

Tema 34

1. Análisis de Componentes Principales.
2. Formulación del Problema, resolución y propiedades.
3. Determinación del número de componentes a considerar.

1. ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales tiene por objetivo reducir la dimensión del problema a costa de una pequeña pérdida de información.

Se analiza si es posible representar adecuadamente la información de p variables con un número menor de variables ($r < p$) construidas como combinación lineal de las originales.

La técnica de componentes principales es debida a Hotelling (1933), aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por K. Pearson (1901). Su utilidad es doble:

- 1.- Permite representar óptimamente en un espacio de dimensión pequeña (r) observaciones de un espacio p -dimensional. Es un primer paso para identificar variables latentes, o no observadas que generan los datos.
- 2.- Permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos.

El método de componentes principales se inscribe dentro de la estadística descriptiva ya que es una herramienta exploratoria. El problema de inferir si las propiedades de reducción de la dimensión encontradas en los datos puedan extenderse a la población de la que provienen los datos se estudia en el análisis factorial.

2.. FORMULACIÓN DEL PROBLEMA, RESOLUCIÓN Y PROPIEDADES

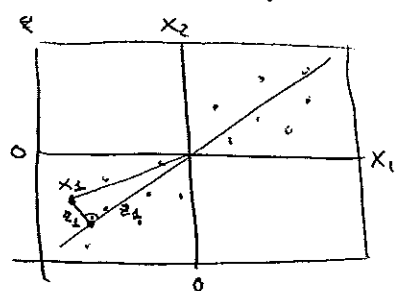
2.1. FORMULACIÓN DEL PROBLEMA

En el análisis de componentes principales se dispone de una muestra de tamaño n con p variables X_1, \dots, X_p (tipificadas o expresadas en desviaciones respecto de su media) inicialmente correlacionadas, para posteriormente obtener a partir de ellas un número $k, 2 \leq k \leq p$ de variables incorrelacionadas Z_1, Z_2, \dots, Z_k que sean combinación ~~lineal~~ lineal de las variables iniciales y que expliquen la mayor parte de su variabilidad.

Euforo descriptivo

Se desea encontrar un subespacio de dimensión menor que p tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible.

Se considere primero una recta, se desea que las proyecciones de los puntos sobre esta recta mantengan lo más posible, sus posiciones relativas.



ej: $(p=2)$

(la recta está entre las dos rectas de regresión)

La condición de que la recta pase cerca de la mayoría de los puntos puede concretarse exigiendo que las distancias entre los puntos originales y sus proyecciones sobre la recta sean lo más pequeñas posibles.

Sea x_i un punto y $a_1 = (a_{11}, \dots, a_{1p})'$ una dirección, definida por un vector \vec{a}_1 de norma unidad, la proyección del

punto x_i sobre esta dirección es el escalar:

$$z_i = a_{i1}x_{i1} + \dots + a_{ip}x_{ip} = \vec{a}_1' x_i$$

y el vector que representa esta proyección es: $z_i \vec{a}_1$

Sea r_i la distancia entre x_i y su proyección sobre la dirección \vec{a}_1 , la mejor recta es la que cumple:

$$\text{minimizar } \sum_{i=1}^n r_i^2 = \min \sum_{i=1}^n \underbrace{|x_i - z_i \vec{a}_1|^2}_{\text{norma euclídea o módulo del vector}}$$

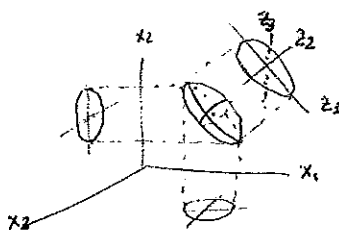
$$\min \sum_{i=1}^n r_i^2 \stackrel{\text{equiv.}}{=} \min \sum_{i=1}^n z_i^2 \stackrel{\text{equiv.}}{=} \text{Maximizar su varianza}$$

\uparrow
 $\frac{x_i' x_i}{n} = z_i^2 + r_i^2$ (te^a Pitágoras)
 \downarrow
 $\hookrightarrow z_i$ su var. de media 0


Teorema estadístico

Representar puntos p dimensionales con la mínima pérdida de información en un espacio de dimensión uno es equivalente a sustituir las p variables originales por una nueva variable, z_1 , que resume óptimamente la información. Esto supone que la nueva variable debe tener globalmente máxima correlación con las originales, es decir, la variable de máxima variabilidad.

La segunda variable z_2 debe ser incorrelada con z_1 y su varianza máxima. En general, la componente z_k ($k < p$) tendrá varianza máxima entre todas las combinaciones lineales de las p variables originales, con la condición de estar incorreladas con las z_1, \dots, z_{k-1} previamente obtenidas.



Geometría geométrica

Si se observa el diagrama de dispersión , los puntos se sitúan siguiendo una elipse y se pueden describir por su proyección en la dirección del eje mayor de la elipse. Puede demostrarse que este eje es la recta que minimiza las distancias ortogonales. En varias dimensiones los se tendrían elipsoides, y la mejor aproximación a los datos es la proporcionada por su proyección sobre el eje mayor del elipsoide.

2.2. ~~CÁLCULO DE U~~

RESOLUCIÓN: CÁLCULO DE LAS COMPONENTES

Cálculo de la primera componente

La primera componente principal se define como la combinación lineal de las variables originales que tiene varianza máxima (z_1)

$$z_1 = X a_1$$

$$z_{1i} = a_{11} x_{1i} + a_{12} x_{2i} + \dots + a_{1p} x_{pi}$$

Como las var. originales tienen media cero, también z_1 tendrá media nula.

$$V(z_1) = \frac{1}{n} \sum_{i=1}^n z_{1i}^2 = \frac{1}{n} z_1' z_1 = \frac{1}{n} a_1' X' X a_1 = a_1' \underbrace{\left[\frac{1}{n} X' X \right]}_{= S \equiv \text{matriz de var.-cov de las observ.}} a_1$$

La primera componente se obtiene de forma que su varianza sea máxima, sujeta a la restricción de que la suma de los pesos (a_{ij}) al cuadrado sea igual a la unidad:

$$\begin{array}{l} \max V(z_1) = a_1' S a_1 \\ \text{s.a.} \quad a_1' a_1 = 1 \end{array} \left\{ \Rightarrow \begin{array}{l} \text{Lagrangiano} \\ L = a_1' S a_1 - \lambda (a_1' a_1 - 1) \end{array} \right.$$

$\frac{\partial L}{\partial a_1} = 2S a_1 - 2\lambda a_1 = 0 \Leftrightarrow S a_1 = \lambda a_1 \Leftrightarrow a_1' S a_1 = a_1' \lambda a_1 = \lambda$
 $(S - \lambda I) a_1 = 0$ $\left\{ \begin{array}{l} \text{la vector propio de } S \\ \text{valor propio de } S \text{ corresp a } a_1 \end{array} \right.$
 $\Rightarrow \lambda$ es la varianza de $z_1 \Rightarrow \lambda$ será el mayor valor propio de S . y su vector asociado a_1 define los coeficientes de cada variable en el primer componente principal.

Cálculo de la segunda componente

La segunda componente, al igual que las restantes, se expresa como combinación lineal de las var. originales:

$$z_2 = X a_2$$

$$z_{2i} = a_{21} x_{1i} + a_{22} x_{2i} + \dots + a_{2p} x_{pi}$$

$$V(z_2) = a_2' \left[\frac{1}{n} X'X \right] a_2 = a_2' S a_2$$

la segunda componente se obtiene maximizando su varianza sujeto a que la suma de los pesos al cuadrado sea igual a la unidad y a que a_1 y a_2 sean ortogonales:

$$\begin{array}{l} \max V(z_2) = a_2' S a_2 \\ \text{s.a.} \quad a_2' a_2 = 1 \\ \quad \quad a_2' S a_1 = 0 \end{array} \left\{ \Rightarrow L = a_2' S a_2 - 2\mu (a_2' S a_1) - \lambda (a_2' a_2 - 1) \right.$$

$$\frac{\partial L}{\partial a_2} = 0 \Leftrightarrow (S - \lambda I) a_2 = 0 \Rightarrow \text{se toma } \lambda \text{ como el segundo}$$

mayor valor propio de S y a_2 su vector propio asociado normalizado.

Generalización

Puede demostrarse análogamente que el espacio de dimensión r que mejor representa a los puntos viene definido por los vectores propios asociados a los r mayores valores propios de S . Estas direcciones se denominan direcciones principales de los datos y a las nuevas variables definidas por ellas componentes principales.

La matriz S tiene rango p , por lo que existen tantas comp. principales como variables obtenidas al calcular los valores propios o raíces características, $\lambda_1, \dots, \lambda_p$, de la matriz de var-cov. de las variables (S) mediante:

$$|S - \lambda I| = 0$$

y sus vectores asociados son:

$$(S - \lambda_i I)a_i = 0$$

Como S def. positiva y simétrica $\Rightarrow \lambda_i > 0$ y reales

las nuevas variables Z están relacionadas con las originales mediante:

$$Z = XA \quad \text{t.q.} \quad A^T A = I$$

→ Calcular las comp. principales equivale a aplicar una transformación ortogonal A a las var. X (ejes originales) para obtener unas nuevas variables Z incorreladas entre sí.

Esta operación puede interpretarse como elegir una nueva ejes coordenados, que coincidan con los "ejes naturales" de los datos.

2.3 PROPIEDADES DE LAS COMPONENTES

Las componentes principales son nuevas variables con las siguientes propiedades:

1.- Conservan la variabilidad inicial

$$\sum_{i=1}^p \text{Var}(x_i) = \sum_{i=1}^p \text{Var}(z_i)$$

2.- la proporción de variabilidad aplicada por una componente es el cociente entre su varianza (\equiv el valor propio asociado al vector propio que la define) y la suma de los valores propios de la matriz

$$\frac{\lambda_i}{\sum_{i=1}^p \lambda_i}$$

3.- las covarianzas entre cada componente principal y las variables X vienen dadas por el producto de las coordenadas del vector propio que define la componente por su valor propio:

$$\text{Cov}(z_i; x_1, \dots, x_p) = \lambda_i a_i = (\lambda_i a_{i1}, \dots, \lambda_i a_{ip})$$

(dem. en D. Peña pg. 146)

4.- la correlación entre una componente principal y una variable X es proporcional al coeficiente de esa variable en la definición del componente

$$\text{Corr}(z_i; x_j) = \frac{\text{Cov}(z_i; x_j)}{\sqrt{\text{Var}(z_i) \cdot \text{Var}(x_j)}} = \frac{\lambda_i a_{ij}}{\sqrt{\lambda_i} s_j} = a_{ij} \frac{\sqrt{\lambda_i}}{s_j}$$

- 5.- las z componentes principales ($z < p$) proporcionan la predicción lineal óptima con z variables del conjunto de variables X .
- 6.- Estandarizando las componentes principales se obtiene la estandarización multivariante de los datos originales

$$Z D^{-1/2} = X A D^{-1/2} \quad \text{,, } D^{-1/2} = \text{matriz de las inversas de las desv. típicas de las compo.}$$

$$Z D^{-1/2} = X S^{-1/2}$$

3.- DETERMINACIÓN DEL NÚMERO DE COMPONENTES A CONSIDERAR

Interpretación de las componentes

Cuando existe una alta correlación positiva entre todas las variables, la primera comp. principal tiene todas sus coordenadas del mismo signo y puede interpretarse como un promedio ponderado de todas las variables, o un factor global de "tamaño".

Las restantes componentes se interpretan como factores "de forma" y típicamente tienen coord positivas y negativas, que implica que contraponen unos grupos de variables frente a otros.

Selección del número de componentes

Se han sugerido distintas reglas para seleccionar el número de componentes:

1. Gráfico de sedimentación: se obtiene al representar en ordenadas los valores característicos λ_i y en abscisas los números de las comp. principales (i) en orden decreciente.

La idea es buscar un "codo" en el gráfico, es decir, un punto a partir del cual los valores propios son aproximadamente iguales.

El criterio es quedarse con un número de componentes que excluya los asociados a valores pequeños y aproximadamente del mismo tamaño.

2. Seleccionar componentes hasta cubrir una proporción determinada de varianza (80% 90%). Esta regla es arbitraria y debe aplicarse con cierto cuidado.

3. Criterio de la media aritmética: desechar las componentes asociadas a valores propios inferiores a una cota, que suele fijarse como la varianza media, $\frac{\sum_{j=1}^p \lambda_j}{p} = \bar{\lambda}$. Cuando se trabaja con la matriz de correlaciones, $\bar{\lambda} = 1$.

1 - Análisis de Componentes Principales.

- Objetivo
- X_1, \dots, X_p var. originales correlacionadas ($\bar{X}_i = 0$)
 $\rightarrow Z_1, \dots, Z_p$ ($r < p$) Z_i c.l. de X_i
 Z_i incorrel.
- Rep. óptima + interpretación
- Descriptiva

2 - Formulación, resolución y propiedades

2.1. Formulación:

- Enfoque descriptivo
 Subespacio dim. menor de proyección óptima.
 Encontrar \vec{a}_1 t.q. $\min \sum r_i^2 = \max \sum Z_i^2$.
- Enfoque estadístico
 Encontrar Z_1 / máx corr con X_i (Azulab.)
 Encontrar Z_2 / máx corr con X_i s.a. incorrel. Z_1 .
 etc.

- Enfoque geométrico

Proyectar sobre ejes ortogonales.

2.2. Resolución:

$$Z_1 = XA_1 \quad V(Z_1) = a_1' \left(\frac{1}{n} X'X \right) a_1 = a_1' S a_1$$

$$\max V(Z_1) \quad \left\{ \begin{array}{l} \lambda_1 \text{ autorvalor} \\ \text{s.a. } a_1' a_1 = 1 \end{array} \right. \quad \left\{ \begin{array}{l} a_1 \text{ autorvector} \end{array} \right.$$

$$Z_2 = XA_2 \quad V(Z_2) = a_2' \left[\frac{1}{n} X'X \right] a_2 = a_2' S a_2$$

$$\max V(Z_2) \quad \left\{ \begin{array}{l} \lambda_2 \text{ autorvalor} \\ \text{s.a. } a_2' a_2 = 1 \end{array} \right. \quad \left\{ \begin{array}{l} a_2 \text{ autorvector} \end{array} \right.$$

etc.

$$Z = XA \quad \text{t.q.} \quad A'A = I$$

2.3. PROPIEDADES:

- 1 - $\sum \text{Var}(X_i) = \sum \text{Var}(Z_i)$.
- 2 - % var. explicada por $Z_i = \lambda_i / \sum \lambda_i$
- 3 - $\text{COV}(Z_i, X_j - \bar{X}_j) = \lambda_i a_j$
- 4 - $\text{CORR}(Z_i, X_j) = a_{ij} \frac{\sqrt{\lambda_j}}{\sqrt{\sum \lambda_j}}$
- 5 - Proyección lineal óptima.
- 6 - Estadística

3 - Número componentes.

- Gráfico scree. $\lambda_i \uparrow \rightarrow i$ (codo)
- % var. explicada. \rightarrow
- Criterio autorval > 1 (correct $\Rightarrow > 1$).