PCA via álgebra linear

PEDRO PATRÍCIO

Sup. que temos n amostras, cada uma com m dados

P. ex., 30 pessoas, cada uma correspondente a 3-uplo com peso, altura, QI.

1) Visualize: como se distribuem as amostras?
2) Que variáveis estão correlacionadas?
3) Quais as variáveis que melhor descrevem o espaço amostral

→ redução de dimensão!
→ detetar de "outliers"

Revisões de álgebra linear

Seja $A_{n\times n}$, define-se o polinómio característico de A como $\Delta_A(\lambda) = |\lambda I - A|$ onde $|\Pi|$ indica o determinante de Π

As raízes de $\Delta_A(\lambda)$ chamam-se valores próprios de A.

Se λ é v.p. de A então $(\lambda I - A)v = 0$ é um sistema possível indeterminado. Ou seja, existe vetor $v \neq 0$ tal que $(\lambda I - A)v = 0$.

Ou seja, para o valor próprio λ, existe $v \neq 0$ tal que $Av = \lambda v$.

v diz-se vector próprio de A associado a λ.

NOTA: mesmo que a matriz seja real, pode ter valores próprios complexos. Por exemplo, $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ tem valores próprios $\lambda = \pm i$

$v = \begin{bmatrix} -i \\ 1 \end{bmatrix}$ é vect. próppr. assoc. a $\lambda = i$

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

Defn. H diz-se hermítica se $H^* = H$, onde $\overline{H^t}$ indica a transposta dos conjugados de H.

Prop. Seja H hermítica.

a) $\sigma(H) \subseteq \mathbb{R}$ onde $\sigma(H)$ indica o conj^to dos valores próprios de H (chamado espectro de H)

b) vectores próprios assoc. a valores próprios distintos são ortogonais 2 a 2.

c) H é diagonalizável à custa de uma matriz unitária ortogonal. i.e., existe $U = (U^{-1})^*$ e D diagonal tal que

$$H = UDU^{-1} \quad (\text{ou equiv. } D = U^{-1}HU)$$

dem a) ~~$\Delta_H(\lambda) = |\lambda I - H| = |(\lambda I - H)^*|$~~

$Hv = \lambda v \Rightarrow v^*(Hv) = v^*(\lambda v) = \lambda v^* v = \lambda \|v\|^2$

~~$v^*(Hv) = (Hv)^* v$~~ Ou seja $v^*(Hv) = \lambda \|v\|^2$

Conjugando ambos os lados $(v^*(Hv))^* = \bar{\lambda} \|v\|$

$\Rightarrow (Hv)^* v = \bar{\lambda} \|v\| \Rightarrow v^* H^* v = \bar{\lambda} \|v\|$

$\Rightarrow v^* H v = \bar{\lambda} \|v\|$

Logo, $\bar{\lambda} \|v\| = \lambda \|v\|$. Como $v \neq 0$ então $\bar{\lambda} = \lambda$, e portanto $\lambda \in \mathbb{R}$

b) $Hv_i = \lambda_i v_i$

$Hv_j = \lambda_j v_j$ Com $\lambda_i \neq \lambda_j$

$v_i^* H v_j = v_i^* \lambda_j v_j = \lambda_j v_i^* v_j$

$v_i^* H v_j = v_i^* H^* v_j = (Hv_i)^* v_j = \lambda_i v_i^* v_j$

$\therefore \lambda_j v_i^* v_j = \lambda_i v_i^* v_j \underset{\lambda_i \neq \lambda_j}{\Rightarrow} v_i^* v_j = 0$

Nota importante: Vectores próprios associados a valores próprios $\neq$s são ortogonais, e formam uma base ortogonal do espaço.

A partir deste momento, supomos que as matrizes são reais.

Teor Se S é simétrica ($S^T = S$) então S é ortogonalmente diagonalizável, $\sigma(S) \subseteq \mathbb{R}$ e os vectores próp. associ. val. pr. ≠s são ortogonais 2a2.

Ou seja, existe $U = (U^{-1})^T$ e D diagonal tais que $S = UDU^{-1}$

As colunas de U são os vectores próprios e D a matriz diagonal com os valores próprios pela mesma ordem escolhida para U.

Defn. Uma matriz simétrica S diz-se ~~tinal~~ semi-definida positiva (SDP) se $\langle v, Sv \rangle \geq 0$, $\forall v$

[~~positiva~~ definida positiva se $\langle v, Sv \rangle > 0$, $\forall v \neq 0$]

Teor. S SDP $\Rightarrow \sigma(S) \subseteq \mathbb{R}_0^+$

S DP $\Rightarrow \sigma(S) \subseteq \mathbb{R}^+$

dem Seja $\lambda \in \sigma(S)$. $0 \leq \langle v, Sv \rangle = \langle v, \lambda v \rangle = \lambda \langle v, v \rangle = \lambda \|v\|^2$

Seja v vect. próp. assoc. λ ($\lambda \in \mathbb{R}$)

$\Rightarrow \lambda \geq 0$ por $v \neq 0$

NOTA $A_{m\times n}$ então AA^T e A^TA são simétricas.

Prop AA^T e A^TA têm os mesmos val. próp. não nulos.

dem $\lambda \neq 0$, $v \neq 0$ t.q. $A^TA v = \lambda v \Rightarrow AA^T(Av) = \lambda(Av)$

Ou seja Av é vect. próp. de AA^T assoc. v.p. λ

Note-se que $Av \neq 0$, senão $A^T(Av) = 0$ e λ seria 0.

Ou seja, λ v.p. A^TA e tb. v.p. AA^T.

Mais, se v é vect.p. A^TA então Av é vect. p. de AA^T.

— □

NOTA Sp A é uma matriz 1000×5.

AA^T é uma matriz 1000×1000.

P/ calcular os valores próp. de AA^T basta calcular os val.p. A^TA, matriz 5×5. Os restantes 995 são nulos!

Coro **Prop** AA^T é SDP

dem $\langle v, AA^T v\rangle = v^T AA^T v = (A^T v)^T (A^T v) = \|A^T v\|^2 \geq 0$

Coro! $\sigma(AA^T) \subseteq \mathbb{R}_0^+$, $\sigma(A^TA) \subseteq \mathbb{R}_0^+$

Um pouco de STATS

Seja A uma variável aleatória medida n vezes, obtendo $a_1, \dots, a_n$.

A média de A : $\mu_A = \frac{1}{n}(a_1 + \dots + a_n)$

(Já falamos distinção entre média e média amostral)

As medições como está distribuídas?

Variância : $\text{var}(A) = \frac{1}{n} \sum_{i=1}^{n} (a_i - \mu_A)^2$

desvio-padrão : $\sigma = \sqrt{\text{Var}(A)}$

Sup que tmos 2 variáveis A, B

$$\text{Cov}(A,B) = \frac{1}{n}\left((a_1-\mu_A)(b_1-\mu_B) + \dots + (a_n-\mu_A)(b_n-\mu_B)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(a_i-\mu_A)(b_i-\mu_B)$$

Sup agora que tmos m variáveis relativas a n leituras $x_1, \dots, x_n$, $x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{im} \end{bmatrix}$ x_{ij} indica o que se obteve no "indivíduo" x_i na variável j

É habitual "recentrar" os dados para a média

$$\mu = \frac{1}{n}(x_1 + \dots + x_n).$$

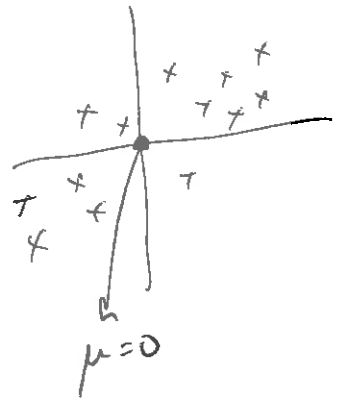
$$B = \begin{bmatrix} | & | & & | \\ x_1-\mu & x_2-\mu & \cdots & x_n-\mu \\ | & | & & | \end{bmatrix}_{m \times n}$$

Matriz de covariância $S = \frac{1}{n} BB^T$

A entrada (i,j) de S é a covariância entre as variáveis i e j

Minimizando os resíduos

Começemos por projectar os dados $\{x_i\}$ num esp. dim 1.



Basta encontrar ~~uma base~~, uma direcção. Para simplificar, suponhamos que os dados estão centrados.

I.e., $\mu_A = 0$

(Para tal, dados $x_1, \dots, x_n$ e média μ basta considerar $x_1-\mu, x_2-\mu, \dots, x_n-\mu$)

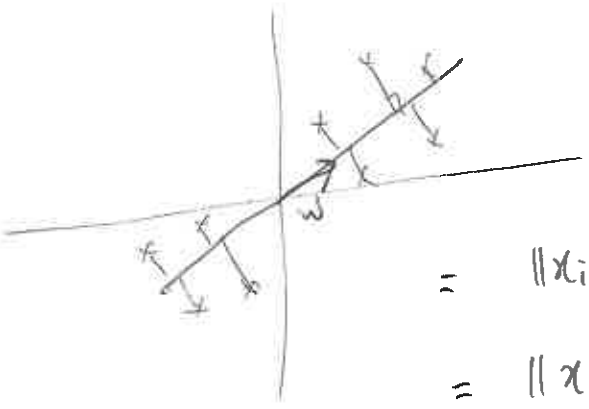
Pretendemos encontrar w c/ $\|w\|=1$ sobre o ~~qual~~ $\langle w \rangle$ projectamos x_i. $\text{proj}_w x_i = (x_i \cdot w)\, w$.

Cálculo da média das projecções: $\frac{1}{n}\sum_{i}^{n} (x_i \cdot w)\, w$

$$= \Big(\underbrace{\Big(\frac{1}{n}\sum x_i\Big)}_{=0} \cdot w \Big) w = 0$$

Ou seja, a média das proj tb é 0.

Quanto vale w por forma a minimizar as distâncias?



Para cada x_i, temos a distância ao quadrado

$$\|x_i - (w\cdot x_i) w\|^2 =$$

$$= \|x_i\|^2 - 2(w\cdot x_i)(w\cdot x_i) + \|w\|^2$$

$$= \|x_i\|^2 - 2(w\cdot x_i)^2 + 1$$

A soma dos resíduos será

$$\text{Res}(w) = \sum_{i=1}^{n} \left(\|x_i\|^2 - 2(w\cdot x_i)^2 + 1 \right) = \left(n + \sum \|x_i\|^2 \right) - 2\sum (w x_i)^2$$

Ora $n + \sum \|x_i\|^2$ não depende de w.

Para minimizarmos $Res(w)$ temos que MAXIMIZAR

$$\sum (w \cdot x_i)^2$$

NOTA maximizar $\sum (w \cdot x_i)^2$ é o mesmo que maximizar

$\frac{1}{n} \sum (w \cdot x_i)^2$, i.e., média de $(w \cdot x_i)^2$

Consideremos $X = \begin{bmatrix} - x_1^T - \\ - x_2^T - \\ \vdots \\ - x_n^T - \end{bmatrix}_{n \times p}$ ($X = B^T$ da pag. 6 !!!)

Então $Xw = \begin{bmatrix} x_1 \cdot w \\ x_2 \cdot w \\ \vdots \\ x_n \cdot w \end{bmatrix}$ e $(Xw)^T = w^T X = [w \cdot x_1 \ \cdots \ w \cdot x_n]$

$$\frac{1}{n} \sum (w \cdot x_i)^2 = \frac{1}{n} (Xw) = \frac{1}{n} [w \cdot x_1 \ w \cdot x_2 \cdots w \cdot x_n] \begin{bmatrix} x_1 \cdot w \\ x_2 \cdot w \\ \vdots \\ x_n \cdot w \end{bmatrix}$$

$$= \frac{1}{n} (Xw)^T (Xw) = \frac{1}{n} w^T (X^T X) w$$

$$= w^T V w \quad \text{c/ } V = \frac{1}{n} X^T X$$

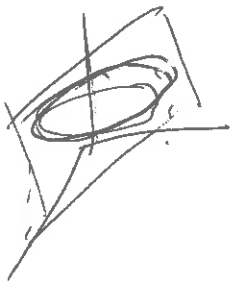
Queremos assim

$$\begin{cases} \text{maximizar } w^T V w \\ \text{sujeito a } \|w\|^2 = 1 \end{cases}$$

Revisão Multiplicadores de LAGRANGE

Exemplo. $\max f(x,y) = x+y$

sujeito a $x^2+y^2 = 1$



$$\mathcal{L}(x,y,\lambda) = f(x,y) - \lambda(x^2+y^2-1)$$
$$= x+y-\lambda(x^2+y^2-1)$$

$\nabla \mathcal{L} = 0$

$\frac{\partial \mathcal{L}}{\partial x} = 0 \Rightarrow 1-2\lambda x = 0 \Rightarrow x = \frac{1}{2\lambda}$

$\frac{\partial \mathcal{L}}{\partial y} = 0 \Rightarrow 1-2\lambda y = 0 \Rightarrow y = \frac{1}{2\lambda}$

$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \Rightarrow x^2+y^2 = 1 \Rightarrow \lambda = \pm\frac{\sqrt{2}}{2}$

$\mathcal{L}(w,\lambda) = f(w) - \lambda(g(w)-c)$ onde $f(w) = w^T V w$

$g(w) = \|w\|^2 = w \cdot w$

$c = 1$

$0 = \frac{\partial \mathcal{L}}{\partial w} \Rightarrow \frac{\partial f}{\partial w} - \lambda \frac{\partial g}{\partial w}$

$0 = \frac{\partial \mathcal{L}}{\partial \lambda} = -(g(w)-c)$

Como calcular $\frac{\partial (w^T V w)}{\partial w}$?

Caso simples 2x2:

$$w^T V w = [w_1\ w_2]\begin{bmatrix} v_{11} & v_{12} \\ v_{12} & v_{22} \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = [w_1\ w_2]\left(w_1\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} + w_2\begin{bmatrix} v_{12} \\ v_{22} \end{bmatrix}\right)$$

$$= [w_1^2\ \ w_1 w_2]\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} + [w_1 w_2\ \ w_2^2]\begin{bmatrix} v_{12} \\ v_{22} \end{bmatrix}$$

$\frac{\partial w^T V w}{\partial w_1} = 2 w_1 v_{11} + w_2 v_{12} + w_2 v_{12}$

$\frac{\partial w^T V w}{\partial w_2} = w_1 v_{12} + w_1 v_{12} + 2 w_2 v_{22}$

$$\begin{bmatrix} \frac{\partial(\ )}{\partial w_1} \\ \frac{\partial(\ )}{\partial w_2} \end{bmatrix} = \begin{bmatrix} 2w_1v_{11} + 2w_2v_{12} \\ 2w_1v_{12} + 2w_2v_{22} \end{bmatrix} = 2 \begin{bmatrix} v_{11} & v_{12} \\ v_{12} & v_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= 2Vw$$

$$\frac{\partial(w^TVw)}{\partial w} = 2Vw$$

Pela mesma razão, $\frac{\partial(w^Tw)}{\partial w} = 2w$

Portanto $\frac{\partial \mathcal{L}}{\partial w} = 2Vw - 2\lambda w = 0$

$\Rightarrow Vw = \lambda w$ Portanto

w é vector de V assoc val.p λ

$w^TVw = \lambda w^Tw = \lambda \|w\| = \lambda$

Temos assim que escolher o <u>MAIOR</u> valor próprio.

$V_{p\times p}$ SDP, logo diagon., $r(V) \subseteq \mathbb{R}_0^+$

<u>vectores próprios</u> 2 a 2 ortogonais

Componentes principais

A 1ª componente principal (vect prop assoc maior val.p.) indica a direcção de maior variância.

(e assim por diante)