Universidad de Granada

Estadística Multivariante

Prácticas

Pedro Ramos Suárez

Doble Grado de Ingeniería Informática y Matemáticas

7 de diciembre de 2022

Índice

1. PCA

Sea $X = (X_1, \dots, X_n)$. Cuando n es muy grande, tenemos un problema.

Buscamos reducir la dimensión, es decir, buscamos $(U_1, \dots, U_p)$ un $p < n$ que recoja la mayor cantidad de información:

$$U_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n$$
$$\vdots$$
$$U_p = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pn}X_n$$

Test de Bartlett:

$$H_0 \equiv det(R) = 1$$
$$H_1 \equiv det(R) \neq 1$$

Quiero que el test sea significativo (rechazar la hipótesis nula) $\Rightarrow p < 0{,}05 \Rightarrow$ rechazo H_0.

Test Barttlet: $\chi^2_{exp} = 789{,}26$
$p = 1{,}32e - 130 < 0{,}05$ Test significativo.

Con un botplox podemos encontrar los outliers que queremos eliminar.

1.1. Teoría

Sea $X = (X_1, \dots, X_p)$. Busco $U_i = a_i'X, i = 1, \dots, q \leq p$. $(a_i' = a_{i1}, \dots, a_{iq})$.

Tiene que ocurrir que $VarU_i$ sea máxima, y que U_i está incorrelado con $U_j, j = 1, \dots, q$.

Asumimos que X es centrado ($E[X] = 0$), y veremos que a_i son los vectores propios asociados a $R = E[XX']$, y $VarU_i = \lambda_i$ los valores propios asociados.

Nota: Sea $a \in \mathbb{R}^n$ y $A \in M_n(\mathbb{K}) \Rightarrow \frac{\delta a'Aa}{\delta a} = (a + A')a$.

1.1.1. Paso 1

En el primer paso se obtiene la primera componente principal U_1 maximizando su varianza.

Asumimos que a_1 es un vector unitario, es decir,

$$||a_1|| = a_1'a_1 = 1$$

Como $E[X] = 0 \Rightarrow E[a_1'X] = 0$. Entonces:

$$Var[U_1] = E[U_1^2] = E[a_1'Xa_1'X] = E[a_1'XXta_1] = a'E[XX']a_1 = a_1'Ra_1$$

Por lo que el problema queda de la forma:

$$\begin{cases} \text{máx}_{a_1}\, a_1'Ra_1 \\ \text{s.a. } a_1'a_1 = 1 \end{cases}$$

Aplicamos el Teorema de los multiplicadores de Lagrange, reduciendo el problema a:

$$\max_{a_1}\{a_1'Ra_1 - \lambda(a_1'a_1 - 1)\} = F$$

$$\frac{\delta F}{\delta a_1} = 0 \Rightarrow (R + R') - 2\lambda_1 Ia_1 = 0 \Rightarrow 2Ra_1 - 2\lambda_1 Ia_1 = 0 \Rightarrow (R - \lambda_1 I)a_1 = 0$$

Por lo que a_1 es un vector propio asociado a λ_1.

$$Ra_1 - \lambda_1 a_1 = 0 \Rightarrow a_1'Ra_1 = \lambda_1 a_1'a_1 \Rightarrow VarU_1 = \lambda_1 a_1'a_1 = \lambda_1$$

donde en la última igualdad hemos usado que a_1 es unitario.

1.1.2. Paso 2

Para calcular la segunda componente principal, U_2 tenemos:

$$\begin{cases} \text{máx}\, Var[U_2] \\ \text{s.a. } ||a_2|| = a_2 a_2' = 1 \\ cov(U_1, U_2) = 0 \end{cases} \Rightarrow \begin{cases} \text{máx}_{a_2}\, a_2' R a_2 \\ \text{s.a. } a_2' a_2 = 1 \\ a_1' R a_2 = 0 \end{cases}$$

Volvemos a aplicar el Teorema de los multiplicadores de Lagrange:

$$F = a_2' R a_2 - \lambda_2 (a_2' a_2 - 1) - \mu (a_1' R a_2)$$

$$\frac{\delta F}{\delta a_2} = 0 \Rightarrow 2Ra_2 - 2\lambda_2 I a_2 - \mu R a_1 = 0$$

Multiplicamos todo por a_1' a la izquierda:

$$2Ra_1' a_2 - 2\lambda_2 a_1' a_2 - \mu a_1' R a_1 = 0$$

Como a_1 y a_2 son perpendiculares, $a_1' a_2 = 0$, y como $a_1' R a_1 = Var[U_1] \neq 0$, tenemos:

$$0 - 0 - \mu a_1' R a_1 = 0 \Rightarrow \mu = 0$$

Y entonces:

$$2Ra_2 - 2\lambda_2 a_2 = 0 \Rightarrow (R - \lambda_2) a_2 = 0$$

Por lo que a_2 es el valor propio asociado a λ_2.

1.1.3. Paso 3

Para calcular la tercera componente principal, U_3 tenemos:

$$\begin{cases} \text{máx}\, Var[U_3] \\ \text{s.a. } ||a_3|| = a_3 a_3' = 1 \\ cov(U_1, U_3) = 0 \\ cov(U_2, U_3) = 0 \end{cases} \Rightarrow \begin{cases} \text{máx}_{a_3}\, a_3 R a_3 \\ \text{s.a. } a_3' a_3 = 1 \\ a_1 R a_3 = 0 \\ a_2 R a_3 = 0 \end{cases}$$

Volvemos a aplicar el Teorema de los multiplicadores de Lagrange:

$$F = a_3' R a_3 - \lambda_3 (a_3' a_3 - 1) - \mu_1 a_1' R a_3 - \mu_2 a_2' R a_3$$

$$\frac{\delta F}{\delta a_3} = 0 \Rightarrow 2Ra_3 - 2\lambda_3 I a_3 - \mu_1 R a_1 - \mu_2 R a_2 = 0$$

Multiplicamos todo por a_1' a la izquierda:

$$2Ra_1' a_3 - 2\lambda_3 a_1' a_3 - \mu_1 a_1' R a_1 - \mu_2 a_1' R a_2 = 0$$

Como a_1, a_2 y a_3 son perpendiculares, $a_1' a_2 = 0$ y $a_1' a_3 = 0$, y como $a_1' R a_1 = Var[U_1] \neq 0$ tenemos:

$$0 - 0 - \mu_1 Var[U_1] - 0 = 0 \Rightarrow \mu_1 = 0$$

Multiplicando todo por a_2' a la derecha:

$$2Ra_2' a_3 - 2\lambda_3 a_2' a_3 - \mu_1 a_2' R a_1 - \mu_2 a_2' R a_2 = 0$$

Como a_1, a_2 y a_3 son perpendiculares, $a_2' a_1 = 0$ y $a_2' a_3 = 0$, y como $a_2' R a_2 = Var[U_2] \neq 0$ tenemos:

$$0 - 0 - 0 - \mu_2 Var[U_2] = 0 \Rightarrow \mu_2 = 0$$

Entonces:

$$2Ra_3 - 2\lambda_3 I a_3 - \mu_1 R a_1 - \mu_2 R a_2 = 0 \Rightarrow 2Ra_3 - 2\lambda_3 a_3 = 0 \Rightarrow (R - \lambda_3) a_3 = 0$$

Por lo que a_3 es el valor asociado a $\lambda 3$.

2. Análisis Factorial

$$X_{p\times 1} = A_{p\times k}F_{k\times 1} + L_{p\times 1}$$

$$X = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{p1} & \dots & a_{pk} \end{pmatrix} \begin{pmatrix} F_1 \\ \vdots \\ F_k \end{pmatrix} + L_{p\times 1}$$

Haciendo el desarrollo matricial:

$$X_1 = a_{11}F_1 + \cdots + a_{1k}F_k + L_1$$

$$X_2 = a_{21}F_1 + \cdots + a_{2k}F_k + L_2$$

Igualdad fundamental:

$$\Sigma = E[XX'] = AA' + D$$

Demostración:

$$E[XX'] = E[(AF + L)(AF + L)'] = E[(AF + L)(F'A' + L')] = E[AFF'A' + AFL' + LF'A' + LL'] =$$

$$= E[AFF'A'] + E[AFL'] + E[LF'A'] + E[LL'] = AE[FF']A' + AE[FL'] + E[LF']A' + E[LL'] =$$

$$= AI_kA' + A0 + 0A' + D = AA' + D = \Sigma$$

De la expresión:

$$\Sigma = AA' + D$$

tenemos:

$$\begin{pmatrix} \sigma_{11} & \dots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \dots & \sigma_{pp} \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{p1} & \dots & a_{pk} \end{pmatrix} \begin{pmatrix} a_{11} & \dots & a_{p1} \\ \vdots & \ddots & \vdots \\ a_{1k} & \dots & a_{pk} \end{pmatrix} + \begin{pmatrix} d_1 \\ \vdots \\ d_p \end{pmatrix}$$

Por lo que:

$$\sigma_{11} = \sum_{j=1}^{k} a_{1j}^2 + d_1$$

$$\vdots$$

$$\sigma_{ii} = \sum_{j=1}^{k} a_{ij}^2 + d_i$$

$$\sigma_{ij} = \sum_{l=1}^{k} a_{il}a_{lj}$$

2.1. Ejemplo

$$X = (X_1, X_2, X_3) \Rightarrow \Sigma = S = \begin{pmatrix} 1 & 0{,}83 & 0{,}78 \\ 0{,}83 & 1 & 0{,}67 \\ 0{,}78 & 0{,}83 & 1 \end{pmatrix}$$

Ajustar el modelo factorial con $K = 1$:

$$X_{3\times 1} = A_{3\times 1}F_{1\times 1} + L_{3\times 1}$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} F_1 + \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix}$$

Tenemos que resolver:

$$\Sigma = AA^t + D$$

$$\begin{pmatrix} 1 & 0{,}83 & 0{,}78 \\ 0{,}83 & 1 & 0{,}67 \\ 0{,}78 & 0{,}67 & 1 \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & a_{31} \end{pmatrix} + \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix}$$

De donde obtenemos las siguientes 6 ecuaciones con 6 incógnitas:

$$\begin{aligned} a_{11}^2 + d_1 &= 1 \\ a_{21}^2 + d_2 &= 1 \\ a_{31}^2 + d_3 &= 1 \\ a_{11}a_{21} &= 0{,}83 \\ a_{11}a_{31} &= 0{,}78 \\ a_{21}a_{31} &= 0{,}67 \end{aligned}$$

Por lo que el modelo queda ajustado:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0{,}983 \\ 0{,}844 \\ 0{,}793 \end{pmatrix} F + \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix}$$

Ajustar el modelo factorial con $K = 2$:

$$X_{3\times 1} = A_{3\times 2}F_{2\times 1} + L3 \times 1$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} + \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix}$$

Tenemos que resolver:

$$\Sigma = AA^t + D$$

$$\begin{pmatrix} 1 & 0{,}83 & 0{,}78 \\ 0{,}83 & 1 & 0{,}67 \\ 0{,}78 & 0{,}67 & 1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{pmatrix} + \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix} =$$

$$= \begin{pmatrix} a_{11}^2 + a_{12}^2 + d_1 & a_{11}a_{21} + a_{12}a_{22} & a_{11}a_{31} + a_{12}a_{32} \\ a_{21}a_{11} + a_{22}a_{12} & a_{21}^2 + a_{22}^2 + d_2 & a_{21}a_{31} + a_{22}a_{32} \\ a_{31}a_{11} + a_{32}a_{12} & a_{31}a_{21} + a_{32}a_{22} & a_{31}^2 + a_{32}^2 + d_3 \end{pmatrix}$$

De donde obtenemos las siguientes 9 ecuaciones con 9 incógnitas:

$$\begin{aligned} a_{11}^2 + a_{12}^2 + d_1 &= 1 \\ a_{21}^2 + a_{22}^2 + d_2 &= 1 \\ a_{31}^2 + a_{32}^2 + d_3 &= 1 \\ a_{11}a_{21} + a_{12}a_{22} &= 0{,}83 \\ a_{21}a_{11} + a_{22}a_{12} &= 0{,}83 \\ a_{11}a_{31} + a_{12}a_{32} &= 0{,}78 \\ a_{31}a_{11} + a_{32}a_{12} &= 0{,}78 \\ a_{21}a_{31} + a_{22}a_{32} &= 0{,}67 \\ a_{31}a_{21} + a_{32}a_{22} &= 0{,}67 \end{aligned}$$

2.1.1. Rotaciones

G isometría $\Rightarrow G^{-1} = G' \Rightarrow GG' = I_k$. Llamando $B = AG$, tenemos:

$$\Sigma = AGG'A' + D = BB' + D$$

$$X = AF + L = AGG'F + L = BF_G + L$$

donde $F_G = G'F$.

3. Práctica final

Para los valores perdidos, creamos una nueva columna que sea *NombreColumna_NA*, en la que tendremos 0 si el valor existe, o 1 si el valor es NA.

Luego comparamos esta columna con otra columna previamente existente, y podemos comparar la media de esta columna cuando el valor es 0 con respecto a la media cuando el valor es 1 realizando un test como el de Student o el de Wiluxom.

$$\begin{cases} H_0 \equiv \mu Variable1 = \mu Variable2 \\ H_1 \equiv \mu Variable1 \neq \mu Variable2 \end{cases}$$

Si tomamos un dataset sin valores perdidos, podemos tomar una variable y eliminar el 5 % de los datos para realizar el análisis.

Realizamos el apartado 1 d) antes del 1 c).

Contraste de normalidad:

$$\begin{cases} H_0 \equiv \text{ Siguen normalidad} \\ H_1 \equiv \text{ No siguen normalidad} \end{cases}$$

Para ver si hay correlación, aunque hay que hacer el test de Barlett, también hay que poner salidas gráficas, como mapas de calor.

En materiales y métodos, si tratamos con personas tiene sentido poner la media y desviación de edades, y la distribución de sexo.

En la sección de resultados tenemos que aportar resultados de forma objetiva, sin ninguna opinión, y en la discusión asociamos estos resultados a los datos.

4. Análisis Discriminante

$Y = \{1, \ldots, k\}$ cualitativa, $X_1, \ldots, X_n$ continuas.
Clasificar en los niveles de Y según los valores $(X_1, \ldots, X_n)$.

$$P(^{Y=j}/_{X=x}) = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

Utilizamos el Teorema de Bayes: Sean A y B dos sucesos.

$$P(^{A}/_{B}) = \frac{P(^{A}/_{B}) \cdot P(B)}{P(A)} = \frac{\frac{P(A \cap B)}{P(B)} \cdot P(B)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

ADL con un sólo regresor:

$$P[^{Y=K}/_{X=x}] = \frac{P[^{X=x}/_{Y=k}] \cdot P[Y = k]}{P[X = x]}$$

Llamamos $\pi_k = P[Y = k]$.

$$P[^{Y=k}/_{X=x}] = \frac{P[^{X=x}/_{Y=k}] \cdot \pi_k}{P[X = x]} = \frac{P[^{X=x}/_{Y=k}]\pi_k}{\sum_{j=1}^{K} \pi_j P[^{X=x}/_{Y=j}]}$$

Asumimos que $^{X}/_{Y=k} \sim N(\mu_k, \sigma_k) \Rightarrow$

$$\Rightarrow f_{X/_{Y=k}}(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\{-\frac{(x-\mu_k)^2}{2\sigma_k^2}\}$$

Supongamos además ahora homogeneidad de varianzas:

$$\sigma_1 = \cdots = \sigma_k = \sigma$$

Supongo que $Y = \{1, 2\}$. Calcular $\log(\frac{P[^{Y=1}/_{X=x}]}{P[^{Y=2}/_{X=x}]})$.

$$P[^{Y=1}/_{X=x}] = \frac{P[^{X=x}/_{Y=1}]\pi_1}{\pi_1 P[^{X=x}/_{Y=1}] + \pi_2 P[^{x=x}/_{Y=2}]}$$

$$P[^{Y=2}/_{X=x}] = \frac{P[^{X=x}/_{Y=2}]\pi_2}{\pi_1 P[^{X=x}/_{Y=1}] + \pi_2 P[^{x=x}/_{Y=2}]}$$

$$\log(\frac{P[^{Y=1}/_{X=x}]}{P[^{Y=2}/_{X=x}]}) = \log(P[^{Y=1}/_{X=x}]) - \log(P[^{Y=2}/_{X=x}]) =$$

$$= \log(\frac{\pi_1}{\sqrt{2\pi}\sigma} e^{-\frac{x-\mu_1)^2}{2\sigma^2}}) - \log(\frac{\pi_2}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}) =$$

$$= \log(\frac{\pi_1}{\pi_2}) - \frac{(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_2)^2}{2\sigma^2} = \frac{x(\mu_1 - \mu_2)}{\sigma^2} + \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} + \log(\frac{\pi_1}{\pi_2})$$

Si la expresión anterior es mayor que 0, es decir, si $\frac{P[^{Y=1}/_{X=x}]}{P[^{Y=2}/_{X=x}]}$ es mayor que 1, entonces clasificamos $Y = 1$, y si no, $Y = 2$.

Nota: Como $\frac{P[^{Y=1}/_{X=x}]}{P[^{Y=2}/_{X=x}]}$ es la división de dos probabilidades, el valor no tiene que estar en $[0, 1]$, por lo que no lo podemos llamar la probabilidad.

Test de Normalidad Univariante:

$$\begin{cases} H_0 \equiv \text{ Los datos siguen una normal} \\ H_1 \equiv \text{ Los datos no siguen una normal} \end{cases}$$

Test de Normalidad Multivariante:

$$\begin{cases} H_0 \equiv \text{ Hay normalidad multivariante} \\ H_1 \equiv \text{ No hay normalidad multivariante} \end{cases}$$

Test de Mordia, H-Z, Royston...

Para homogeneidad de varianzas, el test de Levene.

Si nos falla algún contraste, continuamos pero con el cuadrático en lugar del lineal ya que es más robusto.