Análisis en Componentes Principales

Dr. Oldemar Rodríguez Rojas

29 de mayo de 2008

Contents

Análisis en Componentes Principales (ACP)

El Análisis de Componentes Principales (ACP) es una técnica proveniente del análisis exploratorio de datos cuyo objetivo es la síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante una tabla de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible. El ACP es uno de los métodos más utilizados en Minería de Datos en países como Francia. Fue primeramente introducido por Pearson en 1901 y desarrollado independientemente en 1933 por Hotelling y la primera implementación computacional se díio en los años 60. Fue aplicado para analizar encuestas de opinión pública por Jean Pages. Como ya se mencionó el objetivo es construir un pequeño número de nuevas variables (componentes) en las cuales se concentre la mayor cantidad posible de información, como se ilustra en la Figura 1.

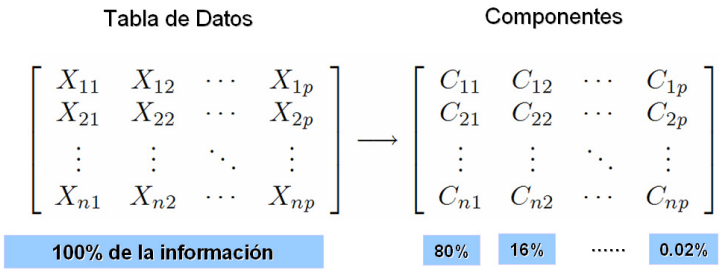

FIGURE 1. Transformación de las variables originales en componentes.

Estas nuevos componentes principales o factores son calculados como una combinacin lineal de las variables originales, y además serán linealmente independientes. Un aspecto clave en ACP es la interpretación, ya que ésta no viene dada a priori, sino que será deducida tras observar la relación de los componentes principales con las variables originales, para esto hay que estudiar tanto el signo como la magnitud de las correlaciones, como vermos en detalle más adelante. Esto no siempre es fácil, y será de vital importancia el conocimiento que el experto tenga sobre la materia de investigación.

Los n individuos de una tabla de datos se pueden ver como una nube de puntos en $\mathbb{R}^p$, como se ilustra en la Figura 2-a, con su centro de gravedad localizado en el origen, y lo que se busca es un subespacio $q-$dimensional L de $\mathbb{R}^p$, usualmente un plano (ver Figura 2-b), tal que la proyección ortogonal de los n puntos sobre L (ver Figura 2-c) tienen varianza máxima, lo cual permitirá el estudio de relaciones, clases, etc. entre los individuos (filas) de

la tabla de datos.

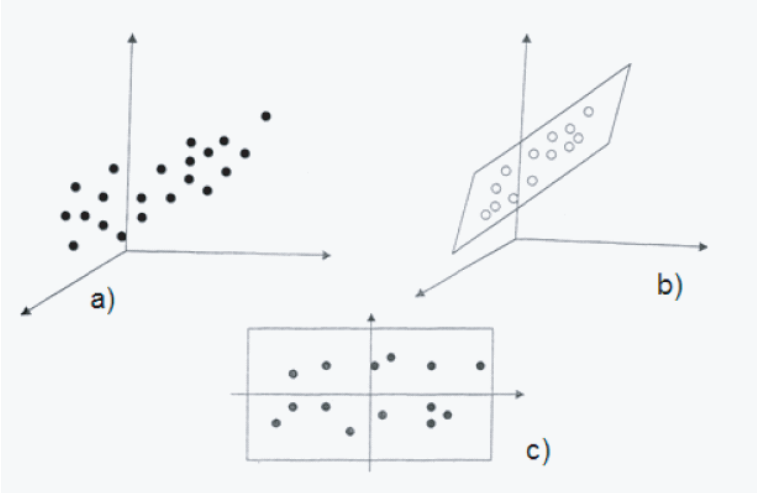

FIGURE 2. Proyección de los individuos en el plano de varianza máxima

1. Los datos

Se parte de una tabla de datos:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix} \hookleftarrow \text{individuo } i$$

que se puede transformar en la siguiente matriz de distancias:

$$D = \begin{pmatrix} d_{11} & \cdots & d_{1j} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ d_{i1} & \cdots & d_{ij} & \cdots & x_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nj} & \cdots & d_{nn} \end{pmatrix}$$

2. El problema

- Se trata de sintetizar los datos contenidos en una tabla de datos X en un conjunto más pequeño de nuevas variables $C^1, C^2, \ldots$ llamadas componentes principales, manteniendo la información escencial de X.
- Así, en la etapa 1 del algoritmo se encuentra una **variable sintética** C^1, la primera componente principal, la cual es combinación lineal de las variables originales X^j, es decir:

$$C^1 = a_{11}X^1 + \cdots + a_{1j}X^j + \cdots + a_{1m}X^m,$$

  donde X^j es la columna j de X. Esto significa que el valor de C^1 para el individuo $i-$ésimo está dado por:

$$C_i^1 = a_{11}x_{i1} + \cdots + a_{1j}x_{ij} + \cdots + a_{1m}x_{im},$$

- Generalmente esta primer componente principal, C^1, no es suficiente para condensar la información contenida en X, por lo que se construye una segunda componente principal C^2, luego una tercera C^3 y así sucesivamente.
- En general en la etapa k, se construye la componente principal $k-$ésima dada por:

$$C^k = a_{k1}X^1 + \cdots + a_{kj}X^j + \cdots + a_{km}X^m.$$

- Matricialmente se tiene que:

$$C^k = Xa^k,$$

  donde:

$$a^k = \begin{pmatrix} a_{k1} \\ \vdots \\ a_{kj} \\ \vdots \\ a_{km} \end{pmatrix}.$$

- a^k se llama el $k-$**ésimo factor**.
- Los factores a_{kj} constituyen un sistema de pesos para las variables, los cuales indican cuanto aporta cada variables a la construcción de la componente.

- Algunos factores a_{kj} serán negativos y otros serán positivos. El valor de cada peso por si solo no es importante, sino la relación con respecto a los otros pesos. Para evitar un problema de escalas se impone la siguiente restricción:

$$\sum_{j=1}^{m}(a_{kj})^2 = 1.$$

3. Cálculo de los factores y de las componentes principales

Como en regresión, el ACP puede ser presentado tanto en el espacio de las variables como en el espacio de los individuos.

3.1. En el espacio de los individuos

- Se supondrá que las variables están centradas y reducidas.
- $V = \frac{1}{n}X^tX$ es la matriz de varianzas–covarianzas. Como las variables están centradas y reducidas entonces $V = R$, la matriz de correlaciones, pues:

$$v_{ij} = cov(X^i, X^j) = \frac{cov(X^i, X^j)}{\sigma_{X^i}\sigma_{X^j}} = R(X^i, X^j).$$

- Por lo tanto el espacio de las filas de X en $\mathbb{R}^m$ es el espacio de individuos cuyo origen será el centro de la nube de puntos.
- El objetivo del ACP es describir de manera sintética la nube de individuos.

Teorema 1 En la etapa 1 de un ACP se calcula el eje D_1 que pasa por el origen para el cual la dispersión de la nube de puntos sea máxima, este eje D_1 pasa entonces lo más cerca posible de la nube de puntos, es decir, el promedio de las distancias al cuadrado de los n puntos de la nube y el eje D_1 es minimal (ver Figura ??).

Sea a^1 es vector director normado (norma 1) del eje (recta) D_1 **entonces**: a_1 es el vector propio asociado al valor propio más grande de la matriz de V de varianzas–covarianzas.

Teorema 2 En la etapa 2 de un ACP se calcula el eje D_2 que pasa por el origen para el cual la dispersión de la nube de puntos sea máxima, este eje D_2 pasa entonces lo más cerca posible de la nube de puntos, es decir,

el promedio de las distancias al cuadrado de los n puntos de la nube y el eje D_2 es minimal.

Sea a^2 es vector director normado (norma 1) del eje (recta) D_2 el cual será ortogonal al vector a^1 construido en la etapa 1, **entonces**: Se tiene el siguiente problema de optimización:

$$\begin{array}{ll} \text{máx} & \frac{1}{n}\left(a^2\right)^t X^t X a^2 \\ \text{sujeto} & \begin{cases} \left(a^2\right)^t a^2 = 1 \\ \left(a^2\right)^t a^2 = 0 \end{cases} \end{array}$$

cuya solución es el vector propio asociado al segundo valor propio más grande de la matriz de V de varianzas–covarianzas.

Teorema 3 En la etapa k de un ACP se calcula el eje D_k que pasa por el origen para el cual la dispersión de la nube de puntos sea máxima, este eje D_k pasa entonces lo más cerca posible de la nube de puntos, es decir, el promedio de las distancias al cuadrado de los n puntos de la nube y el eje D_k es minimal.

Sea a^k es vector director normado (norma 1) del eje (recta) D_k el cual será ortogonal al vector $a^r \ \forall\, r < k$ construidos en las etapas $1, 2, \ldots, k-1$ **entonces**: Se tiene el siguiente problema de optimización:

$$\begin{array}{ll} \text{máx} & \frac{1}{n}\left(a^k\right)^t X^t X a^k \\ \text{sujeto} & \begin{cases} \left(a^k\right)^t a^k = 1 \\ \left(a^k\right)^t a^k = 0 \text{ para } r = 1, 2, \ldots, k-1 \end{cases} \end{array}$$

cuya solución es el vector propio asociado al $k-$ésimo valor propio más grande de la matriz de V de varianzas–covarianzas.

4. En el espacio de las variables

Teorema 4 En la etapa 1 de un ACP se calcula una variable sintética (eje) C^1 que resuma lo mejor posible las variables originales, es decir, de tal manera que:

$$\sum_{j=1}^{m} R^2(C^1, X^j) \text{ sea máxima.}$$

Entonces: C^1 es el vector propio de $\frac{1}{n}XX^t$ asociado al valor propio más grande.

Teorema 5 En la etapa k de un ACP se calcula una variable sintética (eje) C^k que resuma lo mejor posible las variables originales y que no esté correlacionada las primeras $k-1$ componentes principales (variables sintéticas) ya calculadas, es decir, de tal manera que:

$$\begin{aligned} &\text{máx} \quad \sum_{j=1}^{m} R^2(C^k, X^j) \\ &\text{sujeto} \quad R^2(C^k, C^r) = 0 \text{ para } r = 1, 2, \ldots, k-1 \end{aligned}$$

Entonces: C^k es el vector propio de $\frac{1}{n}XX^t$ asociado al $k-$ésimo valor propio más grande.

5. Equivalencia de los dos análisis – Relaciones de dualidad

- Espacio de los individuos $\mapsto \frac{1}{n}X^tX$ que es tamaño $m \times m$.
- Espacio de las variables $\mapsto \frac{1}{n}XX^t$ que es tamaño $n \times n$.

Usualemente el número de variables es menor que el número de individuos, por supondremos en adelante sin pérdidad de generalidad que $m < n$.

Teorema 6 [Relaciones de Dualidad]

1. Si v_k es el $k-$ésimo vector propio de norma 1 asociado a λ_k de la matriz $\frac{1}{n}XX^t$ entonces:

   $$u_k = \frac{X^t v_k}{\sqrt{n\lambda_k}},$$

   es el $k-$ésimo vector propio de norma 1 asociado a λ_k de la matriz $\frac{1}{n}X^tX$.

2. Si u_k es el $k-$ésimo vector propio de norma 1 asociado a λ_k de la matriz $\frac{1}{n}X^tX$ entonces:

   $$v_k = \frac{Xu_k}{\sqrt{n\lambda_k}},$$

   es el $k-$ésimo vector propio de norma 1 asociado a λ_k de la matriz $\frac{1}{n}XX^t$.

6. Varianza explicada por cada eje

Teorema 7

- $\frac{1}{n}X^tX$ y $\frac{1}{n}XX^t$ tienen los mismos valores propios, $\beta_1, \beta_2, \ldots, \beta_m$.
- Además el rango de ambas matrices es $n - m$ y los últimos $n - m$ valores propios de $\frac{1}{n}XX^t$ son nulos.

Teorema 8

$$\sum_{k=1}^{m} \beta_k = m.$$

El ACP tiene m etapas, en cada etapa se construye un resumen de la tabla X, menos interesante que el construido en la etapa anterior.

- ¿Cómo medir la calidad de la etapa k?
- En la etapa k, el criterio del ACP es maximizar:

  $$\frac{1}{n}\sum_{i=1}^{n}(C_i^k)^2,$$

  como:

  $$\frac{1}{n}\sum_{i=1}^{n}(C_i^k)^2 = \frac{1}{n}(a^k)^tX^tXa^k = (a^k)^t\beta_k a^k = \beta_k.$$

- Entonces β_k es la varianza explicada por el eje $k-$ésimo, es decir por C^k.
- Como:

  $$\sum_{k=1}^{m} \beta_k = m,$$

  se tiene que:

  $$\frac{\beta_k}{m} = \% \text{ de la varianza explicada por el eje } C^k = \% \text{ de INERCIA.}$$

- Por ejemplo, la **inercia** explicada por el plano principal, ejes 1 y 2 es:

  $$\frac{\beta_1 + \beta_2}{m}.$$

7. Gráficos y su interpretación

7.1. Representación de los individuos

Recordemos que para calcular las coordenadas de un individuos se tiene que (La matriz X se supone centrada y reducida):

- $C^s = Xa^s$ donde a^s es el vector propio de $R = \frac{1}{n}X^tX$ asociado a λ_s.
- De donde:

$$C_i^s = a_1^s X_{i1} + \cdots + a_j^s X_{ij} + \cdots + a_m^s X_{im},$$

es decir:

$$C_i^s = \sum_{j=1}^{m} X_{ij} a_j^s$$

Análogamente:

- $C^r = Xa^r$ donde a^r es el vector propio de $R = \frac{1}{n}X^tX$ asociado a λ_r.
- De donde:

$$C_i^r = a_1^r X_{i1} + \cdots + a_j^r X_{ij} + \cdots + a_m^r X_{im},$$

es decir:

$$C_i^r = \sum_{j=1}^{m} X_{ij} a_j^r$$

Gráficamente se ilustra como sigue:

- Así, dos individuos i y j cuyas proyecciones son cercanas son "semejantes"en la nube de puntos.
- Para proyectar un individuo en suplementario $s = (s_1, \ldots, s_m)$ simplemente se centra y reduce como si fuera la última fila de X, como sigue:

$$\tilde{s} = \left(\frac{s_1 - \bar{X}^1}{\sigma_1}, \ldots, \frac{s_m - \bar{X}^m}{\sigma_m} \right),$$

donde $\bar{X}^j$ es la media de la columna $j-$ésima de la matriz X. Entonces las coordenadas se calculan como sigue:

$$C_i^s = \sum_{j=1}^{m} \tilde{s}_j a_j^s$$

7.2. Calidad de la representación de un individuo

- En el espacio de los individuos se tienen 2 bases ortonormales:

  1. La base original, en la cual las coordenadas del individuo i son:

  $$i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{im}).$$

  2. La base construida por los m factores, en la cual las coordenadas del individuo i son:

  $$i = (C_i^1, \ldots, C_i^k, \ldots, C_i^m),$$

  entonces la distancia del punto al origen se puede medir con ambas representaciones, lo que implica que:

  $$\sum_{j=1}^{m} (X_{ij})^2 = \sum_{k=1}^{m} (C_i^k)^2.$$

- De modo que el individuo i tiene una buena representación en el eje r si $(C_i^r)^2$ tiene un valor importante respecto a la suma $\sum_{j=1}^{n} (X_{ij})^2$.

- Por lo que la calidad de la representación del individuo i sobre el eje r está dada por:

  $$\frac{(C_i^r)^2}{\sum_{j=1}^{m} (X_{ij})^2} = \% \text{ del individuo } i \text{ representado en el eje } r.$$

- Lo anterior es útil para qué tan bien está representado un individuo en un eje o plano.

7.3. Las contribuciones de los individuos a la varianza total

- La varianza total en la etapa r es igual a:

  $$\frac{1}{n} \sum_{i=1}^{n} (C_i^r)^2 = \beta_r.$$

- La parte de esta varianza explicada por el individuo i es:

  $$\frac{1}{n} (C_i^r)^2$$

- Entonces, la contribución del individuo i a la varianza total del eje r está dada por:

$$\frac{(C_i^r)^2}{n\beta_r} = \% \text{ de contribución del individuo } i \text{ a la formación del eje } r.$$

- Lo anterior es útil para intepretar los ejes.

7.4. Representación de las variables

- La coordenada de la variable X^j sobre el eje r está dada por:

$$R(X^j, C^r),$$

que es el coeficiente de correlación entre la variable $j-$ésima y la componente principal $r-$ésima.

- Entonces las coordenadas de X^j sobre la base de componentes principales son:

$$(R(X^j, C^1), \ldots, R(X^j, C^s), \ldots, R(X^j, C^m)),$$

esto implica que:

$$\sum_{k=1}^{m} R^2(X^j, C^k) = 1$$

- Por lo que si se usan solamente 2 componentes C^r y C^s se tiene que:

$$R^2(X^j, C^s) + R^2(X^j, C^r) \leqslant 1.$$

- Por esta razón las variables pueden ser representadas en un círculo de radio 1 como se ilustra a continuación:

Teorema 9 [Cálculo de las correlaciones]

$$\begin{pmatrix} R(X^1, C^r) \\ \vdots \\ R(X^j, C^r) \\ \vdots \\ R(X^m, C^r) \end{pmatrix} = \sqrt{\beta_r} \cdot a^r = \begin{pmatrix} \sqrt{\beta_r} a_1^r \\ \vdots \\ \sqrt{\beta_r} a_j^r \\ \vdots \\ \sqrt{\beta_r} a_m^r \end{pmatrix},$$

donde a^r es el $r-$ésimo vector propio de $R = \frac{1}{n} X^t X$ asociado a λ_r.

- Por dualidad, en el espacio de las variables, para calcular las coordenadas (correlaciones) se podría diagonalizar la matriz $H = \frac{1}{n} X X^t$ (que es tamaño $n \times n$) y proceder a calcular dichas coordenadas de manera completamente análoga al caso de los individuos.

  Es decir, suponiendo que la matriz X está centrada y reducida, y si denotamos por $Z = X^t$ entonces:

  $R^s = Z a^s$ donde a^s es el vector propio de $H = \frac{1}{n} X X^t$ asociado a λ_s.

  De donde:

$$R_i^s = a_1^s Z_{i1} + \cdots + a_j^s Z_{ij} + \cdots + a_n^s Z_{in},$$

  es decir:

$$R_i^s = \sum_{j=1}^{n} Z_{ij} a_j^s$$

- Para proyectar una variable suplememtaria:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

  primero se centra y se reduce respecto a sí misma como sigue:

$$y^c = \begin{pmatrix} \frac{y_1 - \bar{y}}{\sigma_y} \\ \frac{y_2 - \bar{y}}{\sigma_y} \\ \vdots \\ \frac{y_n - \bar{y}}{\sigma_y} \end{pmatrix}$$

  y luego se calculan las correlaciones de y^c con las componentes principales, de manera análoga a proyectar una columna de X.

- **INTERPRETACIÓN**

  - Si la proyección de X^j está cercana al borde del círculo (la suma de las correlaciones al cuadrado está cerca de 1), significa que está bien representada en ese plano, pues tendría fuerte correlación con las 2 componentes (o con alguna de ellas) y por la tanto la correlación con las demás componentes es débil.

- Si dos variables X^j y $X^{j'}$ están cercanas al borde del círculo, entonces el ángulo G entre la proyección de estas dos variables será muy cercano al ángulo que ambas variables tienen en la nube de puntos (variables) y así el coseno de G será muy cercano a la correlación entre ambas variables (ver el siguiente gráfico), luego la interpretación es la siguiente:
  - Si X^j y $X^{j'}$ están cercanas entre si, entonces X^j y $X^{j'}$ son fuerte y positivamente correlacionadas.
  - Si el ángulo entre X^j y $X^{j'}$ es cercano a los 90° entonces NO existe ninguna correlación entre ambas variables.
  - Si X^j y $X^{j'}$ están opuestas al vértice (origen) entonces existe una correlacin fuerte y negativa entre X^j y $X^{j'}$.

7.5. Interpretación de la dualidad en los gráficos