Método de Componentes Principales

1. *Introducción*
2. *Metodología Estadística*
   2.1 Estimación de las componentes principales
   2.2 Retención de las componentes principales
   2.3 Interpretación de las componentes principales
   2.4 Puntuaciones o Scores.

1. Introducción

El Análisis de Componentes Principales (ACP), es un método de análisis estadístico multivariante que estudia la dependencia estructural y las interrelaciones entre un conjunto de variables. Pertenece a las técnicas de análisis multivariadas exploratorias o descriptivas.

El método se basa en transformar el conjunto de variables originales X1, X2, ... ,Xp en un nuevo conjunto de variables ortogonales Z1, Z2, ... , Zp, denominadas componentes principales, que se caracterizan por estar incorrelacionadas entre sí y son expresadas como una combinación lineal de las variables originales.

1. Introducción

El objetivo del ACP:

- Reducir la dimensionalidad de las variables originales, tratando de explicar la mayor parte de la variabilidad total del conjunto de variables originales con el menor número posible de componentes principales.
- El ACP determina los pesos que tienen c/u de las variables en cada componente, es decir las CP se explican en función de variables observables. También puede ser aplicado como datos de entrada para otras técnicas (cluster, regresión, etc).

2.1 Estimación de las Componentes Principales

La j-ésima componente principal Zj es expresa como una combinación lineal de las p variables originales, esto es:

$$Zj = \underline{L}j\ X \qquad j=1, 2, \ldots , p$$

La estimación de la j-ésima componente principal Zj consiste en hallar la combinación lineal, tal que se maximice su variancia en orden decreciente y sujeta a la condición que los Lj sean vectores ortonormales:

$$Max\{Var(Zj)\} = Max\{\underline{L}j' \sum \underline{L}j \ / \ \underline{L}j\underline{L}i=1 \text{ si } i=j \text{ y } 0 \text{ si } i\neq j\}$$

Aplicando estimadores de Lagrange se tiene la siguiente igualdad:

$$\left| \sum - \lambda j\ I \right| \ \underline{L}j = 0$$

2.1 Estimación de las Componentes Principales

Estimación en la Muestra

Para el cálculo de los autovalores λj y autovectores Lj muestrales se utiliza el Teorema de Descomposición Espectral. Siendo S la matriz de variancias-covariancia muestral que estima a $\sum$, entonces expresando en términos de los estimadores se tiene:

$$\left| S - \hat{\lambda}_j I \right| \hat{L}_j = 0$$

donde $\hat{\lambda}_j \; y \; \hat{L}_j$ son los estimadores del autovalor λj y autovector Lj asociado respectivamente.

2.1 Estimación de las Componentes Principales

Propiedades de las Componentes Principales

1. La varianza de la componente principal j es igual a su autovector asociado.

$$Var\left(Z_j\right) = L'\Sigma L = \lambda_j$$

2. Las varianzas (autovalores) están en orden decreciente.

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p \geq 0$$

3. La suma de las varianzas de las componentes es igual a la suma de las varianzas de las variables originales.

$$\sum_{j=1}^{p} Var\left(Z_j\right) = \sum_{j=1}^{p} \lambda_j = \text{Traza}\left(\Sigma\right) = \sum_{j=1}^{p} \sigma_j^2$$

4. Si las variables originales están tipificadas:

$$\sum_{j=1}^{p} Var\left(Z_j\right) = \sum_{j=1}^{p} \lambda_j = \text{Traza}\left(R\right) = \sum_{j=1}^{p} 1 = 1+1+\ldots+1 = p$$

2.1 Estimación de las Componentes Principales

Propiedades de las Componentes Principales

1. Proporción explicada por la componente j respecto a la variación total.

$$PVE = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100, \text{ si las variables originales se han tipificado} \rightarrow PVE = \frac{\lambda_j}{p} \times 100$$

2. Proporción explicada por las k primeras componentes respecto a la variación total.

$$PVE_K = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \times 100, \text{ si las variables originales se han tipificado} \rightarrow PVE_K = \frac{\sum_{i=1}^{k} \lambda_i}{p} \times 100$$

2.2 Retención de las Componentes Principales

El objetivo del ACP es reducir la dimensionalidad de p variables originales a m componentes principales (m<p), entonces el problema es como determinar m, esto es, el número de componentes principales a ser retenidas.

A continuación se presenta varios criterios:

1) Criterio práctico: Retener las dos primeras componentes. El análisis gráfico en dos dimensiones es más fácil.
2) Criterio de la media aritmética
3) Gráfico de Sedimentación

2.2 Retención de las Componentes Principales

1. Criterio práctico: Retener las dos primeras componentes. El análisis gráfico en dos dimensiones es más fácil.
2. Criterio de la media aritmética: se retienen las componentes cuyos autovalores (eigenvalores, raíces características o varianzas) sea mayor que el promedio de todos los autovalores.

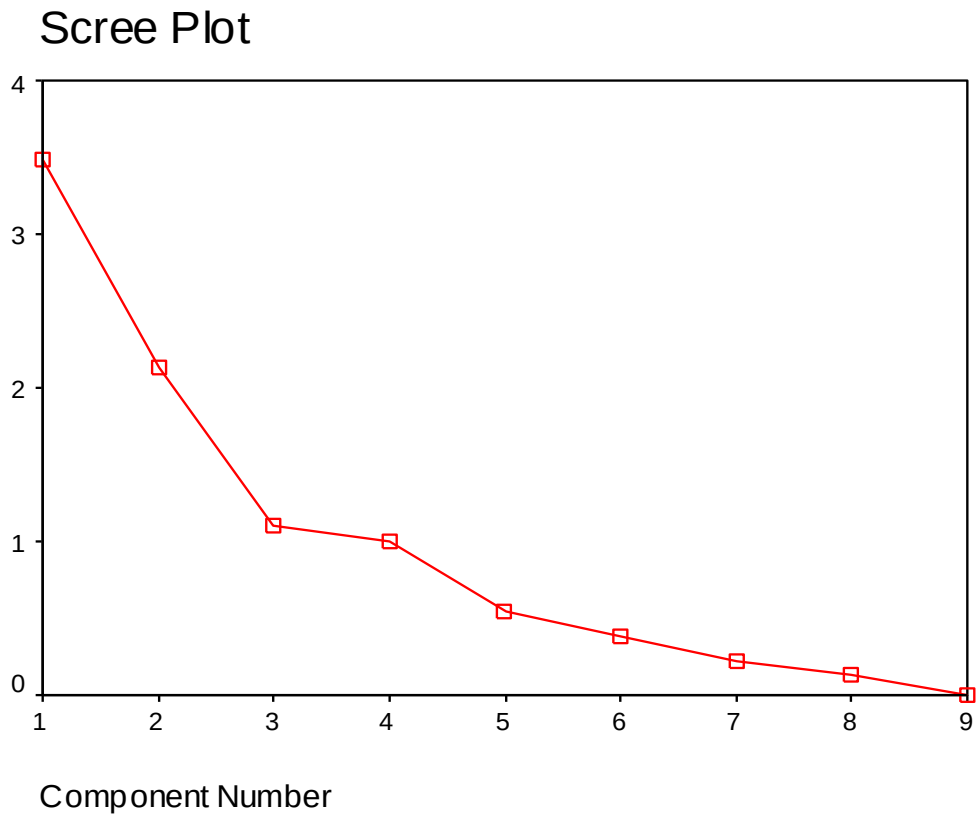
   Si $\lambda_r > \bar{\lambda} = \frac{\sum_{i=1}^{p} \lambda_i}{p}$ se retiene la componente *r*. Cuando las variables originales se han tipificado, se retiene la componente *r* si $\lambda_r > \bar{\lambda} = \frac{\sum_{i=1}^{p} \lambda_i}{p} = \frac{p}{p} = 1, \quad [\lambda_r > 1]$

3. Criterio del gráfico de sedimentación: Es la descripción gráfica del criterio de la media aritmética.

2.2 Retención de las Componentes Principales

4) Gráfico de Sedimentación

El gráfico de sedimentación (Scree plot), se obtiene al representar en ordenadas los autovalores y en abscisas los números de los componentes principales correspondientes a cada autovalor en orden decreciente. La unión de estos puntos muestra un perfil (a una montaña) con una pendiente fuerte hasta llegar a la base.



2.3 Interpretación de las Componentes Principales

Una componente principal es una combinación lineal de todas las variables, pero puede estar muy correlacionada con algunas variables y menos con otras. Entonces es importante considerar el peso que cada variable original tiene dentro de cada componente y las correlaciones existentes entre variables y CP.

2.3 Interpretación de las Componentes Principales

Correlación entre las componentes principales y las variables originales

La correlación entre la j-ésima variable original y la k-ésima componente principal, representa el grado de asociación entre ellas y su valor cuantificará la proporción de la variación total de la variable original j que es explicada por la componente k. Así se tiene:

$$r_{jh} = \frac{w_{hj}\sqrt{\lambda_h}}{\sqrt{Var(X_j)}}$$

Si la variable X_j esta tipificada: $r_{jh} = w_{hj}\sqrt{\lambda_h}$

donde los w_{hj} son los coeficientes de los autovectores o cargas factoriales.

Estas correlaciones también representan la parte de variancia de cada variable que es explicada cada factor. Se cumple que la suma horizontal de los cuadrados de las cargas factoriales de una variable es igual a uno.

2.4 Puntuaciones de las CP o Scores

Las puntuaciones (scores) son los nuevos valores que son obtenidos como la combinación lineal de las variables originales y los coeficientes (autovectores). Para una matriz de datos de orden (nxp), la matriz de puntuaciones también será de orden (nxp). Se obtienen sustituyéndose el conjunto de variables originales en los componentes principales. Estos scores son hallados con el propósito de realizar un diagnóstico de los objetos o como entrada de datos para otros métodos multivariados.

$$Zj = \underline{L}j\ X \qquad j=1, 2, \ldots , p$$

$$Z1 = \underline{L}1X = l11\ X1 + l21\ X2 + \ .\ .\ .\ + lp1\ Xp$$
$$Z2 = \underline{L}2X = l12\ X1 + l22\ X2 + \ .\ .\ .\ + lp2\ Xp$$
$$\vdots$$
$$Zp = \underline{L}pX = l1p\ X1 + l2p\ X2 + \ .\ .\ .\ + lpp\ Xp$$

Matricialmente : $Znxp = Lpxp \times Xpxn$

Ejemplo 1 de Componentes principales

- Se desea hacer el Análisis de Componentes Principales con las notas de los postulantes a cierta universidad con una muestra de 541 postulantes. Trabaje con el archivo **Postulantes.sav**. Para sus cálculos utilice solamente las notas de los alumnos.

Ejemplo 2 de Componentes principales

Con el archivo **PEA** haga el análisis de componentes principales.

Las variables son las siguientes:

Ejemplo 2 de Componentes principales

- X1=PEA Ocupada de 14 años y más que busca Trabajo por Primera Vez.
- X2=PEA Ocupada de 14 años y más que se dedican a actividades No Declaradas.
- X3=PEA Ocupada de 14 años y más que se dedican a actividades de Organizaciones y Órganos Extra Territoriales.
- X4=PEA Ocupada de 14 años y más que se dedican a actividades de Hogares Privados con Servicio Doméstico.
- X8=PEA Ocupada de 14 años y más que se dedican a actividades de Administración Pública y Defensa, Planes de Seguridad Social de Afiliación Obligatoria.
- X18=PEA Ocupada de 14 años y más que se dedican a actividades de Explotación de Minas y Canteras.
- X20=PEA Ocupada de 14 años y más que se dedican a actividades de Agricultura, Ganadería, Caza y Silvicultura.