Análisis Factorial

Fernando A. Crespo R.

November 16, 2023

6.1 Objetivo Análisis de Componentes Principales

- Principal Component Analysis (PCA) es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información.
- Suponga que existe una muestra con n individuos cada uno con p variables $(X_1, X_2, \ldots, X_p)$, es decir, el espacio muestral tiene p dimensiones. PCA permite encontrar un número de factores subyacentes $(z < p)$ que explican aproximadamente lo mismo que las p variables originales. Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores. Cada una de estas z nuevas variables recibe el nombre de componente principal.

6.1 Objetivo Análisis de Componentes Principales

- Es un método que sirve para reducir dimensiones, perteneciente a los métodos de *unsupervised learning*, porque importa extraer la información importante de las variables predictoras, no la variable a predecir.
- Es útil de usar, antes que otras técnicas de análisis de datos.

6.3 Interpretación Intuitiva de Componentes Principales

- Supongamos unos datos espaciales.

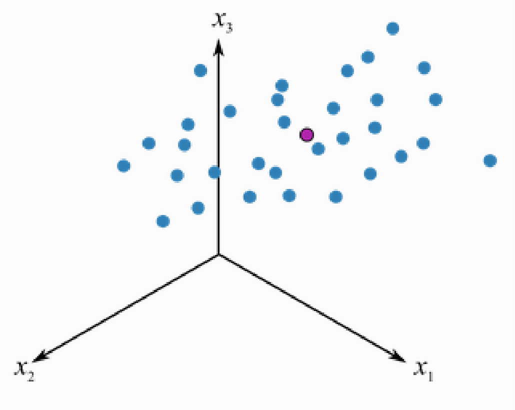

Figure: 6.1.

6.3 Interpretación Intuitiva de Componentes Principales

- Para reducir variables hay que buscar la dirección de máxima información (o varianza).

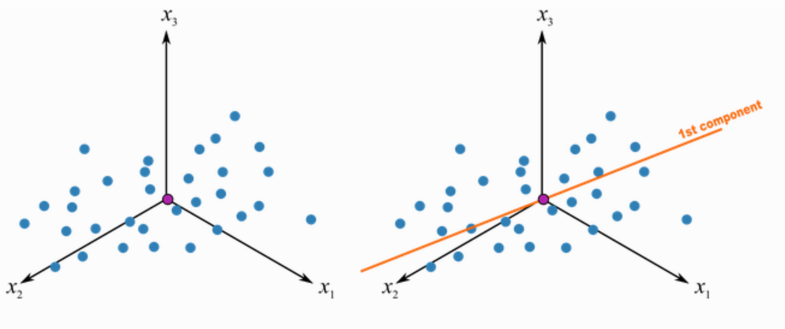

Figure: 6.2.

6.3 Interpretación Intuitiva de Componentes Principales

- Después se ven las cargas sobre esos ejes preferentes.

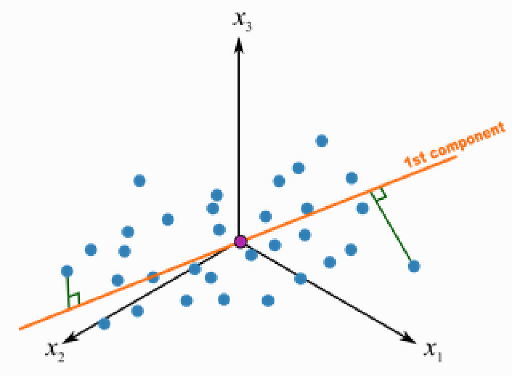

Figure: 6.3.

6.3 Interpretación Intuitiva de Componentes Principales

- Después se busca la siguiente dirección preferente.

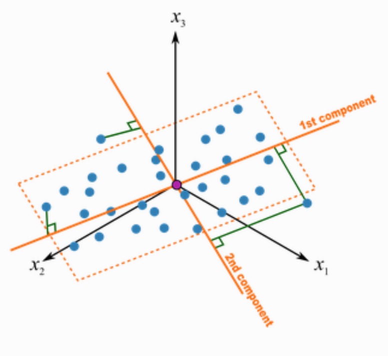

Figure: 6.4.

6.3 Interpretación Intuitiva de Componentes Principales

- Lo que queremos es que la proyección sobre el nuevo eje sea máximo:

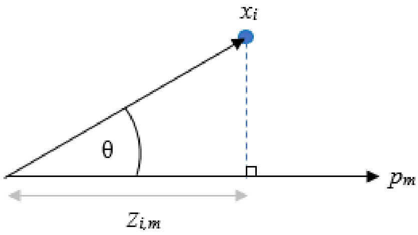

Figure: 6.5.

6.3 Interpretación Intuitiva de Componentes Principales

- La mirada de los datos pueden ser vistos por individuos o por variables.
- Cuando se reducen variables el objetivo es observar los individuos con menos variables.

6.4 Primer Paso

- Normalizar los datos por cada variable, si las unidades de medida son diferentes.

$$y_{ik} = \frac{x_{ik} - \bar{x}}{s_k} \tag{1}$$

6.5 Matriz de correlación

- Antes de despejar la factorización de la matriz, obtenemos los test de Barlett, para ver si las desviaciones estándar son homógeneas, es decir si esférico, y el test KMO para ver si el muestro de la variable es relevante o no para el modelo. A mayor valor, mejor, más relevante la variable para el modelo.
- Estando las variables normalizadas, usamos la matriz de correlación:

$$R = Y^t Y \tag{2}$$

  Con el número de dimensiones igual al número de variables.
- Y a esa matriz le sacamos los valores propios:

$$Ru = \lambda u \tag{3}$$

Cada uno de esos vectores corresponde a los ejes preferenciales de información.

6.5 Matriz de correlación

- Cada λ_i corresponde a una parte de la varianza, y su porcentaje de contribución es:

$$contribución_i = \lambda_i / \sum_i \lambda_i \tag{4}$$