Análisis de componentes principales

Emilio Carrizosa

Sevilla, 9 de marzo de 2023

Análisis de componentes principales

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London* *.

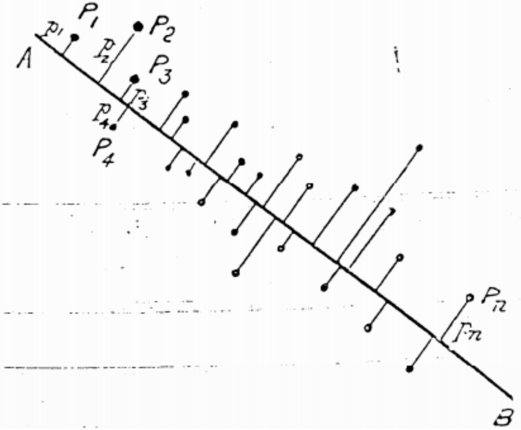
(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$y=a_0+a_1x, \quad \text{or} \quad z=a_0+a_1x+b_1y,$$

$$\text{or} \quad z=a_0+a_1x_1+a_2x_2+a_3x_3+\ldots+a_nx_n,$$

where $y, x, z, x_1, x_2, \ldots x_n$ are variables, and determining the "best" values for the constants $a_0, a_1, b_1, a_0, a_1, a_2, a_3, \ldots a_n$ in relation to the observed corresponding values of the variables. In nearly all the cases dealt with in the text-books

(y' being the ordinate of the theoretical line at the point x which corresponds to y), had we wanted to determine the best-fitting line in the usual manner.



Geometría euclídea básica

Sea $c \in \mathbb{R}^p$, $c^\top c = 1$

- Para $x \in \mathbb{R}^p$, la proyección $\pi_c(x)$ de x en la recta $\{tc : t \in \mathbb{R}\}$: resolviendo $\min_t \underbrace{\|x - tc\|_2^2}_{t^2 \cdot 1 + \|x\|^2 - 2tc'x}$

$t^2 \cdot 1 + \|x\|^2 - 2tc'x \xrightarrow{\frac{d}{dt} = 0} 2t - 2c'x = 0 \Rightarrow t^* = c'x$

$$\pi_c(x) = (c^\top x)c$$

-

$$\|x - \pi_c(x)\|_2^2 = \|x\|_2^2 + \left(c^\top x\right)^2 \|c\|_2^2 - 2x^\top \left(c^\top x\right) c = \|x\|_2^2 - \left(c^\top x\right)^2$$

- Sea X un vector aleatorio en $\mathbb{R}^p$ con $E(X) = 0$

$$E\left(\|X - \pi_c(X)\|_2^2\right) = E\left(\|X\|_2^2\right) - E\left(\left(c^\top X\right)^2\right) = E\left(\|X\|_2^2\right) - var\left(c^\top X\right)$$

$= E(\|X\|^2) - [var(c'X) + E^2(c'X)]$

Primera componente principal

slide anterior

$$\begin{array}{ll} \min & E\left(\|X-\pi_c(X)\|_2^2\right) = \min\left[E(\|X\|_2^2) - Var(c'X)\right] \\ \text{s.a} & c^\top c = 1 \end{array}$$

sin c

- min = max

$$\begin{array}{ll} \max & var\left(c^\top X\right) \\ \text{s.a} & c^\top c = 1 \end{array}$$

$Var[c'X] = c\,Var(X)c' =$

$= c[E(XX') - \underbrace{E(X)E(X)'}_{\text{Hemos supuesto cero}}]c'$

$$\begin{array}{ll} \max & c^\top E\left(XX^\top\right)c \\ \text{s.a} & c^\top c = 1 \end{array}$$

$\Sigma = Var(X) = E(XX') - \cancel{E(X)E(X)}$ 0

$$\begin{array}{ll} \max & c^\top \Sigma c \\ \text{s.a} & c^\top c = 1 \end{array}$$

Primera componente principal

$$\begin{array}{ll} \max & c^\top \Sigma c \\ \text{s.a} & c^\top c = 1 \end{array}$$

Σ semid posi

- Problema de maximización cuadrática convexa con restricciones no lineales
- KKT:
  - $\mathcal{L}(c, \lambda) = c^\top \Sigma c - \lambda(c^\top c - 1)$
  - $\nabla_c \mathcal{L} = 0 : \Sigma c = \lambda c.$
    - λ : autovalor de Σ
    - c : autovector unitario asociado
    - $c^\top \Sigma c = \lambda c^\top c = \lambda$
  - $$\begin{array}{ll} \max & \lambda \\ \text{s.a} & \Sigma c = \lambda c \\ & c^\top c = 1 \end{array}$$

    función objetivo en el óptimo: $c' \Sigma c = c' \lambda c = \lambda$
  - Solución óptima: c : autovector unitario, asociado al mayor autovalor λ_1 de Σ
  - c : Primera componente principal

Bondad de ajuste. Varianza explicada

- Sea c^1 : primera componente principal, asociada al mayor autovalor (λ_1) de Σ

  $E[\sum_j X_j^2] = \sum_j E[X_j]^2 = \sum_j \sigma_j^2 = tr(\Sigma)$

  Sol óptima
- $E\left(\|X - \pi_{c^1}(X)\|_2^2\right) = E\left(\|X\|_2^2\right) - var\left(c^\top X\right) = tr(\Sigma) - \lambda_1$
- $\frac{E\left(\|X-\pi_{c^1}(X)\|_2^2\right)}{E\left(\|X\|_2^2\right)} = 1 - \frac{\lambda_1}{\sum_{j=1}^p \lambda_j}$

  $c'\Sigma c = c'\lambda c = c'c\lambda = \lambda$
- $\frac{\lambda_1}{\sum_{j=1}^p \lambda_j}$: fracción de varianza explicada por la primera componente principal

k componentes principales

Construidas las j primeras componentes principales, $c^1, \ldots, c^j$, la componente principal $j+1$ es la solución óptima del problema de encontrar el vector unitario c, ortogonal a $c^1, \ldots, c^j$, que maximice la varianza de la proyección $c^\top X$:

$$\begin{array}{ll} \max & c^\top \Sigma c \\ & c^\top c^i = 0 \quad i = 1, 2, \ldots, j \\ & c^\top c = 1 \end{array}$$

La componente principal j $(1 \leq j \leq n)$ es el autovector unitario asociado al j-ésimo mayor autovalor de Σ

Nunca hay solución única: por ejemplo, si c^j es componente principal, también lo sería $-c^j$. Abusaremos del lenguaje

Las puntuaciones $\left(c^i)\right)^\top X$, $\left(c^j\right)^\top X$: incorreladas: $cov\left((c^i)^\top X, (c^j)^\top X\right) = E\left((c^i)^\top X\right)\left((c^j)^\top X\right)\big) = E\left((c^i)^\top X\right)\left(X^\top c^j\right)\big) = \lambda_j E\left((c^i)^\top c_j\right) = 0$

Bondad de ajuste. Varianza explicada

Para cada k, $1 \leq k \leq p$,

- las componentes principales $c^1, \ldots, c^k$ son una base ortonormal del espacio V de dimensión k que minimiza la esperanza del cuadrado de la distancia euclídea entre X y su proyección $\pi_V(X)$ sobre V
- $\frac{E\left(\|X-\pi_V(X)\|_2^2\right)}{E\left(\|X\|_2^2\right)} = 1 - \frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^p \lambda_j}$
- $\frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^p \lambda_j}$: fracción de varianza explicada por las k primeras componentes principales

$\underline{X} \rightsquigarrow (c^{1\prime}\underline{X}\, c^1, \ldots, c^{p\prime}\underline{X}\, c^p)$ coordenadas de $\underline{X}$ con base formada por componentes principales

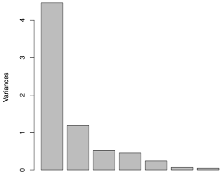
$\underline{X} \rightsquigarrow (c^{1\prime}\underline{X}, \ldots, c^{p\prime}\underline{X})$ (puntuaciones scores)

para $i \neq j$ $\operatorname{cov}(c^{i\prime}\underline{X}, c^{j\prime}\underline{X}) = E[(c^{i\prime}\underline{X} - E(c^{i\prime}\underline{X}))(c^{j\prime}\underline{X} - E(c^{j\prime}\underline{X}))] =$

$= E[c^{i\prime}\underline{X}\; c^{j\prime}\underline{X}] = E[c^{i\prime}\underline{X}\,\underline{X}'c^j]$

El problema de selección del k

- Tomar $k = p$ componentes principales: cambio de sistema de referencia ortonormal
- Cuanto mayor k, mayor varianza explicada (= menor error al reemplazar X por $\pi(X)$), pero peor interpretabilidad de resultados



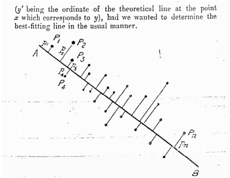
- Criterios:
  - Seleccionar k tal que se explique al menos una fracción de varianza (por ejemplo, el 80%)
  - Seleccionar k tal que la varianza explicada dé el mayor salto al pasar de k a $k+1$
  - ...

El problema de la escala

- Un cambio de escala (AX en lugar de X) hace que cambien las componentes principales: autovectores de $A\Sigma A^\top$ y no de Σ
- Aplicar análisis de componentes principales a AX es lo mismo que aplicarlo a X, pero tomando, en vez de la matriz de covarianzas Σ, la matriz $A\Sigma A^\top$.
- En particular, podemos hacer análisis de componentes principales
  - con la matriz de covarianzas
  - con la matriz de correlaciones

Componentes principales muestrales

Tenemos $x_1, \ldots, x_n \in \mathbb{R}^p$, que entendemos realizaciones de un vector aleatorio X en $\mathbb{R}^p$.

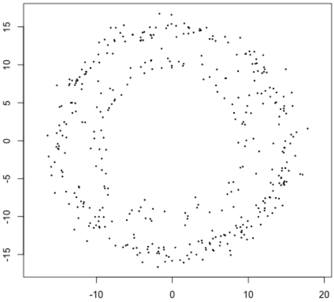


- Las componentes principales $c^1, \ldots, c^k$ son una base ortonormal del espacio V de dimensión k que minimiza la media del cuadrado de la distancia euclídea entre los x_i y su proyección $\pi_V(x_i)$ sobre V
- Se corresponden con los autovectores unitarios asociados a los k mayores autovalores λ_j de la matriz de covarianzas muestrales
- $\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$: fracción de varianza explicada por las k primeras componentes principales

Análisis en componentes principales no lineales. KPCA

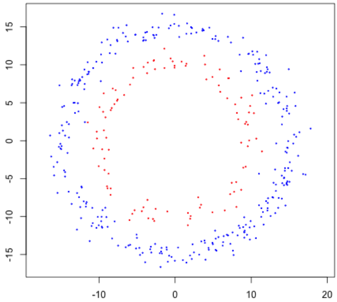
Un ejemplo

- $x_1, \ldots, x_{n_1+n_2} \in \mathbb{R}^{30}$
- $x_i = (10\cos(\theta_i) + \varepsilon_{i1}, 10\sin(\theta_i) + \varepsilon_{i2}, \varepsilon_{i3}, \ldots, \varepsilon_{i\,30})\,,\ i = 1, \ldots, n_1 = 100$
- $x_i = (15\cos\theta_i + \varepsilon_{i1}, 15\sin\theta_i + \varepsilon_{i2}, \varepsilon_{i3}, \ldots, \varepsilon_{i\,30})\,,$
  $i = n_1 + 1, \ldots, n_1 + n_2 = 400$
- $\theta_i \sim \mathcal{U}(0, 2\pi)$, $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$, $i = 1, \ldots, n_1 + n_2$, $j = 1, \ldots, 30$



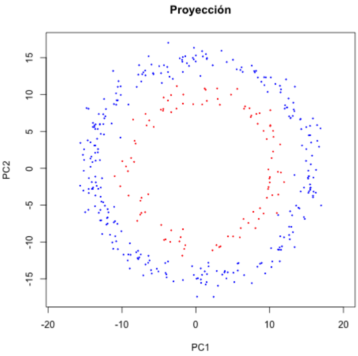
Un ejemplo

- $x_1, \ldots, x_{n_1+n_2} \in \mathbb{R}^{30}$
- $x_i = (10\cos(\theta_i) + \varepsilon_{i1}, 10\sin(\theta_i) + \varepsilon_{i2}, \varepsilon_{i3}, \ldots, \varepsilon_{i\,30})\,,\ i = 1, \ldots, n_1 = 100$
- $x_i = (15\cos\theta_i + \varepsilon_{i1}, 15\sin\theta_i + \varepsilon_{i2}, \varepsilon_{i3}, \ldots, \varepsilon_{i\,30})\,,$
  $i = n_1 + 1, \ldots, n_1 + n_2 = 400$
- $\theta_i \sim \mathcal{U}(0, 2\pi)$, $\varepsilon_{ij} \sim \mathcal{N}(0,1)$, $i = 1, \ldots, n_1 + n_2$, $j = 1, \ldots, 30$



PCA:

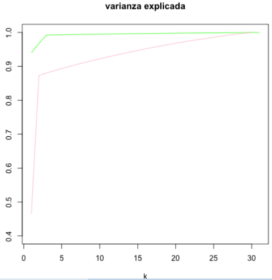
- Fracción de varianza explicada: 0.465 0.873 0.880 0.887 0.893 0.900, 0.906 0.911 0.917 0.922 0.927 0.933 0.938 0.942 ...



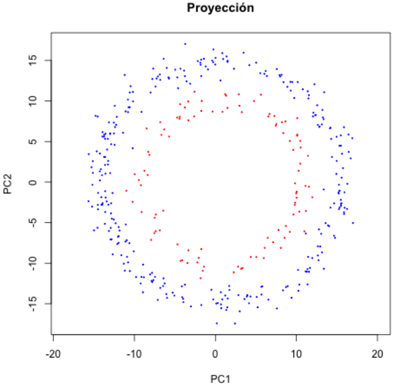
Análisis en componentes principales no lineales. KPCA

Un ejemplo

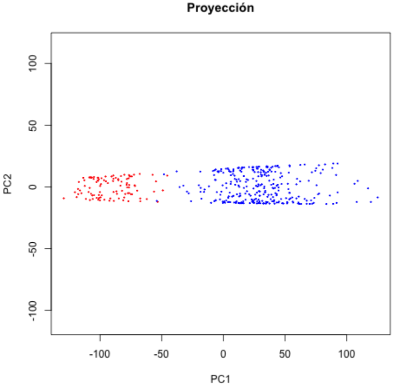
- $x_1, \ldots, x_{n_1+n_2} \in \mathbb{R}^{31}$
- $x_i = \left(10\cos(\theta_i) + \varepsilon_{i1}, 10\sin(\theta_i) + \varepsilon_{i2}, \varepsilon_{i3}, \ldots, \varepsilon_{i\,30}, x_{i1}^2 + x_{i2}^2\right)$, $i = 1, \ldots, n_1 = 100$
- $x_i = \left(15\cos\theta_i + \varepsilon_{i1}, 15\sin\theta_i + \varepsilon_{i2}, \varepsilon_{i3}, \ldots, \varepsilon_{i\,30}, x_{i1}^2 + x_{i2}^2\right)$, $i = n_1 + 1, \ldots, n_1 + n_2 = 400$
- $\theta_i \sim \mathcal{U}(0, 2\pi)$, $\varepsilon_{ij} \sim \mathcal{N}(0,1)$, $i = 1, \ldots, n_1 + n_2$, $j = 1, \ldots, 30$
- Fracción de varianza explicada: $0.940\,0.968\,0.992\,0.993\,\ldots$



PCA en $\mathbb{R}^{30}$



PCA en $\mathbb{R}^{31}$



- Podemos hacer el PCA tras una inmersión de los datos $x_1, \ldots, x_n \in \mathbb{R}^p$ con una función ϕ
- Los datos deben estar centrados:

$$\{x_1, \ldots, x_n\} \hookrightarrow \left\{ \phi(x_1) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), , \ldots, \phi(x_n) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) \right\}$$

- Componentes principales: autovectores de

$$\left(\phi(x_1) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \ldots, \phi(x_n) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right)\left(\phi(x_1) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \ldots, \phi(x_n) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right)^{\top}$$

- Sean $\Phi = (\phi(x_1), \ldots, \phi(x_n))$, $E = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \vdots & & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$. Las componentes principales: autovectores de

No será $\hat{\Sigma}$

sino $\hat{\Sigma} = S$ (no?)

$$\left(\Phi - \Phi\frac{1}{n}E\right)\left(\Phi - \Phi\frac{1}{n}E\right)^{\top} = \Phi\left(I - \frac{1}{n}E\right)^{2}\Phi^{\top}$$

- Sea u : componente principal, esto es: autovector unitario de $\Phi\left(I-\frac{1}{n}E\right)^2\Phi^\top$. Entonces

$$\Phi\left(I-\frac{1}{n}E\right)^2\Phi^\top u=\lambda u$$

- $u=\Phi\underbrace{\frac{1}{\lambda}\left(I-\frac{1}{n}E\right)^2\Phi^\top u}_{\omega}$
- $u=\sum_{i=1}^n \omega_i\phi(x_i)$
- La puntuación de un cierto x:

$$u^\top\left(\phi(x)-\frac{1}{n}\sum_{j=1}^n\phi(x_j)\right)=\sum_{i=1}^n\omega_i\phi(x_i)^\top\phi(x)-\frac{1}{n}\sum_{i,j=1}^n\omega_i\phi(x_i)^\top\phi(x_j)$$

$$\sum_i w_i\phi(x_i)\left[\phi(x)-\frac{1}{n}\sum_j\phi(x_j)\right]$$

Núcleos

$$u^\top \left(\phi(x) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j) \right) = \sum_{i=1}^{n} \omega_i \phi(x_i)^\top \phi(x) - \frac{1}{n} \sum_{i,j=1}^{n} \omega_i \phi(x_i)^\top \phi(x_j)$$

Núcleos

$$u^\top \left(\phi(x) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j) \right) = \sum_{i=1}^{n} \omega_i \phi(x_i)^\top \phi(x) - \frac{1}{n} \sum_{i,j=1}^{n} \omega_i \phi(x_i)^\top \phi(x_j)$$

No necesitamos conocer ϕ. Solo necesitamos conocer la función núcleo (kernel) $K : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$, $K(x,z) = \phi(x)^\top \phi(z)$

Núcleos

$$u^\top \left(\phi(x) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j) \right) = \sum_{i=1}^{n} \omega_i \phi(x_i)^\top \phi(x) - \frac{1}{n} \sum_{i,j=1}^{n} \omega_i \phi(x_i)^\top \phi(x_j)$$

No necesitamos conocer ϕ. Solo necesitamos conocer la función núcleo (kernel) $K : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$, $K(x,z) = \phi(x)^\top \phi(z)$

Puntuación de x según componente principal:

$$u^\top \left(\phi(x) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j) \right) = \sum_{i=1}^{n} \omega_i K(x_i, x) - \frac{1}{n} \sum_{i,j=1}^{n} \omega_i K(x_i, x_j)$$

Núcleos

Posibles núcleos K :

- $K(x,z) = x^\top z$ (lineal)
- $K(x,z) = (1 + x^\top z)^d$ (polinómico)
- $K(x,z) = exp(-\gamma \|x - z\|_2^2)$ (gaussiano; RBF)
- ...

- Para evaluar las puntuaciones no necesitamos ϕ, sino K. Pero para eso necesitamos calcular los autovectores de $\Phi\left(I-\frac{1}{n}E\right)\left(I-\frac{1}{n}E\right)^\top\Phi^\top$

- Si z es autovector de $M^\top M$ con autovalor $\lambda\neq 0$, entonces Mz es autovector de $MM^\top$ con autovalor λ

- Si z es autovector de $\left(I-\frac{1}{n}E\right)^\top \Phi^\top\Phi \left(I-\frac{1}{n}E\right)$ para un autovalor $\lambda>0$, entonces $\Phi\left(I-\frac{1}{n}E\right)z$ es (salvo normalización) componente principal con autovalor λ.
- $\Phi^\top\Phi=\begin{pmatrix} K(x_1,x_1) & \cdots & K(x_1,x_n)\\ \vdots & \ddots & \vdots \\ K(x_n,x_1) & \cdots & K(x_n,x_n)\end{pmatrix}$