UFM UNIVERSIDAD FRANCISCO MARROQUÍN
VERITAS • LIBERTAS • IVSTITIA

FACULTAD de CIENCIAS ECONÓMICAS

Análisis de componentes principales (PCA)

Alan Reyes-Figueroa

Elements of Machine Learning (Aula 06) 07.febrero.2023

Componentes principales

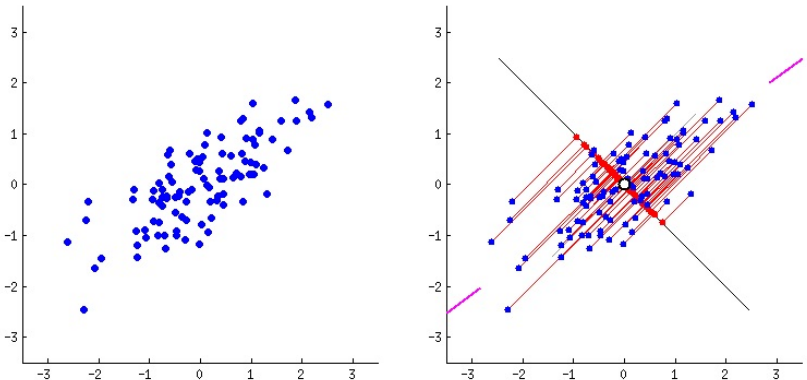
Objetivo: encontrar una estructura subyacente en los datos.

- Proyectar a un subespacio adecuado.

Componentes principales

Objetivo: encontrar una estructura subyacente en los datos.

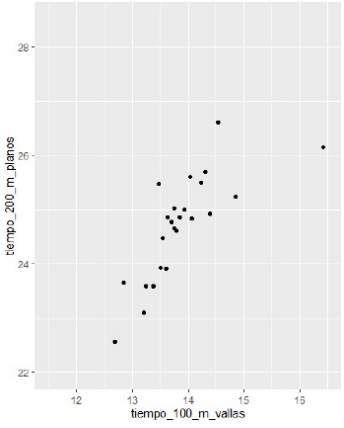
- Proyectar a un subespacio adecuado.



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

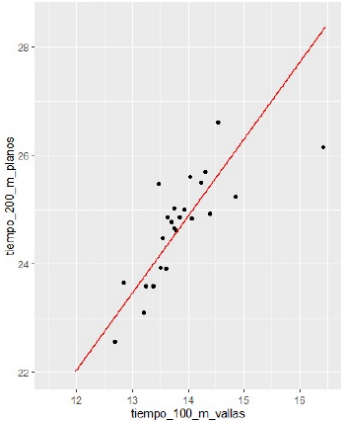


Se observa cierta estructura.

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



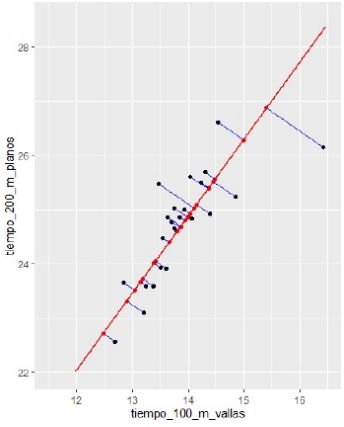
Se observa cierta estructura.

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.

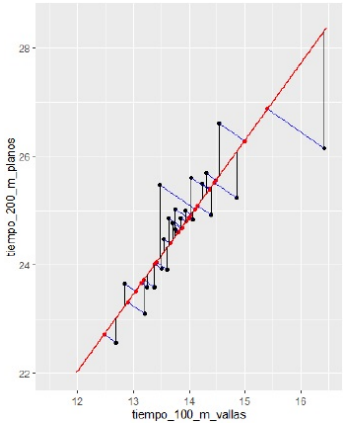
Karl Pearson (1901), describir con una recta.

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.

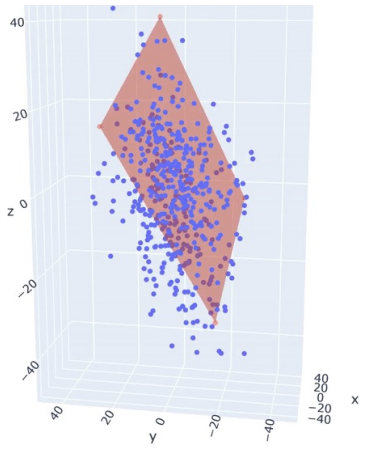
Hotelling (1933), relación entre variables $g(X_1, X_2)$.

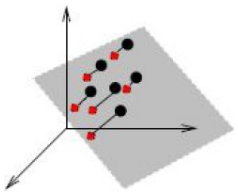
No confundir con regresión, Incorporar incertidumbre.

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

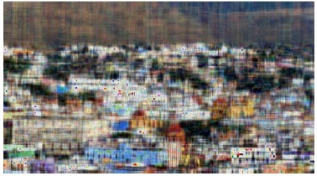




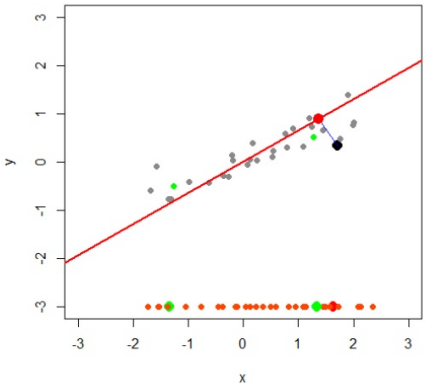
UFM UNIVERSIDAD FRANCISCO MARROQUÍN

Componentes principales

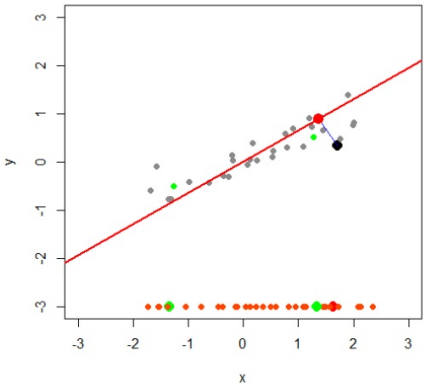
Ejemplo: Compresión de imágenes digitales.



Componentes principales

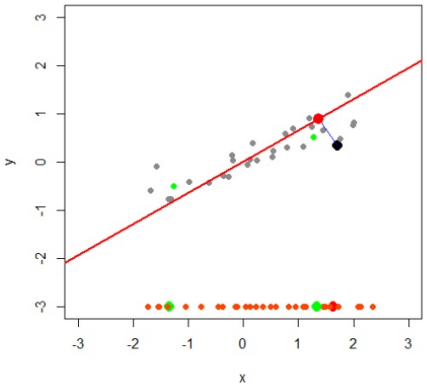


Componentes principales



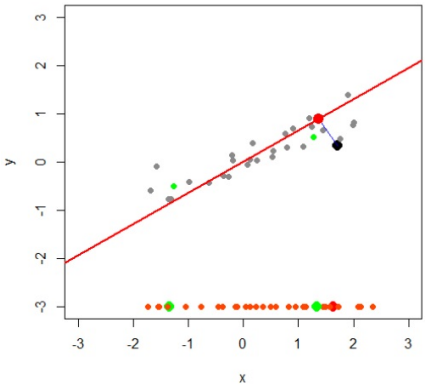
- Buscamos direcciones informativas (estructura)

Componentes principales



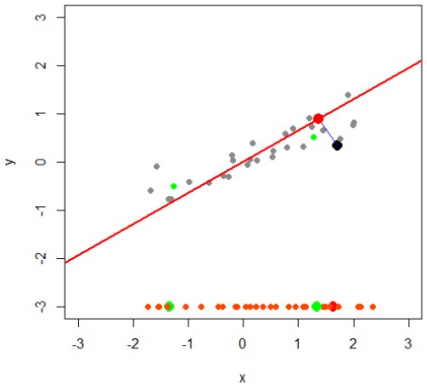
- Buscamos direcciones informativas (estructura)
  informativo = máxima variabilidad

Componentes principales

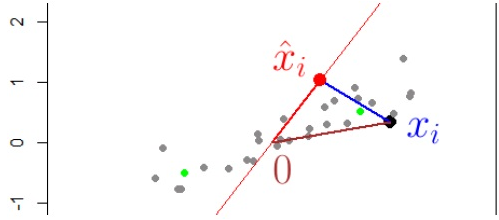


- Buscamos direcciones informativas (estructura)
  informativo = máxima variabilidad
- Buscamos minimizar el error de reconstrucción.

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

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

$$MSS = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i||^2 =$$

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

$$MSS = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i||^2 = \frac{1}{n}\sum_{i=1}^{n} ||(\mathbf{x}_i - \widehat{\mathbf{x}}_i) + \widehat{\mathbf{x}}_i||^2$$

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

$$\begin{aligned} MSS &= \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i||^2 = \frac{1}{n}\sum_{i=1}^{n} ||(\mathbf{x}_i - \widehat{\mathbf{x}}_i) + \widehat{\mathbf{x}}_i||^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \widehat{\mathbf{x}}_i||^2 + \frac{1}{n}\sum_{i=1}^{n} ||\widehat{\mathbf{x}}_i||^2 \end{aligned}$$

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

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

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

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

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 distribución $\mathbb{P}_X$.
- Supondremos también que $\mathbb{E}(X) = \mathbf{0}$ (los datos están centrados). En consecuencia, $Var(X) = \mathbb{X}^T\mathbb{X}$.

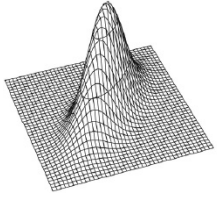
Componentes principales

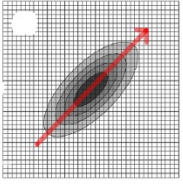
Caso particular 1D: (proyectamos a un subespacio 1-dimensional).

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. 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

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

Componentes principales

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.