Análisis de componentes ppales (ACP) o (PCA)

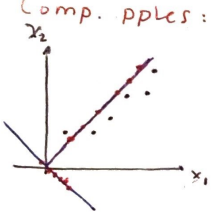
Thursday, March 31, 2022 6:52 PM

Análisis de componentes ppales (ACP) ó (PCA)

Buscamos explicar la estructura de varianza - cov. de los datos y reducir dimensionalidad.

- Si hay p-vars, necesito p componentes ppales xa capturar TODA la info.
- En general, se puede capturar gran parte de la info con $k << p$ comp. ppales.
- La BD se convierte en n mediciones de k "vars".
- Normalm. se usa PCA antes de otros métodos (reg. lín)

Comp. ppales: Las c.p son combinac°es lín de las p v.a's. $x_1, .., x_p$ q' max la varianza.



Interp geométrica.

ej) c.p define una dir dada $a_i = (a_{i1}, .., a_{ip})$ y se busca encontrar las dirs q' explican la mayor ♀ de variabilidad de los datos.

→ Aquí tenemos 2 vars y 2 dirs (\ y /)
La proyec° de los vals (..) en esas dirs están como: • • •
Busco cuál me explica mejor los vals haciendo uso de la proyec°.

Considere el vec aleat. $\mathbb{X} = \begin{bmatrix} X_1 \\ X_p \end{bmatrix}$ con matriz de cov Σ y vals propios $\lambda_1 \geq \lambda_2 \geq .. \geq \lambda_p \geq 0$.

Las comb. lin:

$$Y_1 = a_1' \mathbb{X} = a_{11} X_1 + \cdots + a_{1p} X_p$$
$$\vdots$$
$$Y_p = a_p' \mathbb{X} = a_{p1} X_1 + \cdots + a_{pp} X_p$$

$$var(Y_i) = a_i' \Sigma a_i$$
$$Cov(Y_i, Y_k) = a_i' \Sigma a_k$$

Las c.p son las comb. lín no correlacionadas (incl p) q' maximizan la varianza.

c.p: • Incorr.
• Ortogonales entre sí.

ej) la 1era c.p = máx $var(Y_1) = a_1' \Sigma a_1$

(ACP sólo sirve para cuantitativas).

LAS COMPONENTES SON LOS VECS PROPIOS

Se tiene:

- 1era c.p = $a_1' \mathbb{X}$ dónde a_1 max $var(a_1' \mathbb{X})$
- i-ésima c.p = $a_i' \mathbb{X}$ dónde a_i max $var(a_i' \mathbb{X})$
  sujeto a $cov(a_i' \mathbb{X}, a_k \mathbb{X}) = 0$, $k < i$
  y a_i ortogonal a $a_1, .., a_k$.

Sea Σ la matriz de varianza-cov de $\mathbb{X}$
Suponga q' Σ tiene los pares de vals, vect. propios:

$(\lambda_1, e_1), .., (\lambda_p, e_p)$ $\lambda_1 \geq .. \geq \lambda_p \geq 0$.

★ La i-ésima c.p es:

$$Y_i = e_i' \mathbb{X}$$

$$var(Y_i) = e_i' \Sigma e_i = \lambda_i \qquad i = 1, .., p$$

$Cov(Y_i, Y_k) = 0$ → Las c.p son incorr.

* Si $Y_i = Y_k$ es como si fueran la misma.

Sea Σ la matriz de var-cov de $\mathbb{X}$, con pares de vals y vecs propios.

$(\lambda_1, e_1), .., (\lambda_p, e_p)$ $\lambda_1 \geq .. \geq \lambda_p \geq 0$.

Sean $Y_1, .., Y_p$ las c.p. Entonces,

$$\sigma_{11} + \cdots + \sigma_{pp} = var(X_1) + .. + var(X_p) = \sum var(X_i)$$
$$= \lambda_1 + .. + \lambda_p = \sum var(Y_i)$$

Obs:

1. La varianza total poblacional se preserva.
2. La propor° de la variabilidad total poblacional explicada por la k-ésima C.P es: $\frac{\lambda_k}{\lambda_1 + \cdots + \lambda_p}$ → la C.P. k explica este % de los datos.
3. En muchos casos, para las primeras C.P esta propor° es del 0.8 - 0.9 (con las 4 primeras ya explico poco más del 70% de la info).

(Lo Q de C.P depende del nivel de correlac° de las vars)

4. $e_i = \begin{bmatrix} e_{i1} \\ e_{ip} \end{bmatrix}$, $|e_{ik}|$ "mide" la importancia de la var k en la C.P i.

Sea $Y_i = e_i' X_1, \ldots, Y_p = e_p' X$ las C.P, entonces:

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$ → correla° entre mi nueva Y_i y mi og X_k.

- ρ grande → la X_k tiene mucho peso en Y_i (tiene info).
- $\rho > 0$: Si en X_k hay un val alto, espero un val alto en la C.P Y_i
- $\rho < 0$: Si X_k tiene val alto, espero q' C.P Y_i tenga uno bajo.

Estandarizar?

Si no estandarizo, puedo estar dándole más del peso q' cubería a las vars con mucha variabilidad.

No obstante, deshacerme de esa mayor variabilidad, podría estar quitándome info.

Se tienen las vars: $z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}$, $z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}$

$$Z = (V^{1/2})^{-1}(X - \mu)$$ (matriz de desv. st.)

- $E[Z] = 0$; $\text{cov}(Z) = \rho$

Los comp. ppales de Z se obtienen con los vals y vect propios de ρ.

La i-ésima C.P de Z está dada por:

- $Y_i = e_i' Z = e_i' (V^{1/2})^{-1}(X - \mu)$
- $\sum_p \text{var}(Y_i) = \sum \text{var}(Z_i) = 1 \cdot p = p$.
- $\rho_{Y_i, Z_k} = e_{ik}\sqrt{\lambda_i}$

Prop. de variabilidad explicada por la C.P k

$$= \frac{\lambda_k}{p} = \frac{\lambda_k}{\sum \text{var}(Y_i)}$$