anàlisi de components principals, cal tenir en compte alguns punts importants en aquesta estratègia:

1) Les dades que mesurem es poden organitzar, habitualment, en forma de matriu, en què el nombre de columnes (m) representa les diferents variables que mesurem i el nombre de files (n) representa els elements de la mostra.

2) Cal normalitzar les dades per evitar els efectes que les magnituds de cada variable tindrien sobre l'anàlisi. Així doncs, cal restar per la mitjana de cada variable i dividir per la desviació tipus de cada variable.

3) Es calcula la matriu de covariàncies mitjançant el producte de matrius $\frac{1}{n-1}\mathbf{X}^T\mathbf{X}$. El resultat és una matriu simètrica i definida positiva, en què els elements de la diagonal són tots uns.

4) Es calculen els valors i vectors propis de la matriu de covariàncies i s'ordenen de gran a petit. El vector propi associat al valor propi més gran s'anomena *primera component principal*. El vector propi associat al segon valor propi més gran s'anomena *segona component principal*. I així successivament. En general, els vectors propis de la matriu de covariàncies són les components principals.

5) La projecció de les dades originals sobre l'espai vectorial generat per les components principals permet descobrir característiques que poden passar desapercebudes. Aquesta projecció també permet reduir la dimensionalitat de les dades originals.

En relació amb la descomposició en valors singulars, els elements clau en aquesta estratègia són:

1) La descomposició en valors singulars descompon una matriu $\mathbf{A}$ de dimensió $n \times m$ com a producte de tres matrius, $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, en què $\mathbf{U}$,$\mathbf{V}$ són matrius unitàries i $\mathbf{\Sigma}$ és una matriu diagonal (en la versió reduïda). Els elements de la diagonal s'anomenen *valors singulars*, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, en què r és el rang de la matriu $\mathbf{A}$.

2) El quadrat dels valors singulars $\sigma_1^2, \ldots, \sigma_r^2$ són els valors propis no nuls tant de la matriu $\mathbf{A}\mathbf{A}^T$ com de la matriu $\mathbf{A}^T\mathbf{A}$.

3) Les columnes ortonormals de la matriu $\mathbf{U}$ són els vectors propis de la matriu $\mathbf{A}\mathbf{A}^T$ i reben el nom de *vectors singulars per l'esquerra*.

4) Les columnes ortonormals de la matriu $\mathbf{V}$ són els vectors propis de la matriu $\mathbf{A}^T\mathbf{A}$ i reben el nom de *vectors singulars per la dreta*.

5) Els vectors singulars per l'esquerra i els vectors singulars per la dreta són dues bases que diagonalitzen la matriu $\mathbf{A}$. És a dir, $\mathbf{A}\mathbf{v}_i = \sigma_i\mathbf{u}_i$, $i \leq r$. Equivalentment, $\mathbf{AV} = \mathbf{\Sigma U}$.

6) Una de les aplicacions més interessants de la descomposició en valors singulars és la compressió d'imatges.

Exercicis d'autoavaluació

1. Considereu aquesta matriu, que conté el valor de dues variables per a una mostra de cinc elements:

$$\mathbf{X}_0 = \begin{bmatrix} 5 & -1 \\ 4 & 1 \\ 3 & 0 \\ 2 & 1 \\ 0 & -1 \end{bmatrix}$$

Calculeu la mitjana aritmètica de cada columna per obtenir la matriu centrada $\mathbf{X}$. Calculeu-ne la matriu de covariàncies $\mathbf{C_X}$ i els valors propis λ_1 i λ_2. Quina és la direcció de la primera component principal?

2. Considereu aquesta matriu, que conté el valor de dues variables per a una mostra de sis elements:

$$\mathbf{X}_0 = \begin{bmatrix} 1 & 1 \\ 2 & 0 \\ 3 & 1 \\ 3 & 0 \\ 2 & 1 \\ 1 & 0 \end{bmatrix}$$

Calculeu la mitjana aritmètica de cada columna per obtenir la matriu centrada $\mathbf{X}$. Calculeu-ne la matriu de covariàncies $\mathbf{C_X}$ i els valors propis λ_1 i λ_2. Quina és la direcció de la primera component principal?

3. Considereu aquesta matriu, que conté el valor de tres variables per a una mostra de quatre elements:

$$\mathbf{X}_0 = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 0 & 1 \\ 0 & 2 & -1 \\ 0 & -2 & -1 \end{bmatrix}$$

Calculeu la mitjana aritmètica i la desviació tipus de cada columna per obtenir la matriu normalitzada $\mathbf{X}$. Calculeu-ne la matriu de covariàncies $\mathbf{C_X}$ i els valors propis λ_1,λ_2 i λ_3. Quina és la direcció de la primera component principal?

4. Considereu aquesta matriu:

$$\mathbf{A} = \begin{bmatrix} 0 & 4 \\ 0 & 0 \end{bmatrix}$$

Calculeu els valors propis de la matriu $\mathbf{A}$. Trobeu els valors singulars de la matriu $\mathbf{A}^T\mathbf{A}$. Calculeu $\mathbf{V}$ mitjançant els vectors propis de la matriu $\mathbf{A}^T\mathbf{A}$ i $\mathbf{U}$ mitjançant els vectors propis de la matriu $\mathbf{A}\mathbf{A}^T$. Comproveu que $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

5. Considereu aquesta matriu:

$$\mathbf{A} = \begin{bmatrix} 0 & 4 \\ 1 & 0 \end{bmatrix}$$

Calculeu els valors propis de la matriu $\mathbf{A}$. Trobeu els valors singulars de la matriu $\mathbf{A}^T\mathbf{A}$. Calculeu $\mathbf{V}$ mitjançant els vectors propis de la matriu $\mathbf{A}^T\mathbf{A}$ i $\mathbf{U}$ mitjançant els vectors propis de la matriu $\mathbf{A}\mathbf{A}^T$. Comproveu que $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

6. Considereu aquesta matriu:

$$\mathbf{A} = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}$$

Calculeu $\mathbf{A}^T\mathbf{A}$,$\mathbf{V}$,$\mathbf{\Sigma}$ i $\mathbf{u}_i = \mathbf{A}\mathbf{v}_i/\sigma_i$ i la descomposició en valors singulars (SVD) completa. Comproveu que $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

7. Considereu una matriu qualsevol $\mathbf{A}$. Demostreu que els valors singulars d'aquesta matriu σ_i són únics.

8. Considereu una matriu qualsevol $\mathbf{A}$. Demostreu que el rang de la matriu coincideix amb el nombre de valors singulars no nuls.

9. Considereu una matriu quadrada qualsevol $\mathbf{A}$. Demostreu que el valor absolut del determinant d'aquesta matriu és igual que el producte dels valors singulars, és a dir:

$$|\det(\mathbf{A})| = \prod_{i=1}^{n} \sigma_i$$

Solucionari

1. La mitjana aritmètica de la primera columna és 2.8 i la de la segona, 0. Per tant, la matriu centrada és:

$$\mathbf{X} = \begin{bmatrix} 2.2 & -1 \\ 1.2 & 1 \\ 0.2 & 0 \\ -0.8 & 1 \\ -2.8 & -1 \end{bmatrix}$$

La matriu de covariàncies és:

$$\mathbf{C_X} = \frac{1}{4}\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 3.70 & 0.25 \\ 0.25 & 1.00 \end{bmatrix}$$

Observeu que la matriu de covariàncies no té uns en la diagonal perquè no hem dividit les columnes per les desviacions tipus.

Els valors propis de la matriu de covariàncies són:

$$\lambda_1 = 3.722953$$

$$\lambda_2 = 0.977047$$

Fixeu-vos que

$$\lambda_1 + \lambda_2 = 4.7 = \mathrm{tr}(\mathbf{C_X}).$$

Finalment, la direcció de la primera component principal és

$$p_1^T = [-0.99581173, -0.09142755].$$

La resolució en R d'aquest problema es pot veure a la figura 15.

2. La mitjana aritmètica de la primera columna és 2 i la de la segona, 0.5. Per tant, la matriu centrada és:

$$\mathbf{X} = \begin{bmatrix} -1 & 0.5 \\ 0 & -0.5 \\ 1 & 0.5 \\ 1 & -0.5 \\ 0 & 0.5 \\ -1 & -0.5 \end{bmatrix}$$

Figura 15

```
> a<-c(5,-1,4,1,3,0,2,1,0,-1)
> X0<-matrix(a,ncol=2,nrow=5,byrow=TRUE)
> X0[,1]<-X0[,1]-mean(X0[,1])
> X0[,2]<-X0[,2]-mean(X0[,2])
> X<-X0
> X
     [,1] [,2]
[1,]  2.2   -1
[2,]  1.2    1
[3,]  0.2    0
[4,] -0.8    1
[5,] -2.8   -1
> CX<-t(X)%*%X/4
> CX
     [,1] [,2]
[1,] 3.70 0.25
[2,] 0.25 1.00
> eigCX<-eigen(CX)
> eigCX$values
[1] 3.722953 0.977047
> eigCX$vectors
           [,1]        [,2]
[1,] -0.99581173  0.09142755
[2,] -0.09142755 -0.99581173
```

Font: elaboració pròpia

La matriu de covariàncies és:

$$\mathbf{C_X} = \frac{1}{5}\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 0.8 & 0.0 \\ 0.0 & 0.3 \end{bmatrix}$$

Observeu que la matriu de covariàncies no té uns en la diagonal perquè no hem dividit les columnes per les desviacions tipus. Fixeu-vos que la matriu de covariàncies és, en aquest cas, una matriu diagonal.

Els valors propis de la matriu de covariàncies són:

$$\lambda_1 = 0.8$$

$$\lambda_2 = 0.3$$

Fixeu-vos que

$$\lambda_1 + \lambda_2 = 1.1 = \mathrm{tr}(\mathbf{C_X}).$$

Finalment, la direcció de la primera component principal és

$$p_1^T = [-1,0].$$

Observeu que, com que la matriu de covariàncies és diagonal, la primera component principal és igual (pel que fa a la direcció) que la variable que hem considerat a la primera columna de la matriu $\mathbf{X}_0$.

La resolució en R d'aquest problema es pot veure a la figura 16.

Figura 16

```
> a<-c(1,2,3,3,2,1,1,0,1,0,1,0)
> X0<-matrix(a,ncol=2,nrow=6,byrow=FALSE)
> X0[,1]<-X0[,1]-mean(X0[,1])
> X0[,2]<-X0[,2]-mean(X0[,2])
> X<-X0
> X
     [,1] [,2]
[1,]   -1  0.5
[2,]    0 -0.5
[3,]    1  0.5
[4,]    1 -0.5
[5,]    0  0.5
[6,]   -1 -0.5
> CX<-t(X)%*%X/5
> CX
     [,1] [,2]
[1,]  0.8  0.0
[2,]  0.0  0.3
> eigCX<-eigen(CX)
> eigCX$values
[1] 0.8 0.3
> eigCX$vectors
     [,1] [,2]
[1,]   -1    0
[2,]    0   -1
```

Font: elaboració pròpia

3. La mitjana aritmètica de la primera columna és 0; la de la segona, 0, i la de la tercera, 0. Les desviacions tipus de les columnes són 0.8164966,1.632993 i 1.154701, respectivament. Per tant, la matriu esglaonada és:

$$\mathbf{X} = \begin{bmatrix} 1.224745 & 0.000000 & 0.8660254 \\ -1.224745 & 0.000000 & 0.8660254 \\ 0.000000 & 1.224745 & -0.8660254 \\ 0.000000 & -1.224745 & -0.8660254 \end{bmatrix}$$

La matriu de covariàncies és:

$$\mathbf{C_X} = \frac{1}{6}\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Observeu que ara la matriu de covariàncies sí que té uns en la diagonal, ja que hem dividit les columnes per les desviacions tipus. Fixeu-vos que la matriu de covariàncies és, en aquest cas, una matriu diagonal.

Els valors propis de la matriu de covariàncies són:

$$\lambda_1 = 1$$

$$\lambda_2 = 1$$

$$\lambda_3 = 1$$

Fixeu-vos que

$$\lambda_1 + \lambda_2 + \lambda_3 = 3 = \mathrm{tr}(\mathbf{C_X}).$$

Finalment, la direcció de la primera component principal és

$$p_1^T = [0,0,1].$$

En aquest cas, després d'haver obtingut els tres valors propis iguals a 1, no hi ha cap valor més gran que la resta. La segona i tercera components principals són, segons els resultats que obtenim en R:

$$p_2^T = [0,1,0]$$

$$p_3^T = [1,0,0]$$

Per tant, tenint en compte que la matriu de covariància és la matriu identitat, hauria estat més normal definir les components principals així:

$$p_1^T = [1,0,0]$$

$$p_2^T = [0,1,0]$$

$$p_3^T = [0,0,1]$$

La resolució en R d'aquest problema es pot veure a la figura 17.

Figura 17

```
> a<-c(1,-1,0,0,0,0,2,-2,1,1,-1,-1)
> X0<-matrix(a,ncol=3,nrow=4,byrow=FALSE)
> X0[,1]<-(X0[,1]-mean(X0[,1]))/sd(X0[,1])
> X0[,2]<-(X0[,2]-mean(X0[,2]))/sd(X0[,2])
> X0[,3]<-(X0[,3]-mean(X0[,3]))/sd(X0[,3])
> X<-X0
> X
          [,1]      [,2]       [,3]
[1,]  1.224745  0.000000  0.8660254
[2,] -1.224745  0.000000  0.8660254
[3,]  0.000000  1.224745 -0.8660254
[4,]  0.000000 -1.224745 -0.8660254
> CX<-t(X)%*%X/3
> CX
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
> eigCX<-eigen(CX)
> eigCX$values
[1] 1 1 1
> eigCX$vectors
     [,1] [,2] [,3]
[1,]    0    0    1
[2,]    0    1    0
[3,]    1    0    0
```

Font: elaboració pròpia

4. El determinant de la matriu $\mathbf{A}$ és 0 i la traça, també. Per tant, sabem que els valors propis de la matriu $\mathbf{A}$ satisfan:

$$\lambda_1 + \lambda_2 = \text{tr}(\mathbf{A}) = 0$$

$$\lambda_1\lambda_2 = \det(\mathbf{A}) = 0$$

Això implica que $\lambda_1 = \lambda_2 = 0$. Fixeu-vos que el rang de la matriu $\mathbf{A}$ és $r = 1$. D'altra banda, la matriu $\mathbf{A}^T\mathbf{A}$ és igual a:

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 16 \end{bmatrix},$$

que té com a valors propis $\lambda_1 = 16$ i $\lambda_2 = 0$. Per tant, l'únic valor singular és $\sigma_1 = \sqrt{\lambda_1} = 4$. Els vectors propis de la matriu $\mathbf{A}^T\mathbf{A}$ són:

$$\mathbf{v}_1^T = [0,1]$$

$$\mathbf{v}_2^T = [-1,0],$$

de manera que la matriu $\mathbf{V}$ és igual a:

$$\mathbf{V} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

D'altra banda, la matriu $\mathbf{A}\mathbf{A}^T$ és igual a:

$$\mathbf{A}\mathbf{A}^T = \begin{bmatrix} 16 & 0 \\ 0 & 0 \end{bmatrix},$$

que té com a valors propis $\lambda_1 = 16$ i $\lambda_2 = 0$. Per tant, l'únic valor singular és $\sigma_1 = \sqrt{\lambda_1} = 4$. Els vectors propis de la matriu $\mathbf{A}\mathbf{A}^T$ són:

$$\mathbf{u}_1^T = [1,0]$$

$$\mathbf{u}_2^T = [0,1],$$

de manera que la matriu $\mathbf{V}$ és igual a:

$$\mathbf{V} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Es pot comprovar, doncs, que $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$:

$$\begin{bmatrix} 0 & 4 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}^T$$

5. El determinant de la matriu $\mathbf{A}$ és –4 i la traça és 0. Per tant, sabem que els valors propis de la matriu $\mathbf{A}$ satisfan:

$$\lambda_1 + \lambda_2 = \text{tr}(\mathbf{A}) = 0$$

$$\lambda_1\lambda_2 = \det(\mathbf{A}) = -4$$

Això implica que $\lambda_1 = 2$ i $\lambda_2 = -\lambda_1 = -2$. Fixeu-vos que el rang de la matriu $\mathbf{A}$ és $r = 2$. D'altra banda, la matriu $\mathbf{A}^T\mathbf{A}$ és igual a:

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 16 \end{bmatrix},$$

que té com a valors propis $\lambda_1 = 16$ i $\lambda_2 = 1$. Per tant, els valors singulars són $\sigma_1 = \sqrt{\lambda_1} = 4$ i $\sigma_2 = \sqrt{\lambda_2} = 1$. Els vectors propis de la matriu $\mathbf{A}^T\mathbf{A}$ són:

$$\mathbf{v}_1^T = [0, -1]$$

$$\mathbf{v}_2^T = [-1,0],$$

de manera que la matriu $\mathbf{V}$ és igual a:

$$\mathbf{V} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

D'altra banda, la matriu $\mathbf{A}\mathbf{A}^T$ és igual a:

$$\mathbf{A}\mathbf{A}^T = \begin{bmatrix} 16 & 0 \\ 0 & 1 \end{bmatrix}$$

que té com a valors propis $\lambda_1 = 16$ i $\lambda_2 = 1$. Per tant, els valors singulars són $\sigma_1 = \sqrt{\lambda_1} = 4$ i $\sigma_2 = \sqrt{\lambda_2} = 1$. Els vectors propis de la matriu $\mathbf{A}\mathbf{A}^T$ són:

$$\mathbf{u}_1^T = [-1,0]$$

$$\mathbf{u}_2^T = [0, -1],$$

de manera que la matriu $\mathbf{V}$ és igual a:

$$\mathbf{V} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

Es pot comprovar, doncs, que $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$:

$$\begin{bmatrix} 0 & 4 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}^T$$

6. La matriu $\mathbf{A}^T\mathbf{A}$ és igual a:

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix},$$

que té com a valors propis $\lambda_1 = 8$ i $\lambda_2 = 2$. Per tant, els valors singulars són $\sigma_1 = \sqrt{\lambda_1} = 2\sqrt{2}$ i $\sigma_2 = \sqrt{\lambda_2} = \sqrt{2}$. Els vectors propis de la matriu $\mathbf{A}^T\mathbf{A}$ són:

$$\mathbf{v}_1^T = [-\sqrt{2}/2, -\sqrt{2}/2]$$

$$\mathbf{v}_2^T = [-\sqrt{2}/2, \sqrt{2}/2],$$

de manera que la matriu $\mathbf{V}$ és igual a:

$$\mathbf{V} = \begin{bmatrix} -\sqrt{2}/2 & -\sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}$$

La matriu $\mathbf{\Sigma}$ és igual a:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} = \begin{bmatrix} 2\sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix}$$

Calcularem els vectors $\mathbf{u}_i$ seguint la indicació $\mathbf{u}_i = \mathbf{A}\mathbf{v}_i/\sigma_i$. En efecte:

$$\mathbf{u}_1 = \frac{1}{2\sqrt{2}} \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} -\sqrt{2}/2 \\ -\sqrt{2}/2 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

$$\mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Es pot comprovar, doncs, que $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$:

$$\begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2\sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} -\sqrt{2}/2 & -\sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}^T$$

7. Els valors singulars de la matriu $\mathbf{A}$ són les arrels quadrades dels valors propis de la matriu $\mathbf{A}^T\mathbf{A}$. Sabem que els valors propis de qualsevol matriu són únics, la qual cosa implica que els valors singulars de la matriu $\mathbf{A}$ són també únics.

8. Considerem la descomposició en valors singulars completa de la matriu $\mathbf{A}$. Això implica l'existència de dues matrius, $\mathbf{U}$ i $\mathbf{V}$, de rang màxim, ja que estan formades per una base de vectors ortonormals dels corresponents espais vectorials. Aleshores:

$$\text{rang}(\mathbf{A}) = \text{rang}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T) = \text{rang}(\mathbf{\Sigma})$$

El rang de la matriu diagonal $\mathbf{\Sigma}$ és igual que el nombre de valors no nuls de la seva diagonal, és a dir, és el nombre de valors singulars no nuls de la matriu $\mathbf{A}$.

9. Considerem la descomposició en valors singulars completa de la matriu $\mathbf{A}$. Això implica l'existència de dues matrius, $\mathbf{U}$ i $\mathbf{V}$, de rang màxim, ja que estan formades per una base de vectors ortonormals dels corresponents espais vectorials. Tenint en compte que el determinant d'un producte de matrius és igual que el producte dels determinants, i tenint en compte també que el determinant d'una matriu ortogonal és 1, obtenim:

$$|\det(\mathbf{A})| = |\det(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)| = |\det(\mathbf{U})\det(\mathbf{\Sigma})\det(\mathbf{V}^T))| = |\det(\Sigma)| = \prod_{i=1}^{n} \sigma_i$$

Glossari

components principals d'una matriu *f* Vectors propis de la matriu de covariàncies corresponent.

matriu de covariàncies *f* Matriu quadrada que en la posició (*i*,*j*) conté la covariància de les variables *i* i *j*.

matriu unitària *f* Matriu en què el producte escalar de dues columnes diferents és igual a zero i en què el producte escalar de dues columnes iguals equival a 1. La inversa d'una matriu unitària és igual que la seva transposada.

primera component principal *f* Component principal associada al valor propi màxim de la matriu de covariàncies.

valors singulars *m* Arrels quadrades dels valors propis de les matrius $\mathbf{A}\mathbf{A}^T$ i $\mathbf{A}^T\mathbf{A}$.

vectors singulars per l'esquerra *m* Vectors propis de la matriu $\mathbf{A}\mathbf{A}^T$.

vectors singulars per la dreta *m* Vectors propis de la matriu $\mathbf{A}^T\mathbf{A}$.

Bibliografia

Castellet, Manuel; Llerena, Irene (2010). *Álgebra lineal y geometría*. Barcelona: Editorial Reverté.

Jolliffe, Ian T. (2002). *Principal Component Analysis*. Berlín: Springer.

Liesen, Jörg; Mehrmann, Volker (2015). *Linear Algebra*. Berlín: Springer.

Peña, Daniel (2013). *Análisis de datos multivariantes*. Madrid: McGraw-Hill Espanya.

Strang, Gilbert (2016). *Introduction to Linear Algebra*. Wellesley (Massachusetts, Estats Units d'Amèrica): Wellesley-Cambridge Press.