UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Análisis de componentes principales (PCA)

Alan Reyes-Figueroa

Aprendizaje Estadístico (Aula 05) 22.enero.2024

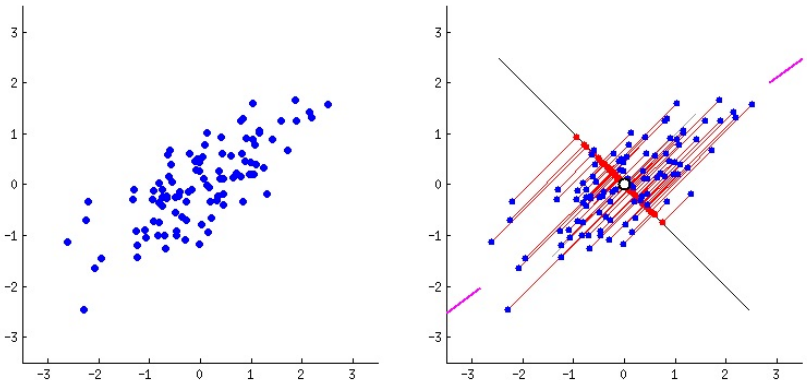
Introducción

1. **Métodos exploratorios y de visualización** (30%)
   Métodos exploratorios para datos multivariados: Visualización y resumen de la dependencia entre variables. Métodos de proyección: Descomposición SVD. Componentes principales (PCA). Re-escalamiento multidimensional. Componentes independientes (ICA). Reducción de la dimensionalidad: Factoración de matrices no-negativas (NNMF). Variables latentes. Otros tópicos.
2. **Aprendizaje no-supervisado** (20%)
   Métodos de agrupamiento. $k-$medias, $k-$medianas, $k-$medoides. Métodos de agrupamiento jerárquico: dendrogramas. Agrupamiento espectral. El método EM.
3. **Aprendizaje supervisado** (50%)
   El clasificador bayesiano. Análisis discriminante. $k-$*nearest neighbors*. Regresión logística. Máquinas de soporte vectorial (SVM). Métodos *kernel*. Árboles de Decisión y random forests. *Bagging* y *Boosting*. Redes neuronales artificiales. Validación cruzada y selección de modelos.
   Mínimos cuadrados. Modelos de regresión lineal (generalizada). Selección de variables. Métodos de regularización: Ridge (L_2), LASSO (L_1), *Elastic-net* (L_0). Criterios de selección de modelos: Mínimos cuadrados parciales. Métodos basados en mezclas. Funciones de base radial (RBFs), mezclas Gaussianas (GMM). Estimaciones empíricas de distribuciones.

Componentes principales

Objetivo: encontrar una estructura subyacente en los datos.

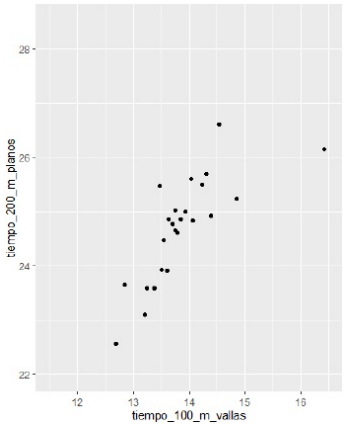
- Proyectar a un subespacio adecuado.



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

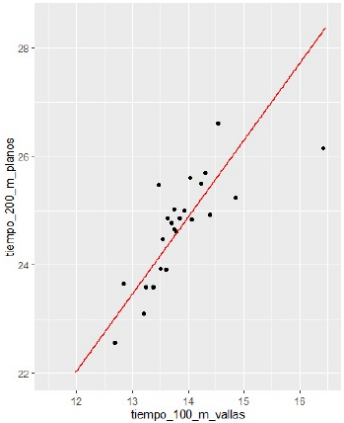


Se observa cierta estructura.

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



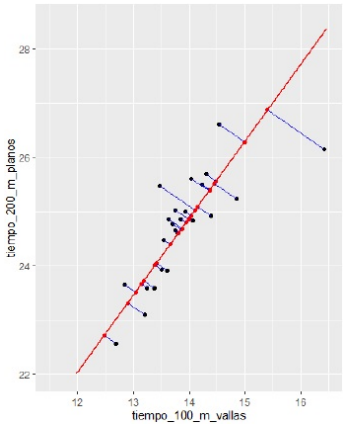
Se observa cierta estructura.

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.

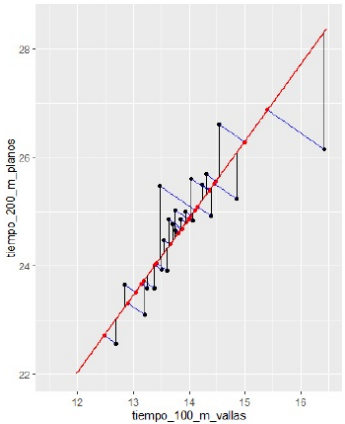
Karl Pearson (1901), describir con una recta.

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.

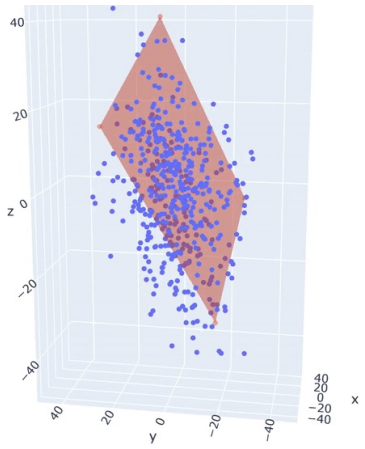
Hotelling (1933), relación entre variables $g(X_1, X_2)$.

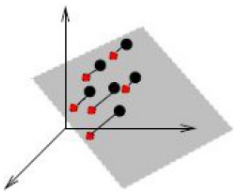
No confundir con regresión, Incorporar incertidumbre.

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

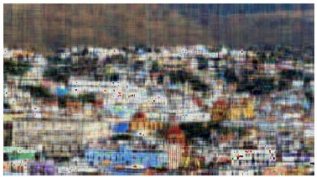




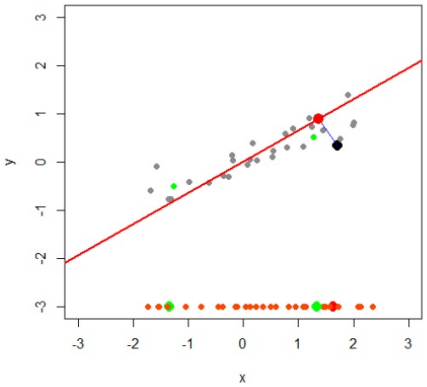
UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Componentes principales

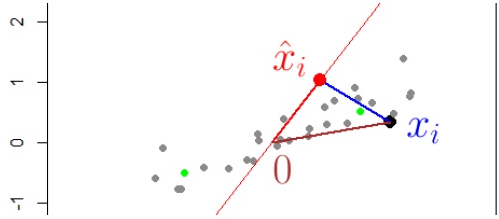
Ejemplo: Compresión de imágenes digitales.



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. $\mathbb{E}(X) = \mathbf{o}$).

$$\begin{aligned} Var(X) &= \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i||^2 = \frac{1}{n}\sum_{i=1}^{n} ||(\mathbf{x}_i - \widehat{\mathbf{x}}_i) + \widehat{\mathbf{x}}_i||^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \widehat{\mathbf{x}}_i||^2 + \frac{1}{n}\sum_{i=1}^{n} ||\widehat{\mathbf{x}}_i||^2 = \textit{Reconstruction error} + Var(X). \end{aligned}$$

$Var(X)$ 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, $Var(X) = \mathbb{X}^T\mathbb{X}$.

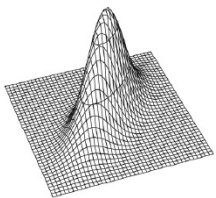
Componentes principales

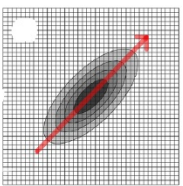
Caso particular 1D: (proyectamos a un subespacio 1-dimensional).

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).





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 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ de A, y

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

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

Teorema (Teorema espectral / Descomposición espectral)

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

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

Teorema Espectral

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

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

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 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

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

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 autovector correspondiente a λ_i, $i = 1, 2, \ldots, d$.

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

Cociente de Rayleigh

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 $\mathbf{y} = \mathbf{e}_1 = (1, 0, \ldots, 0)$, entonces

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

Cociente de Rayleigh

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, o $< r < d$.
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 $Cov(X)$.

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 $\mathbf{u}_1$:

Proyección PCA

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

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 $\mathbf{u}_2, \ldots, \mathbf{u}_d$.

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

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

Proyección PCA

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. $\square$

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.

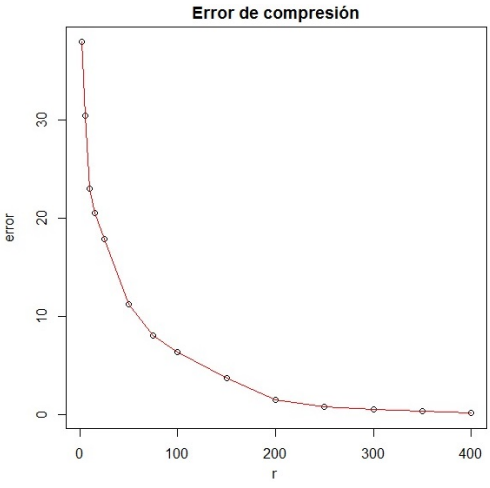
Ejemplos

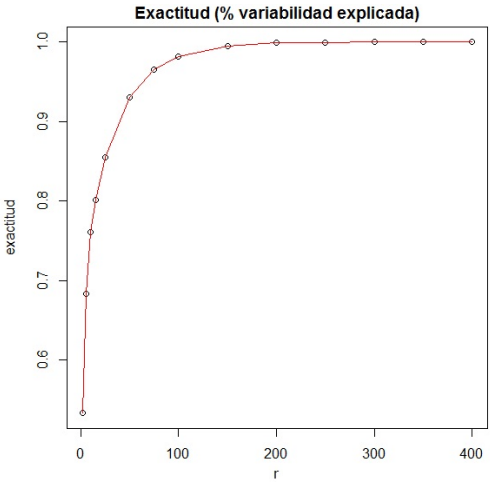
Compresión de imágenes usando PCA.

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

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

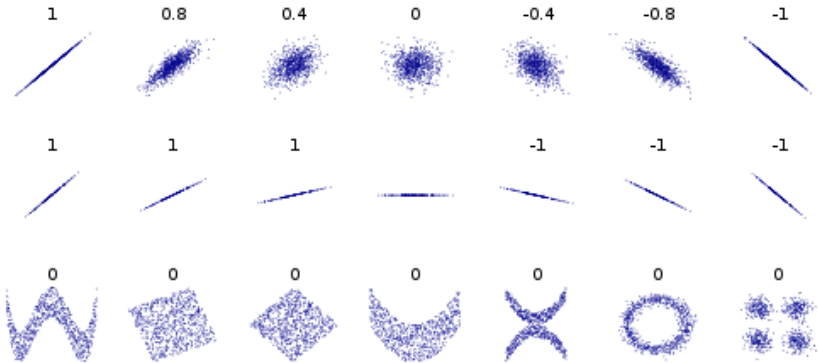
Ejemplos


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

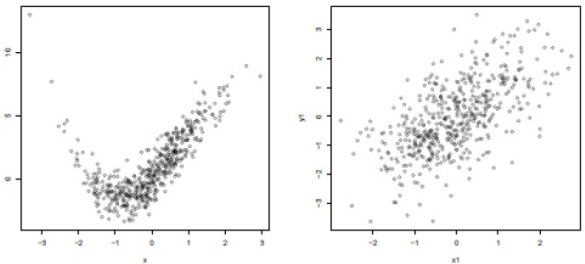

Exactitud (% variabilidad explicada)
exactitud
r
0.6
0.7
0.8
0.9
1.0
0
100
200
300
400

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