SVD no PCA

PEDRO PATRÍCIO

Recorde que temos n "indivíduos", $x_1, \dots, x_n$.
Cada um com m "inputs" ou atributos.
Assumimos que os dados estão centrados, i.e., c/ média 0.
Podemos escrever o indivíduo i como

$$x_i = \begin{bmatrix} x_i^{(1)} \\ \vdots \\ x_i^{(m)} \end{bmatrix}$$ onde $x_i^{(j)}$ denota

o input do atributo j em x_i.

O atributo j toma, assim, os valores

$$x^{(j)} = \left(x_1^{(j)}, x_2^{(j)}, \dots, x_n^{(j)} \right)$$

Se considerarmos a matriz B cujas colunas são os vectores x_i, então $x^{(j)}$ será a linha j de B.

A matriz de covariância é

$$S = \frac{1}{n} BB^T = \begin{bmatrix} \mathrm{Cov}(x^{(1)}, x^{(1)}) & \mathrm{Cov}(x^{(1)}, x^{(2)}) & \\ \mathrm{Cov}(x^{(2)}, x^{(1)}) & \mathrm{Cov}(x^{(2)}, x^{(2)}) & \\ & & \ddots & \\ & & & \mathrm{Cov}(x^{(m)}, x^{(m)}) \end{bmatrix}$$

cuja entrada (i,j) iguala $\mathrm{Cov}(x^{(i)}, x^{(j)})$
$\left(= \mathrm{Cov}(x^{(j)}, x^{(i)}) \right)$

<u>Recorde ainda</u>: 1) MM^T é SDP, i.e., é simétrica e $\sigma(MM^T) \subseteq \mathbb{R}_0^+$
(todos os valores próprios são reais não-negativos)

2) a soma dos valores próprios(*) de uma matriz iguala a soma dos elementos diagonais

(*) - contando multiplicidade

2.2

3) Se $M_{n\times n}$ tem valores próprios $\lambda_1, \dots, \lambda_n$, então $\alpha M_{n\times n}$ tem valores próprios $\alpha\lambda_1, \dots, \alpha\lambda_n$

Se v é vector prop. de M assoc. λ então v é vector prop. de αM assoc. $\alpha\lambda$.

Ou seja, os valores próprios são afectados por α, mas mantém os mesmos vectores próprios.

(repare que v vect. próprio então αv também é vector prop., $\forall \alpha \neq 0$)

4) MM^T e M^TM têm os mesmos valores próprios não nulos, contando com a multiplicidade.

Mais, se v é vector prop. de M^TM assoc. a λ então Mv é " " " MM^T assoc. a λ.

Se u é vect. prop. de MM^T assoc. a λ então M^Tu é " " " M^TM " " λ.

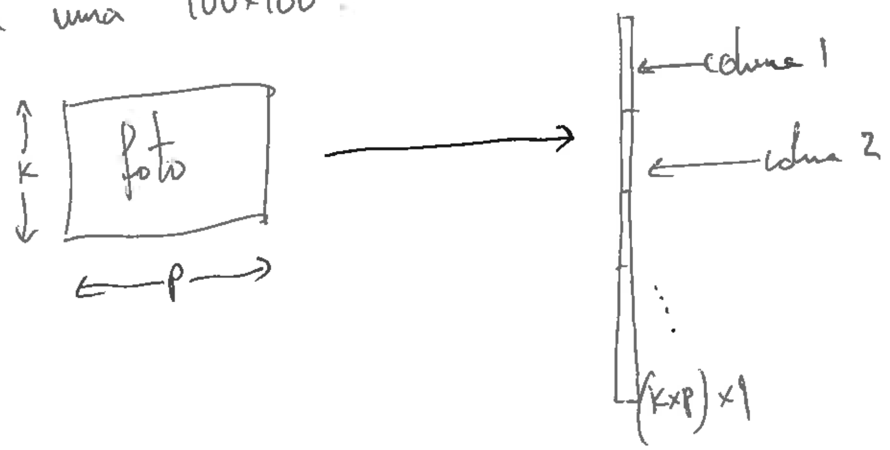
Este facto (4) pode ser muito relevante na capacidade de poder implementar o PCA.

No caso de termos n indivíduos, cada um com m atributos, obtemos a matriz de covariância

$$S = \frac{1}{n} BB^T \text{, de ordem } m \times m.$$

No caso de termos menos atributos do que indivíduos, é preferível usar o PCA de acordo c/ os valores e vectores próprios de S.

Suponha agora que tem 400 fotografias ($n=40$)
cada uma 100×100.



obtemos assim 400 fotografias, sendo cada uma um vector com 10^4 atributos. A matriz de covariância será $S=\frac{1}{400}BB^T$ onde B é uma matriz $10^4\times 400$ e portanto S será do tipo $10^4\times 10^4$.

Por forma a aplicar o PCA, teremos que calcular os maiores valores próprios de S e respectivos vectores próprios!

Ao invés de se utilizar BB^T (a menos por um produto por um escalar – como vimos sabemos como altera o espectro) fazemos o estudo de B^TB. No exemplo, é uma matriz 400×400, com os mesmos valores próprios não-nulos de BB^T, ... e cujos vectores próprios se relacionam.

Repare que se v é vect. prop. de BB^T, $BB^Tv=\lambda v$, então $(B^TB)(B^Tv)=\lambda(B^Tv)$, i.e., B^Tv é vect. prop. de B^TB; reciprocamente, se u é vect.p. de B^TB, $B^TBu=\lambda u$, então $(BB^T)(Bu)=\lambda(Bu)$ e Bu é vect. prop. de BB^T.

Considere a matriz $X_{n\times m}$, com característica 2.4

$$\operatorname{rank}(X) = r = \operatorname{rank}(X^TX) = \operatorname{rank}(XX^T)$$

Sendo X^TX SDP então $\sigma(X^TX) \subseteq \mathbb{R}_0^+$.

Sejam $\lambda_1, \lambda_2, \dots, \lambda_r$ os valores próprios ñ-nulos de X^TX contando as multiplicidades.

Sejam $v_1, \dots, v_m$ vectores próprios associados aos valores próprios $\lambda_1, \lambda_2, \dots, \lambda_r, \lambda_{r+1}=0, \lambda_{r+2}=0, \dots, \lambda_m = 0$.

Como X^TX é simétrica, os m vectores próprios são ortogonais 2 a 2.

Sem perda de generalidade, sup. $v_1, \dots, v_r, \dots, v_m$ são ortogonais 2 a 2 e de norma 1. Ou seja

$v_i \perp v_j$ se $i \neq j$ e $\|v_i\| = 1$, satisfazendo

(i.e., $v_i^T v_j = 0$)

$$X^TX\, v_i = \lambda_i v_i \quad , \quad i = 1, \dots, r$$

$$X^TX\, v_i = 0 \quad , \quad i = r+1, \dots, m \quad \text{i.e., } v_i \in \operatorname{Ker}(X^TX) = \operatorname{Ker}(X)$$

Sejam $\sigma_i = \sqrt{\lambda_i}$, $u_i = \frac{1}{\sigma_i} X v_i$, $i = 1, \dots, r$.

(Recorde que X^TX é SDP, logo os valores próprios ñ-nulos são positivos)

Para $i = 1, \dots, r$, temos

$$\|u_i\| = \frac{1}{\sigma_i}\|Xv_i\| = \frac{1}{\sigma_i}\sqrt{(Xv_i)^T(Xv_i)} = \frac{1}{\sigma_i}\sqrt{v_i^T X^TX v_i}$$

$$= \frac{1}{\sigma_i}\sqrt{v_i^T \lambda_i v_i^T} = \frac{\sqrt{\lambda_i}}{\sigma_i}\sqrt{v_i^T v_i}$$

$X^T X v_i = \lambda_i v_i$

$$= \|v_i\| = 1$$

Ou seja, $\|u_i\| = 1$, $i = 1, \dots, r$

Também temos $u_i \perp u_j$, se $i \neq j$, $i,j = 1, \dots, r$

De facto, $\langle u_i, u_j \rangle = u_i^T u_j = \frac{1}{\sigma_i}(Xv_i)^T \frac{1}{\sigma_j}(Xv_j)$

$$= \frac{1}{\sigma_i \sigma_j} v_i^T \underbrace{X^T X v_j}_{\lambda_j v_j} = \frac{\lambda_j}{\sigma_i \sigma_j} v_i^T v_j = 0$$

pois $v_i^T v_j = \langle v_i, v_j \rangle = 0$ por $v_i \perp v_j$

Temos, portanto, $Xv_i = \sigma_i u_i$

$u_i \perp u_j$, $i \neq j$

$\|u_i\| = 1$, com $i,j = 1, \dots, r$

Ou seja
$$\begin{cases} Xv_1 = \sigma_1 u_1 \\ Xv_2 = \sigma_2 u_2 \\ \vdots \\ Xv_r = \sigma_r u_r \end{cases} \Leftrightarrow X \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_r \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_r \\ | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ & & \ddots \ \sigma_r \\ 0 & 0 \cdots & 0 \end{bmatrix}$$

Recorde que $Xv_{r+1} = Xv_{r+2} = \dots = Xv_m = 0$

e portanto $X \begin{bmatrix} | & | & & | & | & & | \\ v_1 & v_2 & \cdots & v_r & v_{r+1} & \cdots & v_m \\ | & | & & | & | & & | \end{bmatrix} = \begin{bmatrix} | & & | & \\ u_1 & \cdots & u_r & ?? \\ | & & | & \end{bmatrix} \underbrace{\begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & 0 \\ & & & 0 \end{bmatrix}}_{\Sigma}$

onde a matriz $\Sigma = \begin{bmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & 0 & \\ 0 & & & 0 \end{bmatrix}_{n \times m}$

Note que $u_1, \dots, u_r \in \mathbb{R}^n$, $r \leq n$

são 2 a 2 ortogonais e de norma 1.

Sabendo que $u_i = \frac{1}{\sigma_i} X v_i$, $i = 1, \dots, r$, onde v_i é vector prop. de $X^T X$ assoc. ao valor próprio $\lambda_i = \sigma_i^2$,

$$(XX^T) u_i = X X^T \frac{1}{\sigma_i} X v_i = \frac{1}{\sigma_i} X \underbrace{(X^T X v_i)}_{= \sigma_i^2 v_i}$$

$$= \sigma_i X v_i = \sigma_i^2 u_i = \lambda_i u_i$$

Ou seja, se v_i é vect. prop. $X^T X$ assoc. λ_i então u_i é vect. prop. XX^T assoc. $\lambda_i \neq 0$ (ao mesmo valor próprio)

onde $i = 1, \dots, r$.

Sendo XX^T simétrica, então é possível completar o conjunto $\{u_1, \dots, u_r\}$ com $n-r$ vectores próprios de XX^T assoc. $\lambda = 0$, $u_{r+1}, \dots, u_n$, por forma a que $u_1, \dots, u_r, u_{r+1}, \dots, u_n$ sejam ortogonais 2 a 2 e cada um com norma 1.

Obtemos assim

$$X \underbrace{\begin{bmatrix} | & & | \\ v_1 & \cdots & v_m \\ | & & | \end{bmatrix}}_{V_{m \times m}} = \underbrace{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{bmatrix}}_{U_{n \times n}} \underbrace{\begin{bmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \sigma_r & & \\ 0 & & & 0 & \\ & & & & \ddots \; 0 \end{bmatrix}}_{\Sigma_{n \times m}}$$

As colunas de $V_{m\times m}$ são 2 a 2 ortogonais de norma 1, donde $V^{-1} = V^T$

O mesmo se aplica às colunas de $U_{n\times n}$, donde

$$U^{-1} = V^T$$

$$XV = U\Sigma \Rightarrow X = U\Sigma V^{-1} = U\Sigma V^{-1}$$

A esta decomposição chama-se decomposição nos valores singulares. $\sigma_1, \dots, \sigma_r$ são os valores singulares,

$\sigma_i = \sqrt{\lambda_i}$ onde λ_i é valor próprio não nulo de XX^T

i.e., λ_i " " " " " " X^TX

Aplicação no PCA

Suponha que B é a matriz dos dados indivíduo/input da forma usual. Sup. que os dados estão centrados p/ média. (como as linhas indicam os vectores aleatórios, cada linha tem média 0). A matriz de covariância é $S = \frac{1}{n} BB^T$.

Seja $Y = \frac{1}{\sqrt{n}} B^T$; cada coluna terá média 0.

$$Y^TY = \left(\frac{1}{\sqrt{n}} B^T\right)^T \frac{1}{\sqrt{n}} B^T = \frac{1}{n} BB^T = S$$

Recorde que as componentes principais de B são os vectores próprios de S assoc. a valores próprios c/ certa "magnitude".

2.8

Aplicando SVD a Y, $Y = U\Sigma V^{-1}$,

as colunas de V são os vectores próprios de Y^TY

Se $X = B^T$ então $X^TX = nY^TY$, e X^TX e Y^TY

têm os mesmos vectores próprios.

Assim as colunas de V são os vect. prop. de X^TX

Ou seja, basta aplicar SVD a $X = B^T$.

Obtem-se assim $X = U\Sigma V^{-1}$ onde as colunas de

V são vectores próprios de X^TX.

$$B = X^T = (U\Sigma V^{-1})^T = (U\Sigma V^T)^T = V\Sigma U^T$$

Ou seja, as colunas de V são as componentes principais

de B. Aplicar SVD a B ou a $B^T = X$

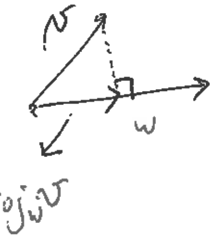
dependerá da dimensão da matriz.

Projecções!

Qual a importância de termos vectores ortogonais 2 a 2?

Sejam $w, v \in \mathbb{R}^n$. A projecção de v ao longo de w, $\text{proj}_w v$ é um vector de $\mathbb{R}^n$ definido por ~~$\text{proj}_w v$~~

$$\text{proj}_w v = \frac{\langle v, w \rangle}{\|w\|^2} w .$$



Dado um espaço vectorial (de dimensão finita) $V \subseteq \mathbb{R}^n$, seja $B = (u_1, \dots, u_k)$ uma base ortogonal de V.
Seja $w \in \mathbb{R}^n$. Define-se projecção ortogonal de w em V como o único $v \in V$ t.q. $v = \arg\min_{v' \in V} \|w - v'\|$

O facto de B ser uma base ortogonal permite calcular a projecção (ortogonal) de forma fácil:

$$\text{proj}_V w = \text{proj}_{u_1} w + \text{proj}_{u_2} w + \dots + \text{proj}_{u_k} w$$

Podemos, assim, calcular a projecção do novo "indivíduo" nas componentes principais, já que são vectores próprios ortogonais 2 a 2 e cada um de norma 1.

Sejam $v_1, \dots, v_K$ componentes principais, i.e., vectores próprios de BB^T assoc. aos maiores K valores próprios da matriz de covariância.

Seja ϕ um novo input, que assumimos centrado na média.

Para calcular a projecção de ϕ em $\langle v_1, \dots, v_K \rangle$, espaço gerado por $v_1, \dots, v_K$, como são ortogonais 2 a 2 e de norma 1, $\text{proj}_{\langle v_1, \dots, v_K \rangle} \phi = \sum_{i=1}^{k} \text{proj}_{v_i} \phi$

$$= \sum_{i=1}^{k} \langle \phi ; v_i \rangle v_i = \sum_{i=1}^{k} (\phi^T v_i) v_i$$

Seja c o vector dos coeficientes da projecção,

$$c = \begin{bmatrix} \phi^T v_1 \\ \phi^T v_2 \\ \vdots \\ \phi^T v_K \end{bmatrix} \qquad \text{Então} \quad \text{proj}_{\langle v_1, \dots, v_K \rangle} \phi = \begin{bmatrix} v_1 & \cdots & v_K \end{bmatrix} \begin{bmatrix} \phi^T v_1 \\ \vdots \\ \phi^T v_K \end{bmatrix}$$

$$= V_K c$$

Este processo permite "identificar" novos indivíduos. Para tal basta projectar o novo indivíduo ϕ (depois de centrado) e procurar qual o indivíduo conhecido cuja projecção minimiza a distância.

Ou seja:

Sejam $v_1, \dots, v_K$ as K componentes principais mais representativas (i.e., assoc. aos K maiores valores próprios)

Seja $V_K = [v_1 \cdots v_K]$ (pode ser obtida por SVD)

Seja ϕ um novo indivíduo (centrado na média!)

Sejam $x_1, \dots, x_n \in \mathbb{R}^m$ os elem.tos da amostra inicial, centrados

(i.e., $B = [x_1 \cdots x_n]$)

Se denotarmos $CS(V_K)$ o espaço gerado pelas colunas de V_K, então

$$\operatorname{proj}_{CS(V_K)} x_i = \langle x_i, v_1 \rangle v_1 + \langle x_i, v_2 \rangle v_2 + \cdots + \langle x_n, v_K \rangle v_K$$

Seja c_i o vector dos coeficientes da projecção

$$c_i = \begin{bmatrix} \langle x_i, v_1 \rangle \\ \langle x_i, v_2 \rangle \\ \vdots \\ \langle x_i, v_K \rangle \end{bmatrix} = \begin{bmatrix} x_i^T v_1 \\ x_i^T v_2 \\ \vdots \\ x_i^T v_K \end{bmatrix}_{K \times 1} \in \mathbb{R}^K, \text{ c/ } K \ll m$$

Façamos o mesmo para ϕ:

$$\operatorname{proj}_{CS(V_K)} \phi = \sum_{i=1}^{K} \langle \phi, v_i \rangle v_i$$ e seja

c o vetor dos coeficientes, $c = \begin{bmatrix} \phi^T v_1 \\ \vdots \\ \phi^T v_K \end{bmatrix}$

Identificamos ϕ com x_j se

$$c_j = \arg\min d(c_i, c) \quad \text{e} \quad d(c_j, c) < L$$

onde L é uma certa tolerância

Podemos usar a distância euclidiana

$$d(v,w) = \| v - w \|_2 = \sqrt{\langle v-w, v-w \rangle}$$

No entanto, há relatos em como a distância de Mahalanobis tem um melhor comportamento:

$$d_M\left(\begin{bmatrix} e_1 \\ \vdots \\ e_K \end{bmatrix}, \begin{bmatrix} e_1^i \\ e_2^i \\ \vdots \\ e_K^i \end{bmatrix} \right) = \sum_{j=1}^{K} \frac{1}{\lambda_j} (e_j - e_j^i)^2$$

Intuitivamente, dá-se mais importância às entradas correspondentes aos maiores valores próprios, quando pretendemos minimizar a distância.

- Dada $X_{n\times m}$, prova em como

$$\ker(X^TX) = \ker(X) \quad \text{onde } \ker(A) \text{ denota}$$

o núcleo de A : $\ker(A) = \{v : Av = 0\}$

Se $v \in \ker(X)$ então $Xv = 0 \Rightarrow X^TXv = 0$

$$\Rightarrow v \in \ker(X^TX)$$

Reciprocamente se $v \in \ker(X^TX)$ então

$X^TXv = 0 \Rightarrow v^TX^TXv = 0$ (multiplicando à esquerda por v^T)

$$\Rightarrow (Xv)^TXv = 0$$

$$\Rightarrow \langle Xv ; Xv \rangle = 0 \Rightarrow \|Xv\| = 0$$

$$\Rightarrow Xv = 0 \Rightarrow v \in \ker(X)$$

- O ponto anterior permite mostrar que $\operatorname{rank}(X) = \operatorname{rank}(X^TX)$.

Recorde que se X é $n\times m$ então

$$m = \operatorname{rank}(X) + \dim \ker(X)$$

Ora X^TX é $m\times m$, o que leva a

$$m = \operatorname{rank}(X^TX) + \dim \underbrace{\ker(X^TX)}_{=\ker(X)} = \operatorname{rank}(X^TX) + \dim\ker(X)$$

Logo $\operatorname{rank}(X) = \operatorname{rank}(X^TX)$.