Análisis de Componentes Principales

- Técnica de análisis multivariante que permite resumir un conjunto de variables, supuestamente relacionadas, en unas pocas, no relacionadas entre sí, pero capaces de capturar la mayor variabilidad de la información original.

- El **objetivo** primordial del ACP es resumir la información original, creando nuevas variables (componentes principales), tales que unas pocas (las primeras) sean capaces de reflejar casi toda la información registrada en los datos originales.

Análisis de Componentes Principales

- Las **componentes** principales se obtienen como combinaciones lineales de las variables originales.
- La primera componente principal es la que captura la mayor variabilidad presente en las variables originales.
- La segunda componente, captura la mayor variabilidad posible de los datos, de entre la que no ha sido extraída por la primera componente. Así sucesivamente, en orden decreciente.
- Los **datos**
- La matriz de datos está formada por N individuos o casos sobre los que se observan p variables, generalmente continuas.

ACP

- Cuándo es apropiado realizar un análisis de componentes principales a los datos?
- Cuando las variables que constituyen el conjunto de datos presenten cierto grado de correlación (variación compartida o redundante) y deseemos resumir la información a unas pocas variables (componentes principales) no correlacionadas entre sí.
- Si las variables no están relacionadas (correlaciones no significativas) no tendrá sentido realizar un análisis de componentes principales, puesto que el resultado mostrará tantas componentes como variables originales.
- Ejemplo: Buscar las Componentes Principales a partir de variables macroeconómicas para encontrar indicadores resumidos de la actividad económica de un país. Las puntuaciones obtenidas por distintos países en esas pocas componentes principales permitirá comparar mejor a los países.
- A veces nos interesa efectuar un ACP para aplicar otras técnicas a las componentes principales, tales como análisis de regresión o cluster.
- La matriz objeto de análisis será la de correlaciones entre las variables originales.

ACP

- Un examen de las correlaciones entre las variables originales permite chequear si es apropiado o no el ACP
- Si entre las variables no hay muestra de asociación o correlación no convendrá utilizar ACP
  - Inspección de la matriz de correlaciones. (viendo si hay subgrupos de variables que están significativamente correlacionadas y también observando el determinante de esta matriz. Si el valor del determinante es muy bajo es indicio de alta correlación entre las variables)
  - Un gráfico de las relaciones entre pares de variables también puede ser útil para visualizar el comportamiento de las relaciones.

ACP

- **Resultados del análisis**
- **Matriz de comunalidades**
- Un resultado útil es el coeficiente de correlación múltiple entre cada variable observada (Xi) y todas las componentes principales. Su valor es 1, dado que toda variable Xi puede expresarse de modo exacto como combinación lineal de las componentes.
- Cuando se realiza **Extracción** de un número de componentes inferior al de variables, el coeficiente R2, que tiene cada variable Xi, pero ahora en función de sólo las componentes extraídas se interpreta como la proporción de varianza explicada por las componentes extraídas. Valores altos suponen un buen resultado del análisis, porque indica que la variabilidad presente en cada variable observada está compartida casi en su totalidad por las componentes extraídas. La variable está bien representada por las componentes
- **La comunalidad** de cada variable Xi se obtiene a partir de la matriz, C, de componentes como la suma de los cuadrados de los elementos de cada fila.

ACP

- **Varianza total explicada**
- Presenta la varianza de cada componente (**autovalor**) en orden descendente de importancia. También se expresan en porcentaje con base la variabilidad total (igual a número de variables cuando las variables están tipificadas). La primera componente es la de mayor varianza, la segunda presenta un valor en la varianza inferior o igual a la primera, así sucesivamente.
- Si el ACP da buenos resultados, unas pocas componentes nos permitirán captar un alto % de la variabilidad de los datos.
- Los autovalores muy bajos sugieren que las componentes correspondientes tienen poca relevancia.
- A veces se seleccionan sólo las componentes principales cuyos autovalores son mayores que 1, pero no siempre ésta es la mejor decisión. El ***gráfico de sedimentación*** puede ayudar a decidir el número de componentes principales o factores a seleccionar.

ACP

- **Matriz de componentes**
- La ***matriz de componentes*** muestra las correlaciones entre cada variable y cada una de las componentes extraídas.
- Puede ser muy útil para ***interpretar*** los componentes, así como su representación gráfica en el **Gráfico de componentes**
- Para que pueda interpretarse mejor lo que representan las componentes, es deseable que cada variable esté relacionada altamente sólo con una componente.
- En la práctica, no siempre se da la situación anterior y es necesario efectuar una rotación de la solución para ver si mejora.
- Una de las rotaciones más usada es la VARIMAX
- La suma de los cuadrados de los elementos de cada fila es la comunalidad correspondiente a la variable representada en dicha fila.
- La suma de los cuadrados de los elementos de cada columna es igual al valor propio o autovalor (varianza) de la componente que representa dicha columna.

ACP

- **ACP con R**
- ***Las funciones básicas de R para una ACP:***
- **princomp**() se puede introducir los datos en forma de data.frame, o matriz de cor o como una función**.**
- **predict** () Permite obtener puntuaciones en las componentes
- **varimax**() proporciona directamente la **matriz de componentes rotadas** ($loadings).
- **summary** () Proporciona un resumen de la importancia de cada componente mediante la variabilidad que describe, cada una en términos absolutos y relativos (en porcentajes) y acumulados por orden.
- **plot** () proporciona un gráfico que visualiza la importancia de cada componente en variabilidad extraída.