

ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

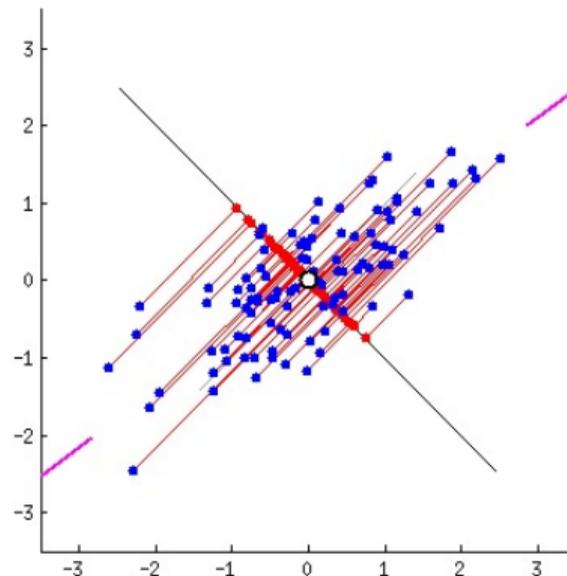
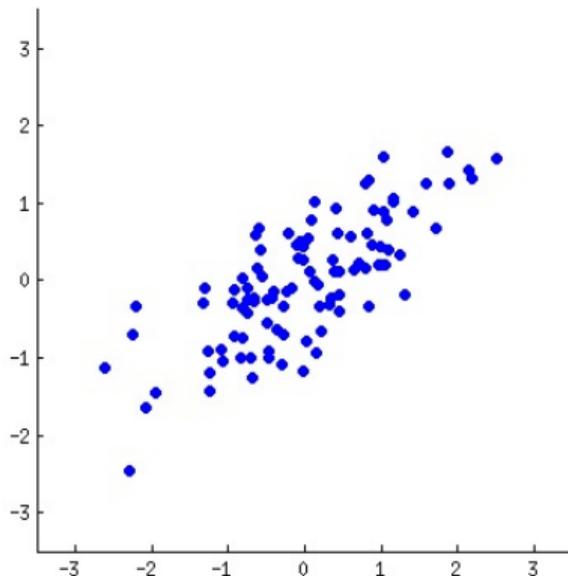
ALAN REYES-FIGUEROA
MÉTODOS NUMÉRICOS II

(AULA 05B) 18.JULIO.2023

Componentes principales

Objetivo: encontrar una estructura subyacente en los datos.

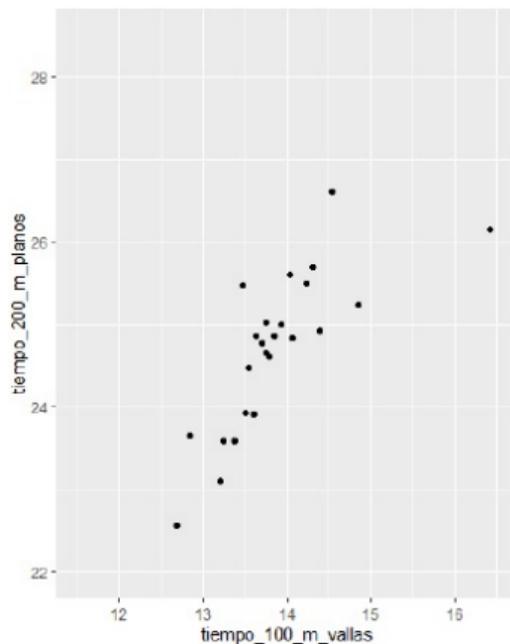
- Proyectar a un subespacio adecuado.



Componentes principales

Ejemplo: Atletismo, pruebas de 100m y 200m.

100m vallas	200m planos
12.69	22.56
12.85	23.65
13.2	23.1
13.61	23.92
13.51	23.93
13.75	24.65
13.38	23.59
13.55	24.48
13.63	24.86
13.25	23.59
13.75	25.03
13.24	23.59
13.85	24.87
13.71	24.78
13.79	24.61
13.93	25
13.47	25.47
14.07	24.83
14.39	24.92
14.04	25.61
14.31	25.69
14.23	25.5
14.85	25.23
14.53	26.61
16.42	26.16

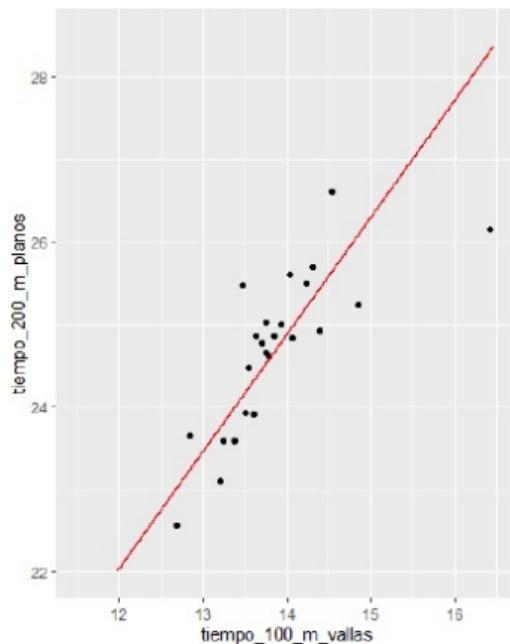


Se observa cierta estructura.

Componentes principales

Ejemplo: Atletismo, pruebas de 100m y 200m.

100m vallas	200m planos
12.69	22.56
12.85	23.65
13.2	23.1
13.61	23.92
13.51	23.93
13.75	24.65
13.38	23.59
13.55	24.48
13.63	24.86
13.25	23.59
13.75	25.03
13.24	23.59
13.85	24.87
13.71	24.78
13.79	24.61
13.93	25
13.47	25.47
14.07	24.83
14.39	24.92
14.04	25.61
14.31	25.69
14.23	25.5
14.85	25.23
14.53	26.61
16.42	26.16



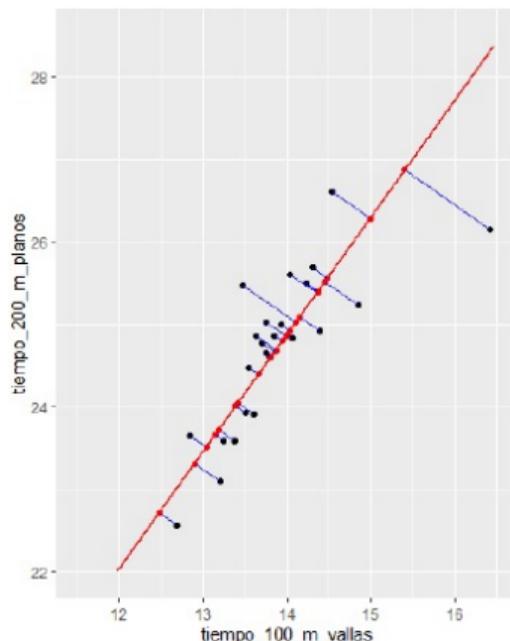
Se observa cierta estructura.

Karl Pearson (1901), describir con una recta.

Componentes principales

Ejemplo: Atletismo, pruebas de 100m y 200m.

100m vallas	200m planos
12.69	22.56
12.85	23.65
13.2	23.1
13.61	23.92
13.51	23.93
13.75	24.65
13.38	23.59
13.55	24.48
13.63	24.86
13.25	23.59
13.75	25.03
13.24	23.59
13.85	24.87
13.71	24.78
13.79	24.61
13.93	25
13.47	25.47
14.07	24.83
14.39	24.92
14.04	25.61
14.31	25.69
14.23	25.5
14.85	25.23
14.53	26.61
16.42	26.16



Se observa cierta estructura.

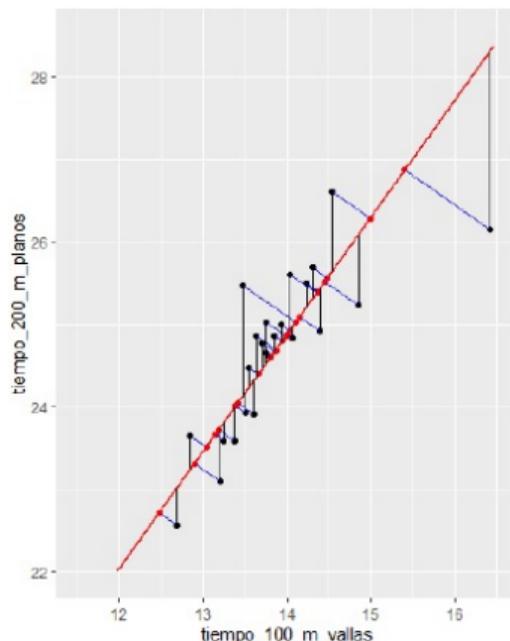
Karl Pearson (1901), describir con una recta.

Hotelling (1933), relación entre variables $g(X_1, X_2)$.

Componentes principales

Ejemplo: Atletismo, pruebas de 100m y 200m.

100m vallas	200m planos
12.69	22.56
12.85	23.65
13.2	23.1
13.61	23.92
13.51	23.93
13.75	24.65
13.38	23.59
13.55	24.48
13.63	24.86
13.25	23.59
13.75	25.03
13.24	23.59
13.85	24.87
13.71	24.78
13.79	24.61
13.93	25
13.47	25.47
14.07	24.83
14.39	24.92
14.04	25.61
14.31	25.69
14.23	25.5
14.85	25.23
14.53	26.61
16.42	26.16



Se observa cierta estructura.

Karl Pearson (1901), describir con una recta.

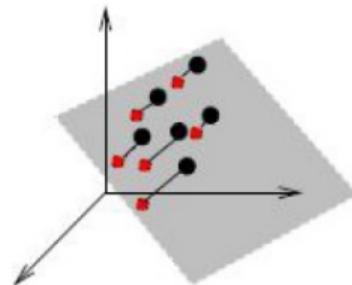
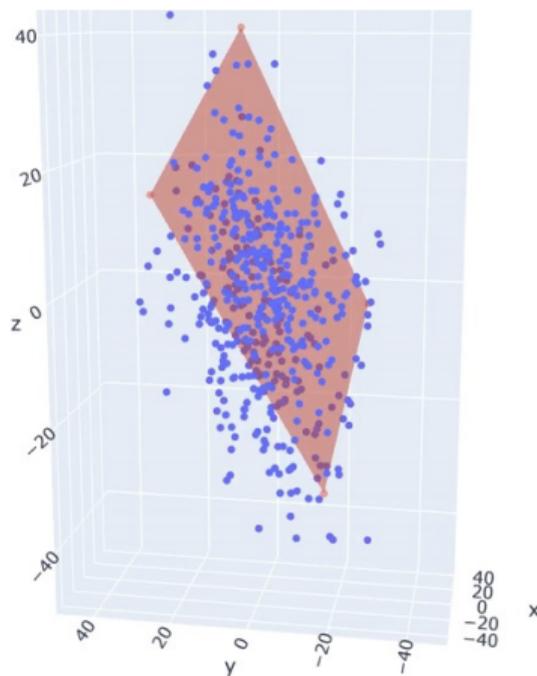
Hotelling (1933), relación entre variables $g(X_1, X_2)$.

No confundir con regresión, Incorporar incertidumbre.

Componentes principales

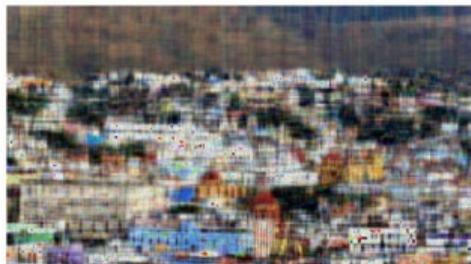
Ejemplo: Atletismo, pruebas de 100m, 200m y salto de longitud.

100m vallas	200m planos	salto long
12.69	22.56	7.27
12.85	23.65	6.71
13.2	23.1	6.68
13.61	23.92	6.25
13.51	23.93	6.32
13.75	24.65	6.33
13.38	23.59	6.37
13.55	24.48	6.47
13.63	24.86	6.11
13.25	23.59	6.28
13.75	25.03	6.34
13.24	23.59	6.37
13.85	24.87	6.05
13.71	24.78	6.12
13.79	24.61	6.08
13.93	25	6.4
13.47	25.47	6.34
14.07	24.83	6.13
14.39	24.92	6.1
14.04	25.61	5.99
14.31	25.69	5.75
14.23	25.5	5.5
14.85	25.23	5.47
14.53	26.61	5.5
16.42	26.16	4.88

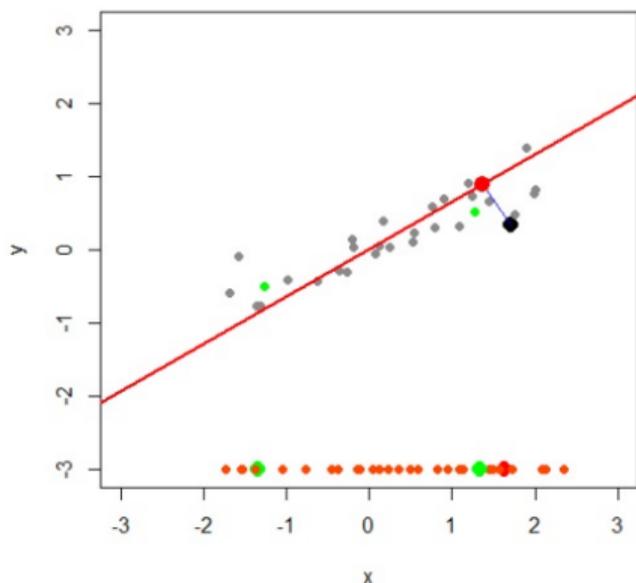


Componentes principales

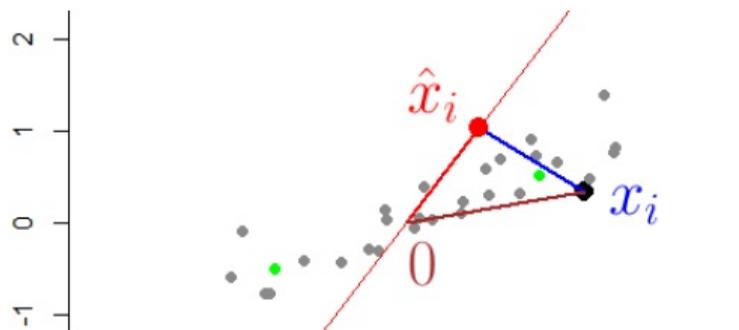
Ejemplo: Compresión de imágenes digitales.



Componentes principales



- Buscamos direcciones informativas (estructura)
informativo = máxima variabilidad
- Buscamos minimizar el error de reconstrucción.



Componentes principales

Obs! Los dos enfoques anteriores son equivalentes.

Prueba:

Denotemos X la v.a. que corresponde a los datos ($X \in \mathbb{R}^2$ en el ejemplo).

Por simplicidad, supongamos que los datos \mathbf{x}_i están centrados (i.e. $\mathbb{E}(X) = \mathbf{0}$).

$$\begin{aligned} MSS &= \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i\|^2 = \frac{1}{n} \sum_{i=1}^n \|(\mathbf{x}_i - \hat{\mathbf{x}}_i) + \hat{\mathbf{x}}_i\|^2 \\ &= \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 + \frac{1}{n} \sum_{i=1}^n \|\hat{\mathbf{x}}_i\|^2 = \text{Reconstruction error} + \mathbb{V}(X). \end{aligned}$$

MSS es fija, luego minimizar el error de reconstrucción equivale a maximizar la varianza de los datos.

Componentes principales

Enfoque probabilístico:

Matriz de datos

$$\mathbb{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1d} \\ X_{21} & X_{22} & \dots & X_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nd} \end{pmatrix}.$$

- Consideramos $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ como variable aleatoria, y los datos $\mathbf{x}_i = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d$, para $i = 1, 2, \dots, n$ como muestra de X .
- Supondremos que conocemos la ley \mathbb{P}_X .
- Supondremos también que $\mathbb{E}(X) = \mathbf{0}$ (los datos están centrados). En consecuencia, $\mathbb{V}(X) = \mathbb{X}^T \mathbb{X}$.

Componentes principales

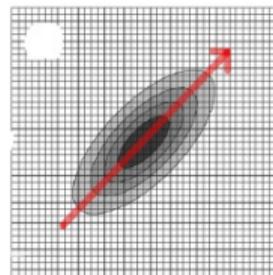
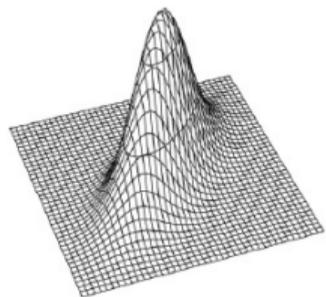
Caso particular 1D: (proyectamos a un subespacio 1-dimensional).

Suponga que proyectamos a un subespacio $\langle l \rangle \Rightarrow \langle l, \mathbb{X} \rangle = l^T \mathbb{X}$.

Buscamos maximizar

$$\max_{\|l\|=1} \text{Cov}(l^T \mathbb{X}) = \max_{l \neq 0} \frac{\text{Cov}(l^T \mathbb{X})}{l^T l} = \max_{l \neq 0} \frac{l^T \text{Cov}(\mathbb{X}) l}{l^T l} = \max_{l \neq 0} \frac{l^T (\mathbb{X}^T \mathbb{X}) l}{l^T l}.$$

(cociente de Rayleigh).



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 $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$ es la matriz diagonal formada por los autovalores $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ de A , y

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

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

Teorema (Teorema espectral / Descomposición espectral)

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

$$\begin{aligned} A &= \begin{pmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \dots & \mathbf{q}_d \end{pmatrix} \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{pmatrix} \begin{pmatrix} \mathbf{q}_1^T \\ \mathbf{q}_2^T \\ \dots \\ \mathbf{q}_d^T \end{pmatrix} \\ &= \sum_{i=1}^d \lambda_i \mathbf{q}_i \mathbf{q}_i^T. \end{aligned}$$

Componentes principales

Comentario:

Para $1 \leq k \leq d$, la suma

$$\hat{A}_k = \sum_{i=1}^k \lambda_i \mathbf{q}_i \mathbf{q}_i^T,$$

es una matriz de rango k siempre que los $\lambda_i \neq 0$ (ya que los \mathbf{q}_i son independientes).

Veremos más adelante, que esta es la mejor aproximación de rango k de la matriz A .

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 $\mathbf{u}_1, \dots, \mathbf{u}_r \in \mathbb{R}^n$ de U son los autovectores de AA^T ,
- las columnas $\mathbf{v}_1, \dots, \mathbf{v}_r \in \mathbb{R}^d$ de V son los autovectores de $A^T A$, $S = \text{diag}(\sigma_1, \dots, \sigma_r)$, $\sigma_i^2 = \lambda_i$, 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 $i = 1, 2, \dots, r$.

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 $A^T A = A^2 = A A^T$ coinciden, y los autovalores de A al cuadrado son los autovalores de $A^T A$.

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

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

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 de A , A con autovalores $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$, y $U = [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_d]$, con \mathbf{u}_i el autovector correspondiente a λ_i , $i = 1, 2, \dots, d$.

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

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

Cociente de Rayleigh

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

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

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 subespacio de proyección.

$$\max_{\|\ell_i\|=1} \mathbb{V}(\ell_i^T X) = \max_{\ell_i \neq 0} \frac{\ell_i^T \text{Cov}(X) \ell_i}{\ell_i^T \ell_i}, \quad \text{sujeito a } \ell_i \perp \ell_1, \dots, \ell_{i-1}, \quad i = 2, 3, \dots, r.$$

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

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

Proyección PCA

Sea $A = \text{Cov}(X)$. 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 $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d$).

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 $\lambda_2 > \dots > \lambda_d$, y $\mathbf{u}_2, \dots, \mathbf{u}_d$.

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 ℓ_3, \dots, ℓ_r . Esto termina la prueba de la descomposición PCA. \square

Aproximaciones de bajo rango

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

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

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

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

Teorema (Eckart-Young)

donde

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

En ese caso, el error de aproximación está dado por

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

o

$$\|A - \hat{A}_r\|_2 = \lambda_{r+1}.$$

Aproximaciones de bajo rango

Obs!

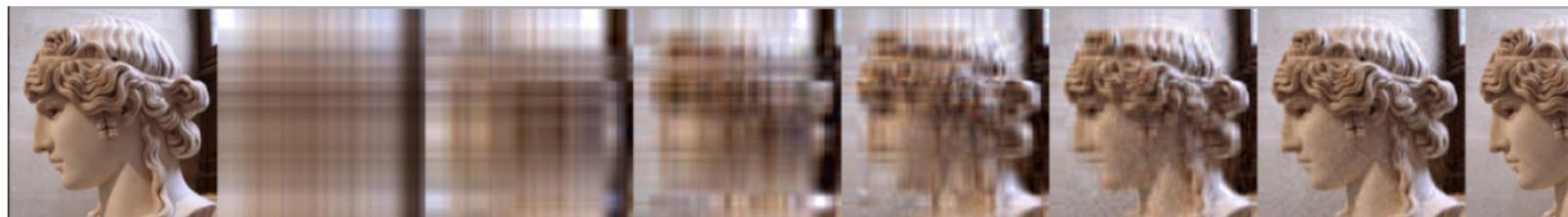
- Las direcciones \mathbf{u}_i se llaman las **componentes principales** de \mathbb{X} .
- La descomposición SVD proporciona un mecanismo para proyectar los datos al “mejor” subespacio de dimensión $r \leq d$. Dicha proyección se obtiene haciendo

$$\mathbb{X}_{proj} = \mathbb{X} V_r.$$

- Los autovalores λ_i de $\mathbb{X}^T \mathbb{X}$ nos proporcionan un mecanismo para medir el error, vía $\|\mathbf{A} - \hat{\mathbf{A}}_r\|_F^2 = \sum_{i=r+1}^d \lambda_i$.
- El cociente $\frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^d \lambda_i}$, $r = 1, 2, \dots, d$, se interpreta como el porcentaje de variabilidad de los datos \mathbb{X} que es explicada por las primeras r componentes principales.
Este se conoce usualmente como **explained variance**.

Ejemplos

Compresión de imágenes usando PCA.

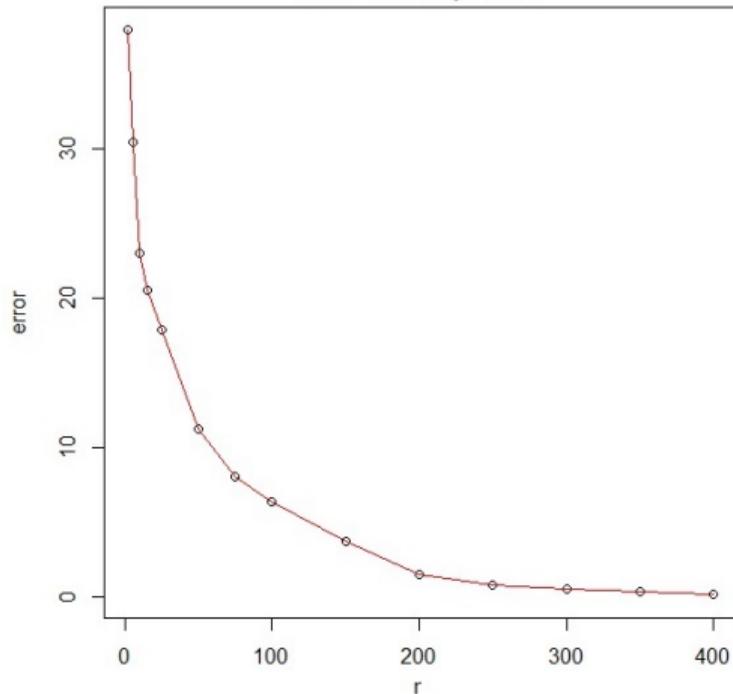


Original $r = 1$ $r = 2$ $r = 4$ $r = 8$ $r = 16$ $r = 32$ $r = 64$

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

Ejemplos

Error de compresión



Exactitud (% variabilidad explicada)

