Análise de Componentes Principais - PCA

Prf. Dr. Osvaldo Vargas Jaques

Ciência da Computação - UEMS

23 de Novembro de 2022

Tópicos

Variáveis

Uma variável aleatória X_i é o vetor linha

$$X_i = \left[\begin{array}{cccc} x_{i1} & x_{i2} & \cdots & x_{in} \end{array} \right] \quad (1)$$

Exemplo

$$X_2 = Peso = \left[\begin{array}{cccccc} 70 & 75 & 80 & 55 & 90 & 80 \end{array} \right], \quad (2)$$

X_2 tem $n = 6$ observações ou instâncias.

Variáveis

Um vetor de variáveis aleatórias X é o vetor

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \quad (3)$$

veja que existem p variáveis (linhas) e cada X_i tem n observações (colunas). Pode-se dizer que X é uma matriz pxn.

Exemplo

$$X = \begin{bmatrix} Altura \\ Peso \\ Cintura \end{bmatrix} = \begin{bmatrix} 1.70 & 1.54 & 1.90 & 1.55 & 1.70 & 1.75 \\ 70 & 75 & 80 & 55 & 90 & 80 \\ 36 & 49 & 40 & 38 & 29 & 35 \end{bmatrix}, \quad (4)$$

neste caso, $p = 3$ variáveis e $n = 6$ observações.

Variância e Covariância

Variância $\sigma^2_{X_i}$ é a medida de variação de uma variável aleatória X_i. Mede o quanto os valores de X_i estão dispersos.

$$\sigma^2_{X_i} = \frac{1}{n}\sum(X_i - \mu_{X_i})^2 = \frac{1}{n}\left[X_i - \mu_{X_i}\right]\left[X_i - \mu_{X_i}\right]', \tag{5}$$

sendo μ_X a média de X e $[X_i - \mu_{X_i}]'$ o transposto de $[X_i - \mu_{X_i}]$.

A raiz quadrada de σ^2_X é o desvio padrão

$$\sigma_X = \sqrt{\frac{1}{n}\sum(X - \mu_X)^2} \tag{6}$$

Variância e Covariância

A covariância σ_{XY} mede o quão fortemente duas variáveis X e Y estão relacionadas.

$$\sigma_{XY} = \frac{1}{n}\sum(X - \mu_X)(Y - \mu_Y) \tag{7}$$

Se $X = Y$, $\sigma_{XX} = \sigma_X^2$, que é a variância de X .

Matriz de Covariância

Na Matriz de Covariância K_{pxp} tem os elementos $C_{X_iX_j}$, sendo X_i e X_j elementos de um vetor de variáveis aleatórias.

Dado o vetor X do exemplo anterior

$$K = \begin{bmatrix} C_{AlturaAltura} & C_{Altura,Peso} & C_{Altura,Cintura} \\ C_{PesoAltura} & C_{PesoPeso} & C_{PesoCintura} \\ C_{CinturaAltura} & C_{CinturaPeso} & C_{CinturaCintura} \end{bmatrix}. \tag{8}$$

$$K = \begin{bmatrix} C_{X_1X_1} & C_{X_1X_2} & \cdots & C_{X_1X_p} \\ C_{X_2X_1} & C_{X_2X_2} & \cdots & C_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{X_pX_1} & C_{X_pX_2} & \cdots & C_{X_pX_p} \end{bmatrix}. \tag{9}$$

Na diagonal principal de K temos a variância de X_i.

Matriz de Covariância aplicando operações matriciais

Dada uma *matriz de dados* $A_{n,p}$, tendo p atributos e n observações, o que temos que fazer é encontrar a média de cada coluna de A, através de algum software próprio para isso, como por exemplo o matlab, scilab, octave ou python. Seja $\vec{\mu}_a = [\mu_1 \ \mu_2 \ \cdots \ \mu_p]$ o vetor linha de médias de cada coluna de A.

Para calcular os desvios a partir da média primeiro precisamos de uma matriz com n linhas do vetor de médias. Seja $\vec{b} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ um vetor com n termos igual a 1.

Matriz de Covariância aplicando operações matriciais

Fazendo

$$M = \vec{b}\vec{\mu}_a = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} \mu_1 & \mu_2 & \dots & \mu_p \end{bmatrix} = \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_p \\ \mu_1 & \mu_2 & \cdots & \mu_p \\ \vdots & \vdots & \ddots & \vdots \\ \mu_1 & \mu_2 & \cdots & \mu_p \end{bmatrix} \tag{10}$$

teremos uma matriz M de n linhas de $\vec{\mu}_a$.

Assim, podemos ter a matriz de covariância C, fazendo

$$C = \frac{1}{n}(A - M)'(A - M) \tag{11}$$

Coeficiente de Correlação

Quando as unidades de medidades das variáveis são diferentes o melhor é usar o coeficiente de correlação que padroniza as variáveis. O coeficiente de correlação entre X e Y é

$$\rho_{XY} = \frac{1}{n} \sum \frac{(X - \mu_X)}{\sigma_X} \frac{(Y - \mu_Y)}{\sigma_Y}. \tag{12}$$

O coeficiente de correlação varia de -1 a 1. Se $X = Y$, $\rho_{XX} = 1$.

Matriz de Correlação

Na Matriz de Correlação R_{pxp} tem os elementos $\rho_{X_iX_j}$, sendo X_i e X_j elementos de um vetor de variáveis aleatórias.

$$R = \frac{1}{n}\sum \frac{\left(X_i - \mu_{X_i}\right)}{\sigma_{X_i}} \frac{\left(X_j - \mu_{X_j}\right)}{\sigma_{X_j}} = \rho_{X_i}\rho_{X_j}. \tag{13}$$

Dado o vetor X da equação 4

$$R = \begin{bmatrix} 1 & \rho_{AlturaPeso} & \rho_{AlturaCintura} \\ \rho_{PesoAltura} & 1 & \rho_{PesoCintura} \\ \rho_{CinturaAltura} & \rho_{CinturaPeso} & 1 \end{bmatrix}. \tag{14}$$

Matriz de correlação

$$R = \begin{bmatrix} \rho_{X_1X_1} & \rho_{X_1X_2} & \cdots & \rho_{X_1X_p} \\ \rho_{X_2X_1} & \rho_{X_2X_2} & \cdots & \rho_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_pX_1} & \rho_{X_pX_2} & \cdots & \rho_{X_pX_p} \end{bmatrix}. \tag{15}$$

Na diagonal principal de R temos $\rho_{X_iX_i} = 1$.

Algumas Propriedades Matriciais

Seja uma matriz X e uma matriz Y, tal que Y seja uma transformada de X. Ou seja,

$$Y = BX \tag{16}$$

Então são válidas as seguintes propriedades:

$$\mu_Y = B\mu_X \tag{17}$$

$$K_Y = BK_XB' \tag{18}$$

onde μ é a média das observações, K_Y e K_X, as matrizes de covariâncias.

Algumas Propriedades Matriciais

Seja uma matriz quadrada A_{pxp} então esta matriz possui um escalar λ e um vetor $\overrightarrow{v}$ tal que:

$$A\overrightarrow{v} = \lambda\overrightarrow{v} \tag{19}$$

λ e $\overrightarrow{v}$ são chamados de autovalor e autovetor de A.
Sendo D uma matriz com os autovalores em diagonal e V o conjunto de autovetores associados, valem as seguintes propriedades:

$$AV = VD \tag{20}$$

Algumas Propriedades Matriciais

Sendo $V = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \cdots & \vec{e}_p \end{bmatrix} = \begin{bmatrix} e_{11} & e_{21} & \cdots & e_{p1} \\ e_{12} & e_{22} & \cdots & e_{p1} \\ e_{13} & e_{23} & \cdots & e_{p1} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1p} & e_{2p} & \cdots & e_{pp} \end{bmatrix}$

e $D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$. Repare que cada vetor $\vec{e}_i$ é um vetor coluna, ou simplesmente vetor.

Decomposição Espectral

O espectro de uma matriz é o conjunto de seus autovalores. Um teorema muito importante em álgebra matricial e de fundamental importância em estatística multivariada é o da decomposição espectral, que relaciona a matriz com seus autovalores e autovetores.

Seja K_{pxp} uma matriz de covariâncias e uma matriz ortogonal O_{pxp}, isto é $O^{'}O = OO^{'} = I_{pxp}$, tal que:

$$KO = O\Lambda \tag{21}$$

Decomposição Espectral

Logo, vemos que (21), é semelhante a (20). Multiplicando (21) por O' pela esquerda, resulta no teorema abaixo.

Para toda matriz de covariância K_{pxp} existe uma matriz ortogonal tal que:

$$O'KO = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} = \Lambda \tag{22}$$

onde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ são os autovalores ordenados em ordem decrescente da matriz K_{pxp}. Neste caso, dizemos que a matriz K_{pxp} é similar à matriz Λ.

Decomposição Espectral

Isso implica em dizer que:

I. $\vec{e_i}'\vec{e_i} = 1$

II. $\vec{e_i}'\vec{e_j} = 0,\ i \neq j$

III $det(K) = |K| = \prod_{i=1}^{p} \lambda_i$

IV $traço(K) = \sum_{i=1}^{p} \lambda_i$

Decomposição Espectral

A i-ésima coluna da matriz O é o autovetor normalizado $\vec{e}_i$, corresponde ao autovetor λ_i, que é denotado por $\vec{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{bmatrix}$

Então a matriz O_{pxp} é dada por $O = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \cdots & \vec{e}_p \end{bmatrix}$ e pelo teorema de decomposição espectral, a matriz de covariância pode ser ***decomposta*** como uma soma de várias matrizes $M_{i_{(pxp)}} = \vec{e}_i \vec{e}_i{}'$ multiplicadas pelos seus autovalores λ_i correspondentes. Ou seja, da propriedade (22), resulta que:

$$\begin{aligned} K_X = O \Lambda O' = \sum_{i=1}^{p} \lambda_i \vec{e}_i \vec{e}_i{}' &= \lambda_1 \vec{e}_1 \vec{e}_1{}' + \lambda_2 \vec{e}_2 \vec{e}_2{}' + \cdots + \lambda_p \vec{e}_p \vec{e}_p{}' \\ &= \lambda_1 M_1 + \lambda_2 M_2 + \cdots + \lambda_p M_p \end{aligned} \tag{23}$$

Transformada Karhunen-Loève

Queremos obter um vetor de variáveis aleatórias Y tal que

$$Y = O'X, \tag{24}$$

sendo O o conjunto de autovetores ortogonais de K_X correspondente aos autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ e O' o conjunto transposto. Assim, de (21) $K_X O = O\Lambda$ e de (22)

$$O' K_X O = \Lambda \tag{25}$$

Vemos que (25) corresponde a propriedade (18). Logo,

$$K_y = O' K_X O = \Lambda \tag{26}$$

Transformada Karhunen-Loève

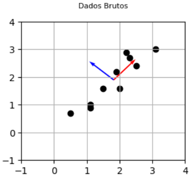
O vetor Y é composto de p combinações lineares das variáveis aleatórias do vetor X, tem vetor de médias igual a $O'\mu$ e matriz de covariâncias Λ_{pxp} que é uma matriz diagonal, cujos elementos $a_{ii} = \lambda_i$, isto é,

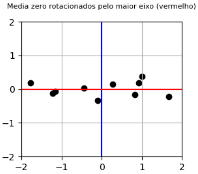
$$\Lambda_{pxp} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} \tag{27}$$

Portanto, as variáveis aleatórias que constituem o vetor Y não são correlacionadas entre si. Deste modo, surge a idéia de utilizar as combinações lineares em Y como uma forma alternativa de se representar a estrutura de covariâncias do vetor X.

Transformada Karhunen-Loève

Os vetores aleatórios, X e Y, tem a mesma variância total e a mesma variância generalizada, sendo que o vetor Y tem a vantagem de ser composto por variáveis aleatórias não correlacionadas, facilitando portanto, a interpretação conjunta dessas. Esta operação é também conhecida como Karhunen-Loève Transform (KLT). Resumidamente, a KLT serve para rotacionar os eixos por ordem decrescente de variância. Em processamento de imagens, isto auxiliaria na identificação de formas, uma vez que estas estarão orientadas decrescentemente por variância de seus eixos.





Transformada Karhunen-Loève

Sabendo que

$$Y = O'X = \begin{bmatrix} \vec{e_1}' \\ \vec{e_2}' \\ \vdots \\ \vec{e_p}' \end{bmatrix} X = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ e_{31} & e_{32} & \cdots & e_{3p} \\ \vdots & \vdots & \ddots & \vdots \\ e_{p1} & e_{p2} & \cdots & e_{pp} \end{bmatrix} X \tag{28}$$

Vemos que a j-ésima componente principal da matriz K_{pxp} é definida como

$$Y_j = \vec{e_j}'X = e_{j1}X_1 + e_{j2}X_2 + \cdots + e_{jp}X_p \tag{29}$$

Transformada Karhunen-Loève

A proporção da variância total de X que é explicada pela i-ésima componente principal é definida como:

$$pv_i = \frac{\lambda_i}{\displaystyle\sum_{j=1}^{p} \lambda_j} \tag{30}$$

A proporção de variância total que é explicada pelas k primeiras componentes principais é definida por:

$$pvt_k = \frac{\displaystyle\sum_{i=1}^{k} \lambda_i}{\displaystyle\sum_{j=1}^{p} \lambda_j} \tag{31}$$

Usando as combinações lineares em Y, surge também a ideia de reduzir o espaço de variáveis, passando da dimensão p para a dimensão k, sendo $k < p$. Portanto,ao invés de utilizar o vetor aleatório original na análise de dados, utiliza-se k combinações lineares de componentes principais, pegando somente os autovetores mais significativos. Esta ordem de significância ou de explicação é dada pelos autovalores ordenados. Este processo é conhecido como Principal Components Anlysis (PCA) ou simplesmente Análise de Componentes Principais.

De (28), se tomarmos os $k < p$ autovalores, teríamos

$$Y = \begin{bmatrix} \vec{e_1}' \\ \vec{e_2}' \\ \vdots \\ \vec{e_k}' \end{bmatrix} X = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ e_{k1} & e_{k2} & \cdots & e_{kp} \end{bmatrix} X \tag{32}$$

Veja que foram retirados os vetores $\vec{e}_{k+1}, \cdots, \vec{e}_p$.

Da operação (32) vemos que que resultaria em um vetor Y_{kxn}, reduzindo assim a dimensão dos dados. Ou seja, ao invés de p variáveis, lidaremos agora com $k < p$ variáveis.

A redução das dimensões facilita a análise de dados. Diferentes critérios podem ser usado na escolha desse k mas um dos modos mais comuns é escolher um k tal que $pvt_k \geq 70\%$.

Geralmente como a PCA e KLT estão interligados, o termo PCA é usado como um sinônimo de ambos. Mas, na realidade, a PCA utiliza a KLT e a Teoria da Decomposição Espectral para reduzir a dimensão dos dados.

Referências

- MINGOTI, Sueli A. **Análise de dados através de estatística multivariada: uma abordagem aplicada.** Belo Horizonte, Editora UFMG, 2005.

- LATTIN, James; CARROL, Douglas C.; GREEN, Paul E. **Análise de dados multivariados.** São Paulo, Cencage Learning, 2011.