MA1034

Análisis de Componentes Principales

Se realizará un PCA sobre el dataset "ma1034-datos-sitprb3D" con columnas ind1, ind2, ind3.

y un subset de 20 filas, para fines prácticos, de sus 201 filas originales.

ind1	ind2	ind3
11.93231	1.80293	4.73133
7.410966	0.623265	4.47807
4.485586	2.532899	1.676949
2.734415	4.566107	4.590162
5.058295	1.848201	2.776555
12.10734	6.397484	5.526137
4.023766	2.863106	2.13785
4.72425	0.216941	3.582792
3.483725	1.500135	1.454271
5.777209	2.754675	1.3248
6.711074	3.318296	4.771564
15.11446	5.056667	4.222457
8.360825	4.36004	3.27595
9.860661	4.90298	2.836638
6.22788	5.769965	2.325835
8.740067	6.812796	4.366692
5.143418	4.134362	2.427982
8.249039	5.31135	3.613925
6.306126	7.765326	4.478954
6.152889	3.86089	3.575643

$$\longrightarrow \mathbb{X} = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \\ \vdots & \vdots & \vdots \\ X_{201} & X_{202} & X_{203} \end{bmatrix}$$

$$\mathbb{X} \in \mathbb{R}^{20\times 3}$$

(Se utilizará negritas de pizarrón para representar matrices)

1) Matriz de datos centrados

a) $\overline{\mathbb{X}} = \frac{1}{n} \cdot \mathbb{X}^T \cdot \mathbb{1}$

- $\mathbb{1}$ es una matriz de $n \times 1$ con solo unos.
- $\mathbb{X}^T$ es la matriz transpuesta de $\mathbb{X}$
- $\overline{\mathbb{X}}$ es el vector columna $m \times 1$ con los promedios de las columnas de $\mathbb{X}$.

$$\mathbb{X}^T \cdot \mathbb{1} = \begin{bmatrix} 142.604301 \\ 76.398415 \\ 68.174556 \end{bmatrix}$$

(Suma de cada columna)

$n = 20$

$$\frac{1}{20} \cdot \begin{bmatrix} 142.604301 \\ 76.398415 \\ 68.174556 \end{bmatrix} = \begin{bmatrix} 7.13021505 \\ 3.81992075 \\ 3.4087278 \end{bmatrix}$$

$$\overline{\mathbb{X}} = \begin{bmatrix} 7.13021505 \\ 3.81992075 \\ 3.4087278 \end{bmatrix}$$

b) $\tilde{X} = X - \mathbb{1} \cdot \bar{X}^T$ — $\bar{X}^T$ es la transpuesta de $\bar{X}$

— $\mathbb{1}$ es el vector columna de 1s de nuevo.

$\mathbb{1} \cdot \bar{X}^T$

$$\begin{bmatrix}1\\1\\1\\\vdots\\1\end{bmatrix} \cdot \begin{bmatrix}7.13021505 & 3.81992075 & 3.4087278\end{bmatrix}$$

$$\mathbb{1} \cdot \bar{X}^T = \begin{bmatrix}7.1302.. & 3.8199... & 3.4087..\\ 7.1302... & 3.8199... & 3.4087...\\ \vdots & \vdots & \vdots\\ 7.1302 & 3.8199... & 3.4087...\end{bmatrix} \quad (20\times 3)$$

Xhat		
4.80209495	-2.0169908	1.3226022
0.28075095	-3.1966558	1.0693422
-2.6446291	-1.2870218	-1.7317788
-4.3958001	0.74618625	1.1814342
-2.0719201	-1.9717198	-0.6321728
4.97712495	2.57756325	2.1174092
-3.1064491	-0.9568148	-1.2708778
-2.4059651	-3.6029798	0.1740642
-3.6464901	-2.3197858	-1.9544568
-1.3530061	-1.0652458	-2.0839278
-0.419141	-0.5016248	1.3628362
7.98424495	1.23674625	0.8137292
1.23060995	0.54011925	-0.1327778
2.73044595	1.08305925	-0.5720898
-0.9023351	1.95004425	-1.0828928
1.60985195	2.99287525	0.9579642
-1.9867971	0.31444125	-0.9807458
1.11882395	1.49142925	0.2051972
-0.8240891	3.94540525	1.0702262
-0.977326	0.04096925	0.1669152

$\tilde{X} = X - \mathbb{1} \cdot \bar{X}^T$

$$\tilde{X} = \begin{bmatrix}\overset{ind1}{X_{11}} & \overset{ind2}{X_{12}} & \overset{ind3}{X_{13}}\\ X_{21} & X_{22} & X_{23}\\ \vdots & \vdots & \vdots\\ X_{20,1} & X_{20,2} & X_{20,3}\end{bmatrix} - \begin{bmatrix}\mu_1 & \mu_2 & \mu_3\\ \mu_1 & \mu_2 & \mu_3\\ \vdots & \vdots & \vdots\\ \mu_1 & \mu_2 & \mu_3\end{bmatrix}$$

(Se resta cada dato de cada columna menos su respectiva media de la columna.)

2) Matriz de Covarianzas

$$\Sigma = \frac{1}{n}\ \underset{\downarrow}{\tilde{X}^T}\cdot\underset{\downarrow}{\tilde{X}} \qquad \Sigma \in \mathbb{R}^{m\times m}$$

$m\times n$ $\quad n\times m$

$$\Sigma = \begin{bmatrix}\overset{ind1}{9.61098624} & \overset{ind2}{2.1256707} & \overset{ind3}{1.96073712}\\ 2.1256707 & 4.06708634 & 0.78425304\\ 1.96073712 & 0.78425304 & 1.45553615\end{bmatrix} \begin{matrix}\text{ind 1}\\ \text{ind 2}\\ \text{ind 3}\end{matrix}$$

3) A partir de la matriz de covarianza, obtener eigenvalores y eigenvectores.

Resolver:

$$\det(S - \lambda I) = 0$$

"Determinante de la matriz de covarianza menos lambda (variable) multiplicado por la matriz de identidad"

Escrito de esta forma:

$$\det\left(\begin{bmatrix} 9.610986 - \lambda & 2.1256707 & 1.960737 \\ 2.1256707 & 4.067086 - \lambda & 0.78425304 \\ 1.960737112 & 0.78425304 & 1.455536 - \lambda \end{bmatrix}\right) = 0$$

Esto sería:

$$(9.610986-\lambda)(4.067086-\lambda)(1.455536-\lambda) +$$
$$(2.1256707)(0.78425304)(1.960737112) +$$
$$(1.960737112)(2.1256707)(0.78425304) -$$
$$(1.960737112)(4.067086-\lambda)(1.960737) -$$
$$(0.78425304)(0.78425304)(9.610986-\lambda) -$$
$$(1.455536-\lambda)(2.1256707)(2.1256707) = 0$$

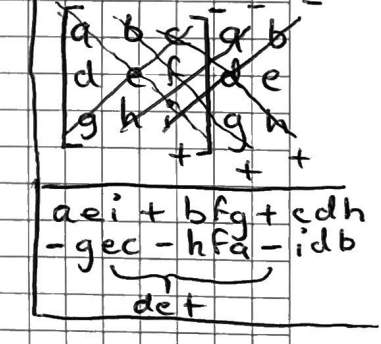
↓

$$(-\lambda^3 + 15.133608\lambda^2 - 58.9976267\lambda + 56.89502) +$$
$$(3.268867369) + (3.268673695) -$$
$$(15.63587167 - 3.844497\lambda) - (5.911264145 - 0.6150528\lambda3\lambda)$$
$$- (6.576009 - 4.518475\lambda) = 0$$

↓

$$(-\lambda^3 + 15.133608\lambda^2 - 50.01960719\lambda + 35.3092) = 0$$
$$\lambda^3 - 15.133608\lambda^2 + 50.01960719\lambda - 35.3092 = 0$$

Regla de Sarrus para determinante (3x3)



$$aei + bfg + cdh - gec - hfa - idb$$

det

Usando una calculadora (WolframAlpha)

Los eigenvalores son:

$\lambda_1 = 0.974983$

$\lambda_2 = 3.35084$

$\lambda_3 = 10.8078$

$\lambda_1 + \lambda_2 + \lambda_3 = 100\%$ (datos)

$\lambda_2 + \lambda_3 = \frac{10.8078 + 3.35084}{15.133623} = \frac{14.15864}{15.133623}$

$= 0.93557 = \boxed{93.5\%}$ (datos)

$\lambda_3 = \frac{10.8078}{15.133623} = 0.714 = \boxed{71.4\%}$

Planteamos ahora, para obtener eigenvectores:

$(S - \lambda_i I)\vec{v} = \vec{0}$

$\downarrow$

$A\vec{v} = \vec{0}$ (Sistema de ecuaciones)

MUY IMPORTANTE

En el contexto de PCA, los eigenvalores obtenidos representan la cantidad de varianza explicada, por cada componente principal. La suma de TODOS, equivale a la varianza total de los datos. En este caso, λ_2 y λ_3 (los mayores eigenvalores) explican aprox. el $\boxed{93.5\%}$ de la "información" (varianza de los datos). Es decir, podemos reducir de 3 dimensiones a 2, sin perder info. significativa.

$\lambda_1 = 3.35084$:

$$\begin{bmatrix} 9.610986 - \lambda & 2.1256707 & 1.960737 \\ 2.1256707 & 4.067086 - \lambda & 0.7842534 \\ 1.960737 & 0.7842534 & 1.455536 - \lambda \end{bmatrix}$$

$\downarrow$

$$\begin{bmatrix} 6.260146 & 2.1256707 & 1.960737 \\ 2.1256707 & 0.716246 & 0.7842534 \\ 1.960737 & 0.7842304 & -1.895304 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

A $\qquad$ $\vec{v}$

Fijando $v_3 = 1$, y resolviendo "hacia atrás":

$\vec{v_2} = \begin{bmatrix} 0.3337 \\ -0.9417 \\ 1 \end{bmatrix}$ $\xrightarrow{\text{Normalizando}}$ $\frac{1}{\sqrt{(0.3337)^2 + (-0.9417)^2 + 1^2}} \cdot \vec{v} = \begin{bmatrix} 0.334 \\ -0.942 \\ -0.044 \end{bmatrix}$

Eigenvector para λ_2

$\lambda_3 = 10.8078$:

$$A = S - \lambda_3 I = \begin{bmatrix} 9.610986 - 10.8078 & 2.1256707 & 1.960737 \\ 2.1256707 & 4.067086 - 10.8078 & 0.78425304 \\ 1.960737 & 0.78425304 & 1.456 - 10.8078 \end{bmatrix}$$

$$A = \begin{bmatrix} -1.1968 & 2.126 & 1.961 \\ 2.126 & -6.7408 & 0.784 \\ 1.961 & 0.784 & -9.3518 \end{bmatrix}$$

$$A\vec{v_3} = \vec{0}$$

$$\underset{A}{\begin{bmatrix} -1.1968 & 2.126 & 1.961 \\ 2.126 & -6.7408 & 0.784 \\ 1.961 & 0.784 & -9.3518 \end{bmatrix}} \underset{\vec{v_3}}{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}} = \underset{\vec{0}}{\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}}$$

NOTA:
Normalizar los vectores asegura que todos tengan la misma escala, sin cambiar dirección.
Esto permitirá proyectar los datos correctamente y mantener la proporción real de varianza.

De nuevo, fijando v_3 como 1, y resolviendo hacia atrás:

$$\vec{v_3} = \begin{bmatrix} 4.19 \\ 1.438 \\ 1 \end{bmatrix} \xrightarrow{\text{Normalizamos}} \frac{1}{\sqrt{(4.19)^2 + (1.438)^2 + (1)^2}} \cdot \vec{v} = \begin{bmatrix} 0.923 \\ 0.317 \\ 0.220 \end{bmatrix}$$

Matriz de eigenvectores λ_2 y λ_3, que explica el 93.5% de la varianza:

$$\begin{matrix} \vec{v_3} & \vec{v_2} \end{matrix}$$
$$\begin{bmatrix} 0.923 & 0.334 \\ 0.317 & -0.942 \\ 0.220 & -0.044 \end{bmatrix}$$

Nota: Se acomoda $\lambda_3 = 10.8078$ primero, por ser el mayor.

4) Proyección final de datos con los componentes principales seleccionados. (Pasar de 3D → 2D)

Finalmente, se proyecta o mapea la matriz inicial de 3 principales indicadores, a una dimensión menor de 2 PC:

$$\mathbb{Z} = \tilde{\mathbb{X}} \cdot \mathbb{W}$$

- $\tilde{\mathbb{X}}$: Matriz centrada de los datos.
- $\mathbb{W}$: Matriz de eigenvectores de proyección normalizado.
- $\mathbb{Z}$: Matriz de componentes principales. (Matriz proyectada)

$$\tilde{\mathbb{X}} \cdot \mathbb{W} = \mathbb{Z}$$

$[20 \times 3] \quad [3 \times 2] \quad [20 \times 2]$ ←

Xhat		
4.80209495	-2.0169908	1.3226022
0.28075095	-3.1966558	1.0693422
-2.6446291	-1.2870218	-1.7317788
-4.3958001	0.74618625	1.1814342
-2.0719201	-1.9717198	-0.6321728
4.97712495	2.57756325	2.1174092
-3.1064491	-0.9568148	-1.2708778
-2.4059651	-3.6029798	0.1740642
-3.6464901	-2.3197858	-1.9544568
-1.3530061	-1.0652458	-2.0839278
-0.419141	-0.5016248	1.3628362
7.98424495	1.23674625	0.8137292
1.23060995	0.54011925	-0.1327778
2.73044595	1.08305925	-0.5720898
-0.9023351	1.95004425	-1.0828928
1.60985195	2.99287525	0.9579642
-1.9867971	0.31444125	-0.9807458
1.11882395	1.49142925	0.2051972
-0.8240891	3.94540525	1.0702262
-0.977326	0.04096925	0.1669152

$$\cdot \begin{bmatrix} 0.923 & 0.334 \\ 0.317 & -0.942 \\ 0.220 & -0.044 \end{bmatrix} =$$

$Z =$

4.0839	3.4457
-0.5190	3.0580
-3.2300	0.4053
-3.5609	-2.2231
-2.6765	1.1932
5.8768	-0.8589
-3.4502	-0.0803
-3.3246	2.5828
-4.5311	1.0533
-2.0450	0.6433
-0.2461	0.2726
7.9405	1.4659
1.2779	-0.0919
2.7377	-0.0831
-0.4529	-2.0907
2.6454	-2.3237
-1.9499	-0.9166
1.5506	-1.0403
0.7255	-4.0389
-0.8524	-0.3724

Y con esto, la matriz proyectada $\mathbb{Z}$ refleja la transformación de los datos originales centrados hacia un nuevo sistema de coordenadas definido por los PC. Conserva la mayor parte de la varianza de los datos en un espacio reducido.