Análisis de componentes principales (PCA)

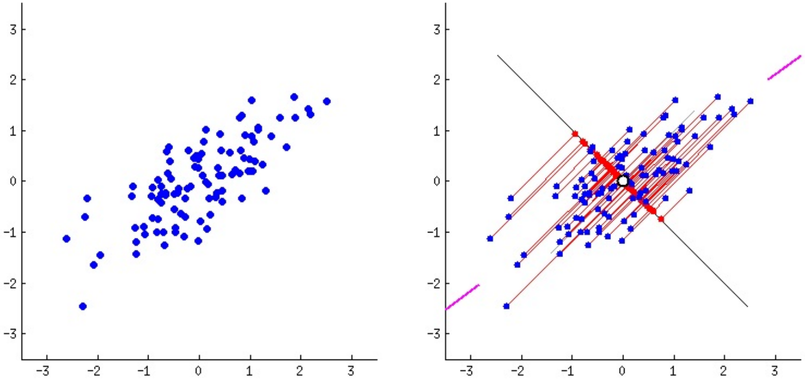
Alan Reyes-Figueroa

Introducción a la Ciencia de Datos (Aula 09) 04.febrero.2021

Componentes principales

Objetivo: encontrar una estructura subyacente en los datos.

- Proyectar a un subespacio adecuado.

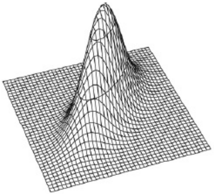


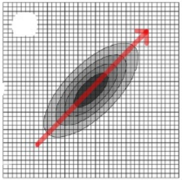
Caso particular 1D: (proyectamos a un subespacio 1-dimensional).

Suponga que proyectamos a un subespacio $\langle \ell \rangle \;\Rightarrow\; \langle \ell, X \rangle = \ell^T X$.
Buscamos maximizar

$$\max_{||\ell||=1} Var(\ell^T X) = \max_{\ell \neq 0} \frac{Var(\ell^T X)}{\ell^T \ell} = \max_{\ell \neq 0} \frac{\ell^T Var(X) \ell}{\ell^T \ell} = \max_{\ell \neq 0} \frac{\ell^T (\mathbb{X}^T \mathbb{X}) \ell}{\ell^T \ell}.$$

(cociente de Rayleigh).





El teorema espectral

Teorema (Teorema espectral / Descomposición espectral)

Sea $A \in \mathbb{R}^{d\times d}$ una matriz simétrica (operador auto-adjunto). Entonces, A admite una descomposición de la forma

$$A = U\Lambda U^T,$$

donde $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)$ es la matriz diagonal formada por los autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d$ de A, y

$$U = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_d \end{pmatrix} \in \mathbb{R}^{d\times d}$$

es una matriz ortogonal cuyas columnas son los autovalores de A, con $\mathbf{u}_i$ el autovalor correspondiente a λ_i, $i = 1, 2, \ldots, d$.

Teorema (Teorema espectral / Descomposición espectral)

En otras palabras, A puede escribirse como una suma de matrices de rango 1

$$A = \sum_{i=1}^{d} \lambda_i \, \mathbf{u}_i \mathbf{u}_i^T.$$

<u>Comentario</u>:

Para $1 \leq k \leq d$, la suma $A = \sum_{i=1}^{k} \lambda_i \, \mathbf{u}_i \mathbf{u}_i^T$, es una matriz de rango k siempre que los $\lambda_i \neq 0$ (ya que los $\mathbf{u}_i$ son independientes).

El teorema espectral

Observaciones:

- Si A es simétrica y semi-definida positiva, existe $A^{1/2}$ tal que $A^{1/2}A^{1/2} = A$.

- Si todos los autovalores de A son no-negativos, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$, entonces $\Lambda^{1/2}$ existe y

$$\Lambda^{1/2} = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)^{1/2} = diag(\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_d^{1/2}).$$

- A partir de la descomposición espectral podemos calcular $A^{1/2}$. De hecho, si $A = U\Lambda U^T$, definimos $A^{1/2} = U\Lambda^{1/2}U^T$, y

$$\begin{aligned} A^{1/2}A^{1/2} &= (U\Lambda^{1/2}U^T)(U\Lambda^{1/2}U^T) = U\Lambda^{1/2}(U^TU)\Lambda^{1/2}U^T \\ &= U\Lambda^{1/2}\Lambda^{1/2}U^T = U\Lambda U^T = A. \end{aligned}$$

Descomposición SVD

Teorema (Descomposición en valores singulares (SVD))

Sea $A \in \mathbb{R}^{n\times d}$ una matriz de rango k. Para todo $1 \leq r \leq k$, existen matrices $U \in \mathbb{R}^{n\times r}$, $S \in \mathbb{R}^{r\times r}$, $V \in \mathbb{R}^{d\times r}$, tales que

$$A = USV^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

con

- *las columnas $\mathbf{u}_1, \ldots, \mathbf{u}_r \in \mathbb{R}^n$ de U son los autovectores de AA^T,*
- *las columnas $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{R}^d$ de V son los autovectores de A^TA, $S = diag(\sigma_1, \ldots, \sigma_r)$, $\sigma_i^2 = \lambda_i$, con λ_i los autovectores de $\mathbf{u}_i$ y de $\mathbf{v}_i$,*
- *Además, $\sigma_i \mathbf{u}_i = A\mathbf{v}_i$ y $\sigma_i \mathbf{v}_i = A^T\mathbf{u}_i$, para $i = 1, 2, \ldots, r$.*

Descomposición SVD

El teorema de descomposición espectral ocurre como un caso particular de la descomposición SVD:

Caso especial: A simétrica

$$A = USU^T = U\Lambda U^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{u}_i^T.$$

En este caso los autovectores de A y $A^TA = A^2 = AA^T$ coinciden, y los autovalores de A al cuadrado son los autovalores de A^TA.

Teorema (Cociente de Rayleigh, caso 1D)

Sea $A \in \mathbb{R}^{d\times d}$ una matriz simétrica, $A \succeq 0$. Entonces, el cociente de Rayleigh

$$\max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

alcanza su máximo exactamente en $\mathbf{x} = \mathbf{u}_1$, el autovector asociado al mayor autovalor λ_1 de A.

Prueba:
Sea $A = U\Lambda U^T$ la descomposición espectral del A, A con autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d \geq 0$, y $U = [\mathbf{u}_1\ \mathbf{u}_2\ \ \ldots\ \ \mathbf{u}_d]$, con $\mathbf{u}_i$ el autovalor correspondiente a λ_i, $i = 1, 2, \ldots, d$.

Cociente de Rayleigh

Tomemos $A^{1/2} = U\Lambda^{1/2}U^T$.
Consideremos el cambio de base $\mathbf{y} = U^T\mathbf{x}$. Entonces

$$\begin{aligned}
\max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} &= \max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T A^{1/2}A^{1/2}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T U\Lambda^{1/2}U^T U\lambda^{1/2}U^T\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T U\Lambda U^T\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \\
&= \max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T U\Lambda U^T\mathbf{x}}{\mathbf{x}^T UU^T\mathbf{x}} = \max_{\mathbf{x}\neq 0} \frac{(U^T\mathbf{x})^T\Lambda(U^T\mathbf{x})}{(U^T\mathbf{x})^T(U^T\mathbf{x})} = \max_{\mathbf{y}\neq 0} \frac{\mathbf{y}^T\Lambda\mathbf{y}}{\mathbf{y}^T\mathbf{y}} = \max_{||\mathbf{y}||=1} \mathbf{y}^T\Lambda\mathbf{y} \\
&= \max_{||\mathbf{y}||=1} \sum_{i=1}^{d} \lambda_i y_i^2 \leq \max_{||\mathbf{y}||=1} \sum_{i=1}^{d} \lambda_1 y_i^2 = \lambda_1.
\end{aligned}$$

Luego, el valor del cociente de Rayleigh, está limitado superiormente por λ_1.

Por otro lado, si $\mathbf{y} = \mathbf{e}_1 = (1, 0, \dots, 0)$, entonces

$$\frac{\mathbf{y}^T \Lambda \mathbf{y}}{\mathbf{y}^T \mathbf{y}} = \mathbf{y}^T \Lambda \mathbf{y} = \mathbf{e}^T \Lambda \mathbf{e}_1 = \sum_{i=1}^{d} \lambda_i \mathbf{e}_{1i}^2 = \lambda_1.$$

Portanto, el cociente de Rayleigh alcanza su máximo en $\mathbf{y} = \mathbf{e}_1$. Volviendo a las coordenadas originales, como $\mathbf{y} = U^T \mathbf{x}$, entonces

$$\mathbf{x} = (U^T)^{-1} \mathbf{e}_1 = U \mathbf{e}_1 = \mathbf{u}_1.$$

De modo que el cociente de Rayleigh alcanza su máximo en $\mathbf{x} = \mathbf{e}_1$, el autovector asociado al mayor autovalor de A. $\square$

Proyección PCA

Caso general: Proyectar a un subespacio r-dimensional.
Buscamos direcciones ortogonales $\{\ell_i\}_{i=1}^r$ que generan el supespacio de proyección.

$$\max_{||\ell_i||=1} Var(\ell_i^T X) = \max_{\ell_i \neq 0} \frac{\ell_i^T Cov(X) \ell_i}{\ell_i^T \ell_i}, \quad \text{sujeto a } \ell_i \perp \ell_1, \ldots, \ell_{i-1}, \; i = 2, 3, \ldots, r.$$

Solución: $\{\ell_i\}$ son los autovectores asociados a los primeros r autovectores de $Cov(X)$.

Prueba: El caso $i = 1$ está resuelto, la proyección se maximiza con el autovector $\mathbf{u}_1$, la primer columna de U en la descomposición SVD de $Cov(X)$.

Proyección PCA

Sea $A = Cov(X)$. Ilustramos ahora como proyectar en la segunda dirección. Para ello, consideramos el espacio ortogonal a $\langle \mathbf{u}_1 \rangle$, esto es, borramos la información de la matriz A en la dirección de $\mathbf{u}_1$:

$$A_2 = A - \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{u}_i^T - \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T = \sum_{i=2}^{d} \lambda_i \mathbf{u}_i \mathbf{u}_i^T.$$

Observe que $A_2 \in \mathbb{R}^{d \times d}$ es una matriz d-dimensional, pero con ceros en toda su primera fila y columna (en la base $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d$.
Luego, podemos considerarla como una matriz $d - 1$-dimensional. La información en el resto de dimensiones no ha cambiado, esto es, los autovalores y autovectores de A_2 son, respectivamente $\lambda_2 > \ldots > \lambda_d$, y $\mathbf{u}_2, \ldots, \mathbf{u}_d$.

Proyección PCA

De ahí, resolver el problema

$$\max_{\ell_2 \neq 0} \frac{\ell_2^T A \ell_2}{\ell_2^T \ell_2}, \quad \text{sujeto a } \ell_2 \perp \mathbf{u}_1,$$

se reduce a

$$\max_{\ell_2 \neq 0} \frac{\ell_2^T A_2 \ell_2}{\ell_2^T \ell_2}.$$

Ya vimos que la solución de este cociente de Rayleigh es dada por $\mathbf{u}_2$, el autovector asociado al mayor autovalor λ_2 de A_2.

Este mismo proceso se generaliza al resto de dimensiones $\ell_3, \ldots, \ell_r$. Esto termina la prueba de la descomposición PCA. $\square$

UNIVERSIDAD DEL VALLE DE GUATEMALA

Teorema (Eckart-Young)

Sea $A \in \mathbb{R}^{n\times d}$, $n \geq d$, una matriz cuya descomposición SVD está dada por

$$A = USV^T = \sum_{i=1}^{d} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

Entonces, la matriz $\widehat{A}_r$ de rango r, $1 \leq r \leq d$, que mejor aproxima A en el sentido de minimizar

$$\min_{rank\ \widehat{A}_r \leq r} ||A - \widehat{A}_r||_F^2$$

se obtiene de truncar la descomposición en valores dingulares de A:

$$\widehat{A}_r = U_r S_r V_r^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

Teorema (Eckart-Young)

donde

$$U_r = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_r],\ \ S_r = diag(\sigma_1, \sigma_2 \ldots, \sigma_r),\ \ V_r = [\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_r].$$

En ese caso, el error de aproximación está dado por

$$||A - \widehat{A}_r||_F^2 = \sum_{i=r+1}^{d} \lambda_i,$$

o

$$||A - \widehat{A}_r||_2^2 = \lambda_{r+1}.$$

Aproximaciones de bajo rango

Obs!

- Las direcciones $\mathbf{u}_i$ se llaman las **componentes principales** de $\mathbb{X}$.
- La descomposición SVD proporciona un mecanismo para proyectar los datos al "mejor" subespacio de dimensión $r \leq d$. Dicha proyección se obtiene haciendo
$$\mathbb{X}_{proj} = \mathbb{X}\, V_r^T.$$
- Los autovalores λ_i de $\mathbb{X}^T\mathbb{X}$ nos proporcionan un mecanismo para medir el error, vía $||A - \widehat{A}_r||_F^2 = \sum_{i=r+1}^d \lambda_i$.
- El cociente $\dfrac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^d \lambda_i}, \quad r = 1, 2, \ldots, d$, se interpreta como el porcentaje de variabilidad de los datos $\mathbb{X}$ que es explicada por las primeras r componentes principales.

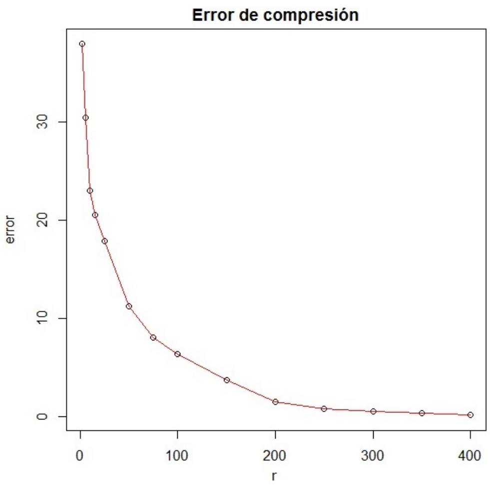
Ejemplos

Compresión de imágenes usando PCA.

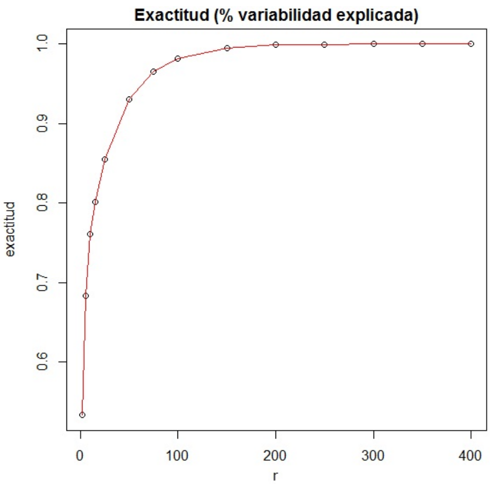
Original $\quad r = 1 \quad r = 2 \quad r = 4 \quad r = 8 \quad r = 16 \quad r = 32 \quad r = 64$

Imagen Original (256×256), aproximaciones con rango = 1, 2, 4, 8, 16, 32, 64.

Ejemplos



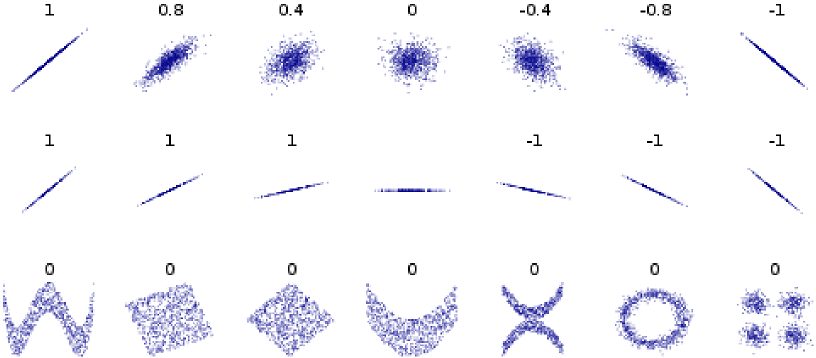
Error de compresión
error
0
10
20
30
0
100
200
300
400
r



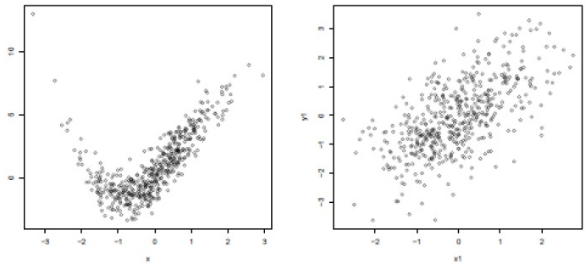
Exactitud (% variabilidad explicada)
exactitud
0.6
0.7
0.8
0.9
1.0
0
100
200
300
400
r

Ejemplos

En PCA la estructura de los datos se capta solamente a través de las matrices $Cov(X)$ o $Corr(X)$.



Ejemplos

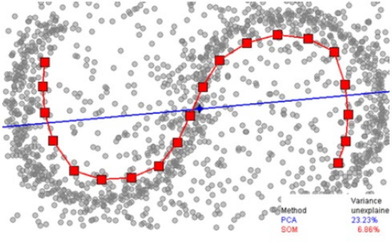


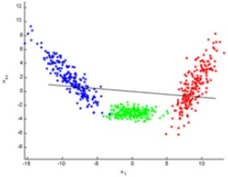
Dos veces misma correlación. =(

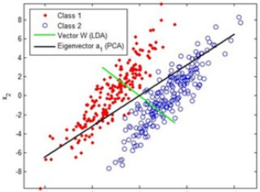
Obs.

- Cuidado con desviaciones fuertes de normalidad.
- Lo ideal es investigar la normalidad de los datos en la práctica, al menos ver si escala es continua, distribución unimodal, simétrica, ...

Contraejemplos







UNIVERSIDAD DEL VALLE DE GUATEMALA