

COMPONENTES PRINCIPALES

Jorge Galbiati R.

El método de Componentes Principales tiene por objeto reducir la dimensionalidad de un problema de múltiples variables, aplicando una sucesión de transformaciones lineales a las variables, de modo que un subconjunto de ellas concentre la mayor parte de la variabilidad contenida en las variables originales.

Supongamos que se tiene un vector aleatorio de dimensión p ,

$$\underline{x} = [x_1 \ x_2 \ \dots \ x_p]' ,$$

que su vector de medias es $\underline{\mu}$ y su matriz de varianzas covarianzas es Σ .

Sea y_1 la variable aleatoria definida por la fórmula

$$y_1 = \underline{\gamma}'_1 (\underline{x} - \underline{\mu}) \quad (1)$$

$\underline{\gamma}_1$ es un vector constante tal que la varianza de y_1 sea máxima, sujeto a la condición de que la norma $\|\underline{\gamma}_1\|$ de $\underline{\gamma}_1$ es 1.

Es decir,

$$\text{Var} (y_1) = \text{máx}_{\underline{\gamma}} \left\{ \text{var} \left[\underline{\gamma}' (\underline{X} - \underline{\mu}) \right] \right\}$$

sujeto a la condición $\|\underline{\gamma}\| = 1$.

Pero

$$\text{var} \left[\underline{\gamma}' (\underline{X} - \underline{\mu}) \right] = \underline{\gamma}' \text{var}(\underline{X} - \underline{\mu}) \underline{\gamma} = \underline{\gamma}' \Sigma \underline{\gamma} \quad \text{y} \quad \|\underline{\gamma}\| = \underline{\gamma}' \underline{\gamma}$$

Luego, equivalentemente, $\underline{\gamma}_1$ se obtiene resolviendo el problema:

Maximizar $\underline{\gamma}' \Sigma \underline{\gamma}$ con respecto de $\underline{\gamma}$ sujeto a la condición $\underline{\gamma}' \underline{\gamma} = 1$

Por multiplicadores de lagrange, esto se logra minimizando la función

$$\Psi(\underline{\gamma}, \lambda) = \underline{\gamma}' \underline{\Sigma} \underline{\gamma} - \lambda(\underline{\gamma}' \underline{\gamma} - 1)$$

Esta es una función polinómica, luego se deriva con respecto de $\underline{\gamma}$ y λ , se igualan a cero las derivadas, y se resuelve es sistema resultante para obtener el mínimo:

$$1.- \quad \frac{\partial \Psi}{\partial \underline{\gamma}} = 2 \underline{\Sigma} \underline{\gamma} - 2\lambda \underline{\gamma} = 0 \quad p \text{ ecuaciones (} p \text{ dimensión del vector aleatorio)}$$

$$2.- \quad \frac{\partial \Psi}{\partial \lambda} = \underline{\gamma}' \underline{\gamma} - 1 = 0 \quad 1 \text{ ecuación}$$

De la ecuación 1, $\underline{\Sigma} \underline{\gamma} = \lambda \underline{\gamma}$, lo que significa que λ es un valor propio de $\underline{\Sigma}$ y su respectivo vector propio es $\underline{\gamma}$, que está estandarizado a norma 1 (estandarizado).

La varianza resulta ser $\underline{\gamma}' \underline{\Sigma} \underline{\gamma} = \lambda \underline{\gamma}' \underline{\gamma} = \lambda$, luego $\underline{\gamma}$ es el vector propio normalizado asociado al mayor valor propio λ_1 , de $\underline{\Sigma}$, y que denominaremos $\underline{\gamma}_1$.

Recordar que $\underline{\Sigma}$ es semidefinida positiva, luego todos sus valores propios son mayores o iguales a cero.

Además, por ser $\underline{\Sigma}$ simétrica, sus vectores propios son ortogonales entre si.

DEFINICION:

El vector aleatorio $y_1 = \underline{\gamma}'_1 (\underline{x} - \underline{\mu})$ de la transformación (1) se denomina *Primera Componente Principal* de \underline{x} . La dirección definida por el vector $\underline{\gamma}_1$ se denomina *Primer Eje Principal*.

La segunda componente principal es la variable aleatoria $y_2 = \underline{\gamma}'_2 (\underline{x} - \underline{\mu})$, en que $\underline{\gamma}_2$ es el vector propio asociado al segundo valor propio λ_2 , etc. hasta la p-ésima componente principal.

PROPIEDADES:

$$1.- \quad E(y_i) = 0$$

$$2.- \quad Var(y_i) = \lambda_i \quad i \text{-ésimo valor propio de } \underline{\Sigma}$$

$$3.- \quad Cov(y_i, y_j) = 0 \quad \text{para } i \neq j$$

$$4.- \quad Var(y_1) \geq Var(y_2) \geq \dots \geq Var(y_p) \geq 0$$

$$5.- \quad \sum_{i=1}^p var(y_i) = \text{Traza}(\underline{\Sigma})$$

$$6.- \quad \prod_{i=1}^p var(y_i) = \det(\underline{\Sigma})$$

Las transformaciones y_i se pueden escribir todas juntas en forma condensada

$$\underline{y} = \Gamma'(\underline{x} - \underline{\mu})$$

en que \underline{y} es el vector de componentes principales $\underline{y} = [y_1, y_2, \dots, y_p]'$

y Γ la matriz ortogonal $p \times p$ tal que sus columnas son los vectores propios de Σ , en orden descendente de sus valores propios.

Observar que Γ es ortogonal, luego $\Gamma\Gamma' = \Gamma'\Gamma = I$. Esto significa que la transformación de componentes principales, $\underline{y} = \Gamma'(\underline{x} - \underline{\mu})$, consiste en un centrado del vector de observaciones, al restar $\underline{\mu}$, para luego efectuar una rotación (la multiplicación de un vector por una matriz ortogonal corresponde a una rotación del vector en el espacio). Esta rotación busca "ver" el vector desde el punto en que aparece con mayor variabilidad, por lo tanto, más informativo.

También se observa que Γ y Γ' son matrices mutuamente inversas.

Entonces la transformación inversa, que convierte las componentes principales en las observaciones originales, es :

$$\underline{x} = \Gamma\underline{y} + \underline{\mu}$$

pues que Γ y Γ' son matrices mutuamente inversas.

Los registros γ_{ij} de Γ se denominan cargas (loadings)

Las propiedades 1, 2 y 3 dadas anteriormente, se pueden expresar en términos matriciales:

1. $E(\underline{y}) = 0$
2. y 3. $\text{Var}(\underline{y}) = \Lambda$, con $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_p\}$

También se puede agregar que la matriz de covarianzas entre \underline{x} e \underline{y} es

$$\text{Cov}(\underline{x}, \underline{y}') = \Gamma\Lambda$$

Entonces la covarianzas entre la variable x_i (la coordenada i-ésima de \underline{x}) y la j-ésima componente principal está dada por $\gamma_{ij}\lambda_j$

La matriz de correlaciones entre \underline{x} e \underline{y} está dada por $D^{-\frac{1}{2}}\Gamma\Lambda^{\frac{1}{2}}$ en que

$$\Lambda^{\frac{1}{2}} = \text{diag}\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_p}\} \quad \text{y} \quad D^{-\frac{1}{2}} = \text{diag}\left\{\frac{1}{\sqrt{\sigma_{11}}}, \frac{1}{\sqrt{\sigma_{22}}}, \dots, \frac{1}{\sqrt{\sigma_{pp}}}\right\}$$

en que σ_{ii} es la varianza de la variable original x_i . Recordar que λ_j es la varianza de la j -ésima componente principal y_j .

Desarrollando la expresión matricial para la matriz de correlaciones obtenemos las correlaciones individuales:

$$\text{corr}(x_i, y_j) = \gamma_{ij} \sqrt{\frac{\lambda_j}{\sigma_{ii}}}$$

COMPONENTES PRINCIPALES MUESTRALES

Si se tiene una matriz de datos $X_{n \times p}$, las componentes principales se obtienen a partir de la matriz de varianzas - covarianzas muestral $S = X'(I - \frac{1}{n}\mathbf{1}\mathbf{1}')X$, en que I es la matriz identidad p -dimensional $\text{diag}\{1, 1, \dots, 1\}$ y $\mathbf{1}$ es un vector p -dimensional en que todos sus elementos son unos.

O bien se pueden obtener las componentes principales a partir de la matriz de correlaciones muestral $R = D^{-1}SD^{-1}$ en que

$$D^{-1} = \text{diag}\left\{\frac{1}{\sqrt{s_{11}}}, \frac{1}{\sqrt{s_{22}}}, \dots, \frac{1}{\sqrt{s_{pp}}}\right\}$$

siendo los s_{ii} las varianzas muestrales. La matriz $H = (I - \frac{1}{n}\mathbf{1}\mathbf{1}')$ se denomina Matriz de Centrado, porque lo que hace es quitarle a cada elemento de la matriz de datos X , el promedio muestral de la variable respectiva (el promedio de la columna).

Las componentes principales no son invariantes con respecto de cambios de escala de las variables. Por lo tanto si se usa S o se usa R , los resultados obtenidos serán diferentes.

¿Cuándo se usa una u otra matriz? Si la escala de medida de todas las variables es la misma, basta con S . En cambio, si las escalas de medida son muy diferentes dominarán las variables con magnitudes mayores. Por lo tanto, en estos casos, conviene usar la matriz de correlaciones R .

ALGORITMO DE CALCULO

De forma análoga al caso de las Componentes Principales poblacionales, visto arriba, dada una matriz de datos, para obtener las componentes principales (muestrales) se deben calcular los valores y vectores propios de S o de R . Los valores propios son las varianzas de las Componentes Principales. Los vectores propios son los ejes principales. Las coordenadas del i -ésimo vector propio son las cargas o coeficientes de la i -ésima componente principal.

Los valores de las componentes principales (también llamados Scores) se obtienen multiplicando las variables centradas por los vectores propios:

La j -ésima componente principal es el vector n -dimensional dado por

$$y_j = HX\underline{\gamma}_j$$

en que $\underline{\gamma}_j$ es el j -ésimo vector propio, H es la matriz de centrado

$$H = (I - \frac{1}{n}\mathbf{1}\mathbf{1}')$$

Expícitamente, la observación i -ésima de la j -ésima variable se expresa como función lineal de las p variables originales de la i -ésima observación:

$$y_{ij} = \gamma_{i1}(x_{i1} - \bar{x}_1) + \gamma_{i2}(x_{i2} - \bar{x}_2) + \dots + \gamma_{ip}(x_{ip} - \bar{x}_p)$$

La covarianza muestral entre x_i y la componente principal y_j está dada por el producto

$$cov(x_i, y_j) = \gamma_{ij}\lambda_j,$$

el producto de la coordenada i -ésima de la j -ésima componente principal γ_{ij} por el j -ésimo valor propio λ_j

si se usa S , ó $\gamma_{ij}\sqrt{\lambda_j s_{ii}}$ en el caso en que se usa R .

La correlación muestral entre x_i e y_j está dada por

$$corr(x_i, y_j) = \gamma_{ij}\sqrt{\frac{\lambda_j}{s_{ii}}}$$

si se usa S , ó $\gamma_{ij}\sqrt{\lambda_j}$ en el caso en que se usa R .

Reducción de dimensionalidad

Las componentes principales tienen un orden descendente de importancia, debido a que la varianza que contienen va también en orden descendente, y las variables que tienen poca varianza, tienen poca información acerca de los sujetos observados.

En las aplicaciones, se suele conservar las primeras componentes y descartar las últimas, con el objeto de reducir la dimensionalidad del problema. Las que se descartan constituyen información que se pierde.

Según la forma que tienen las cargas en las fórmulas de transformación de las observaciones en las componentes, se les puede dar interpretaciones a las componentes principales. Por ejemplo, si las cargas son todas positivas, es un promedio ponderado entre las variables cuyas cargas sean significativamente grandes. Si hay positivas y negativas, es un contraste o comparación (entre las variables que tienen cargas positivas y las que tienen cargas negativas, con magnitudes significativas).

Para decidir cuántas conservar se debe llegar a un compromiso entre el grado de simplificación del problema y la cantidad de información perdida.

Existen algunos criterios para decidir cuántas componentes principales conservar, y cuántas descartar:

1) Observar el gráfico de barras que muestra las magnitudes de las varianzas de las componentes, en orden descendente (*scree plot*). Este puede sugerir si hay un punto que separe las "grandes" de las "pequeñas". Este se denomina criterio de Cattell.

2) Incluir las componentes que expliquen hasta un 90% de la varianza acumulada. Este criterio tiende a incluir muchas componentes.

3) Excluir aquellas componentes cuyas varianzas sean menos que el promedio. Este es el llamado criterio de Kayser, y tiende a incluir muy pocas variables.

En cada caso se verá cuál criterio será más apropiado, o si se aplicará un criterio de compromiso entre ellos.

EJEMPLO

Supongamos que se tiene una matriz de datos de n observaciones de dimensión 3, es decir, tres variables.

$$X = \begin{bmatrix} 24 & 8 & 7 \\ 14 & 8 & 5 \\ 32 & 9 & 10 \\ 28 & 10 & 7 \\ 23 & 12 & 9 \\ \cdot & \cdot & \cdot \\ 21 & 11 & 8 \end{bmatrix}$$

Se calculan el vector de medias y la matriz de varianzas-covarianzas que supondremos dan los siguientes resultados:

$$\bar{x} = \begin{bmatrix} 22,1 \\ 11,4 \\ 9,3 \end{bmatrix} \quad S = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{bmatrix}$$

Los valores propios son $\lambda_1 = 4$, $\lambda_2 = 3$, $\lambda_3 = 1$ y sus respectivos vectores propios son

$$v_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad v_3 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

Los valores que aparecen como factores, en términos decimales (aproximado a tres decimales) son

$$\frac{1}{\sqrt{3}} = 0,577 \quad \frac{1}{\sqrt{2}} = 0,707 \quad \frac{1}{\sqrt{6}} = 0,408$$

Los ejes principales son v_1 , v_2 y v_3 , en ese orden. Los porcentajes de varianza respectivos de las componentes principales son 50%, 37,5% y 12,5%. Las varianzas acumuladas son 50% para la primera, 87,5% para las primeras dos y 100% para las tres.

Se puede observar que la suma de las varianzas de las tres componentes principales es igual a la suma de las varianzas de las observaciones originales.

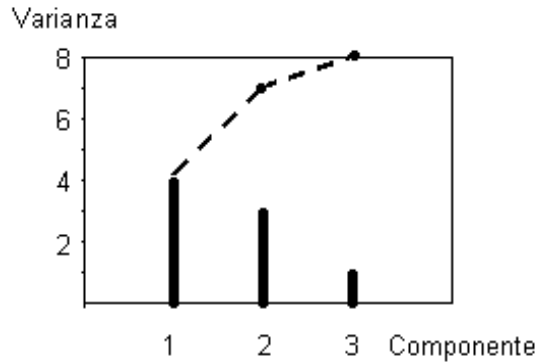


Figura 1: Scree plot.

Las transformaciones de componentes principales $y_i = \gamma'_i(\underline{x} - \underline{\mu})$ expresadas de manera explícita son:

$$y_1 = 0,577x_1 + 0,577x_2 + 0,577x_3 - 24,696$$

$$y_2 = 0,707x_1 - 0,707x_3 - 9,050$$

$$y_3 = 0,408x_1 - 0,816x_2 + 0,408x_3 - 3,509$$

El término constante $-24,696$ corresponde a

$$\gamma'_1 \bar{\underline{x}} = 0,577 \times 22,1 + 0,577 \times 11,4 + 0,577 \times 9,3$$

Los otros se obtienen en forma análoga.

La primera componente se puede interpretar como un promedio entre las tres variables x_1 , x_2 y x_3 . La segunda como una comparación entre la primera variable, x_1 , y la tercera, x_3 . La tercera componente principal como una comparación entre la segunda variable, x_2 , y el promedio de las otras dos.

El siguiente gráfico (screeplot) exhibe las tres varianzas, mostrando que la primera componente contiene más del doble de la varianza de cada una de las restantes. También se muestran las varianzas acumuladas :

Según los criterios presentados para la eliminación de componentes, claramente se puede eliminar la última sin gran pérdida de información, quedándonos con las dos primeras.

Los escores de las 6 observaciones que se observan en la matriz de datos, correspondientes a las dos primeras componentes principales, se calculan utilizando las fórmulas de cálculo de las componentes. Para la primera observación son

$$y_{11} = 0,577 \times 24 + 0,577 \times 8 + 0,577 \times 7 - 24,696 = -2,19$$

$$y_{12} = 0,707 \times 24 - 0,707 \times 7 - 9,050 = 2,97$$

Calculando los dos escores para las demás observaciones, se obtienen los siguientes resultados, expresados como matriz de datos:

$$Y = \begin{bmatrix} -2,19 & 2,97 \\ -9,11 & -2,68 \\ 4,73 & 6,50 \\ 1,27 & 5,80 \\ 0,69 & 0,85 \\ \cdot & \cdot \\ -1,62 & 0,14 \end{bmatrix}$$

Estos pares se pueden graficar en un plano formado por las dos primeras componentes principales (*Primer Plano Principal*), que muestra las proyecciones de los puntos sobre este plano, que es aquel en que se observa la mayor dispersión. El gráfico se denomina *Biplot*.

Las proyecciones de los ejes originales x_1 , x_2 y x_3 sobre el primer plano principal se obtienen multiplicando los vectores elementales

$$\underline{e}_1 = [1, 0, 0]' \quad \underline{e}_2 = [0, 1, 0]' \quad \underline{e}_3 = [0, 0, 1]' \quad \text{por } \gamma_1$$

y por γ_2 , obteniéndose así los extremos de los vectores elementales.

Los resultados son $[0,577; 0,707]$, $[0,577; 0]$ y $[0,577; -0,707]$ para cada uno de las proyecciones de los ejes originales, los que se muestran en el biplot.

Se observa una mayor dispersión de las observaciones que vemos en la matriz de datos, en el sentido del eje de la primera componente. Las observaciones aparecen numeradas, según el orden en que se ven en la matriz de datos. Los tres vectores rotulados \underline{x}_1 , \underline{x}_2 y \underline{x}_3 son las proyecciones de los vectores unitarios que siguen la dirección de los tres ejes correspondientes a las variables originales, sobre el primer plano principal.

Cálculo de las correlaciones de las variables originales x con las componentes principales y , utilizando la fórmula $corr(x_i, y_j) = \gamma_{ij} \sqrt{\frac{\lambda_j}{s_{ii}}}$:

Recordar que la matriz de transformación, cuyas columnas son los vectores propios normalizados de la matriz de varianzas-covarianzas S , y cuyos elementos son las *cargas*, es

$$\Gamma = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

Los valores propios λ son $\lambda_1 = 4$, $\lambda_2 = 3$, y $\lambda_3 = 1$.

Las varianzas de las variables originales son $\sigma_{11} = 3$, $\sigma_{22} = 2$, y $\sigma_{33} = 3$.

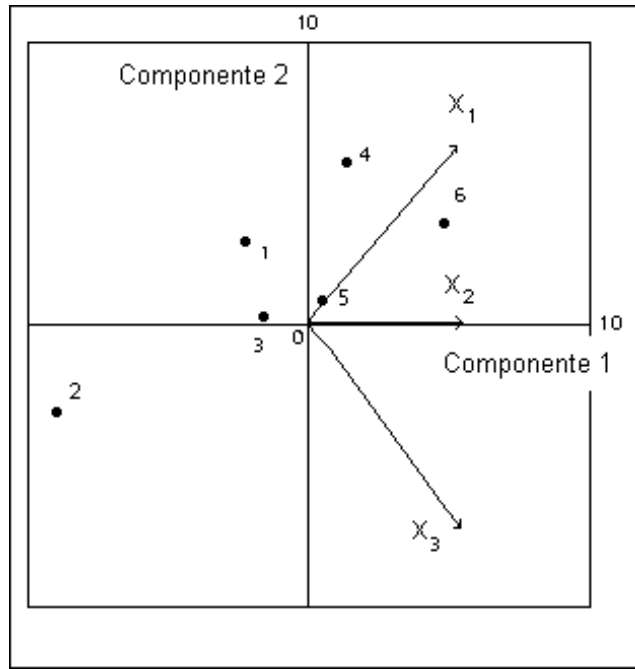


Figura 2: Biplot.

Con estos valores obtenemos la matriz de correlaciones siguiente, en que el elemento (i, j) es la correlación $corr(x_i, y_j)$:

$$Corr(\underline{x}, \underline{y}) = \begin{bmatrix} \frac{2}{3} & \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} \\ \frac{\sqrt{2}}{\sqrt{3}} & 0 & -\frac{1}{\sqrt{3}} \\ \frac{2}{3} & -\frac{\sqrt{3}}{2} & \frac{1}{3\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0,667 & 0,707 & 0,236 \\ 0,816 & 0 & -0,577 \\ 0,667 & -0,866 & 0,236 \end{bmatrix}$$

Se puede observar que la primera variable x_1 está correlacionada positivamente con las dos primeras componentes principales, lo que se refleja en el gráfico Biplot, en que aparece aproximadamente diagonal en el primer cuadrante. La segunda variable está correlacionada con la primera componente, y tiene cero correlación con la segunda. Por eso en el gráfico aparece paralela al eje de la primera componente. La tercera variable parece correlacionada positivamente con la primera componente, y negativamente con la segunda. De ahí que aparezca casi diagonal, pero en el cuadrante 4, en que la segunda componente es negativa.