Análisis de Componentes Principales

Juan Sosa, PhD

I - 2018

Análisis de Componentes Principales (ACP)

Objetivo

- Reducir la dimensionalidad de la data reteniendo tanta de la variabilidad original como sea posible.
- Generar un nuevo conjunto de variables no correlacionadas (componentes principales) que retengan la mayor cantidad de variabilidad de la data original.

Observaciones

- Útil cuando las variables están correlacionadas.
- Técnica de carácter exploratorio/descriptivo, mas NO probabilístico!
- La distribución probabilística de $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]^T$ no es de interés.
- Input para otro método (e.g., análisis de regresión).

Método

Formulación

$$
\begin{aligned}
Y_1 &= a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p = \boldsymbol{a}_1^T \boldsymbol{X} \\
Y_2 &= a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p = \boldsymbol{a}_2^T \boldsymbol{X} \\
&\vdots \\
Y_p &= a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p = \boldsymbol{a}_p^T \boldsymbol{X}
\end{aligned}
$$

Coeficientes (cargas/pesos)

a_{jk}: contribución/aporte de la variable X_k en la componente Y_j.

- Las cargas son escogidas de tal manera que las componentes tengan la **mayor variabilidad posible** (en orden decreciente) y además sean **mutuamente incorrelacionadas**.
- Para que la variabilidad de las componentes no sea arbitrariamente grande se impone la **restricción** $\|\boldsymbol{a}_j\| = 1$.

Computo

Para **maximizar** una función de varias variables con **restricciones** se utiliza el método de los **multiplicadores de Lagrange**:

- $\mathbb{V}\text{ar}\left[Y_j\right] = \lambda_j$ i.e., j-ésimo **valor propio** de $\mathbf{S}$,
- $\boldsymbol{a}_j = \boldsymbol{e}_j$, i.e., j-ésimo **vector propio** de $\mathbf{S}$,

donde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ y $\boldsymbol{e}_j \cdot \boldsymbol{e}_k = 0$.

$$\mathbf{S}\,\boldsymbol{e}_j = \lambda_j\,\boldsymbol{e}_j$$

Observaciones

- La media de los datos no influye en el proceso (centrar la data).
- $\lambda_1 + \lambda_2 + \ldots + \lambda_p = s_1^2 + s_2^2 + \ldots + s_p^2$.
- La interpretación de las componentes Y_j no es directa.
- El ACP no es invariante a la escala de medición.
- Las variables X_j con mayores varianzas tienden a dominar.
- La estructura de las componentes depende de la escala de medición.
- Usar $\mathbf{R}$ cuando hay problemas de conmensurabilidad.

En la práctica

Cuántas componentes son necesarias?

- Graficar λ_j frente a j (*scree diagram*) y excluir aquellas componentes que no provoquen un cambio drástico en la curva.
- Escoger las primeras componentes que acumulen entre 70% y 90% de la variabilidad.
- Excluir la componente Y_j si $\lambda_j < \bar{\lambda}$.
- Usando $\mathbf{R}$, excluir la componente Y_j si $\lambda_j < 0.7$.

Puntajes (*scores*)

$$\mathbb{Y} = \mathbb{X}\mathbf{A}$$

donde $\mathbf{A} = [\boldsymbol{a}_1|\boldsymbol{a}_2|\cdots|\boldsymbol{a}_m]$ y m es el número de componentes seleccionadas.

Biplot (Gabriel, 1981; Gower y Hand, 1986)

- Representación bidimensional gráfica de individuos y variables.
- La longitud del vector representa la varianza.
- El ángulo entre los vectores representa la correlación.