MAT-269: Método de Componentes Principales

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Objetivo:

Reemplazar observaciones p-dimensionales por k combinaciones lineales de las variables donde k es mucho más pequeño que p. La elección de k debe explicar razonablemente la proporción de dispersión total $\operatorname{tr}\boldsymbol{S}$.

Definición 1:

Sea $\boldsymbol{x}$ un vector aleatorio con media $\boldsymbol{\mu}$ y dispersión $\boldsymbol{\Sigma}$. Considere $\boldsymbol{T} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \dots, \boldsymbol{t}_p)$ una matriz ortogonal, tal que

$$\boldsymbol{T}^\top \boldsymbol{\Sigma} \boldsymbol{T} = \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_p),$$

donde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Sea

$$\boldsymbol{y} = \boldsymbol{T}^\top (\boldsymbol{x} - \boldsymbol{\mu}),$$

es decir $y_j = \boldsymbol{t}_j^\top (\boldsymbol{x} - \boldsymbol{\mu})$, es llamado el j-ésimo componente principal de $\boldsymbol{x}$ para $j = 1, \dots, p$ y $\boldsymbol{z}_j = y_j / \lambda_j^{1/2}$ es llamado el j-ésimo componente principal estandarizado.

Resultado 1:

Los y_j son no correlacionados y $\operatorname{var}(y_j) = \lambda_j$. En efecto,

$$\begin{aligned}\operatorname{Cov}(\boldsymbol{y}) &= \operatorname{Cov}(\boldsymbol{T}^\top(\boldsymbol{x} - \boldsymbol{\mu})) = \boldsymbol{T}^\top \operatorname{Cov}(\boldsymbol{x})\boldsymbol{T} \\ &= \boldsymbol{T}^\top \boldsymbol{\Sigma} \boldsymbol{T} = \boldsymbol{\Lambda}.\end{aligned}$$

Además, se tiene que

$$\operatorname{Cov}(\boldsymbol{z}) = \operatorname{Cov}(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{y}) = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{I}.$$

Propiedad 1:

Los componentes principales $y_j = \boldsymbol{t}_j^\top(\boldsymbol{x} - \boldsymbol{\mu})$, $j = 1, \dots, p$ tienen las siguientes propiedades:

(i) Para cualquier vector $\boldsymbol{a}_1$ de largo $\|\boldsymbol{a}_1\| = 1$, var$(\boldsymbol{a}_1^\top \boldsymbol{x})$ alcanza su valor máximo λ_1 cuando $\boldsymbol{a}_1 = \boldsymbol{t}_1$.

(ii) Para cualquier vector unitario tal que $\boldsymbol{a}_j^\top \boldsymbol{t}_i = 0$ $(i = 1, \dots, j-1)$ var$(\boldsymbol{a}_j^\top \boldsymbol{x})$ alcanza su valor máximo λ_j cuando $\boldsymbol{a}_j = \boldsymbol{t}_j$.

(iii) $\sum_{j=1}^p \text{var}(y_j) = \sum_{j=1}^p \text{var}(x_j) = \operatorname{tr} \boldsymbol{\Sigma}$.

Previo:

Considere $\boldsymbol{T}^\top\boldsymbol{A}\boldsymbol{T} = \boldsymbol{\Lambda}$ con $\boldsymbol{T}$ matriz ortogonal y $\boldsymbol{\Lambda}$ matriz diagonal. Sea

$$\boldsymbol{x} = \boldsymbol{T}\boldsymbol{y} = y_1\boldsymbol{t}_1 + \cdots + y_p\boldsymbol{t}_p.$$

Entonces

$$\frac{\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top\boldsymbol{x}} = \frac{\boldsymbol{y}^\top\boldsymbol{T}^\top\boldsymbol{A}\boldsymbol{T}\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}} = \frac{\boldsymbol{y}^\top\boldsymbol{\Lambda}\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}} = \frac{\sum_{i=1}^p \lambda_i y_i^2}{\boldsymbol{y}^\top\boldsymbol{y}},$$

como $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, tenemos

$$\frac{\sum_{i=1}^p \lambda_i y_i^2}{\boldsymbol{y}^\top\boldsymbol{y}} \leq \lambda_1 \frac{\sum_{i=1}^p y_i^2}{\boldsymbol{y}^\top\boldsymbol{y}} = \lambda_1,$$

con la igualdad sólo si $y_1 = 1, y_2 = 0, \ldots, y_p = 0$, es decir para $\boldsymbol{x} = \boldsymbol{t}_1$.

Demostración:

Note que $\operatorname{var}(\boldsymbol{a}_1^\top \boldsymbol{x}) = \boldsymbol{a}_1^\top \operatorname{Cov}(\boldsymbol{x})\boldsymbol{a}_1 = \boldsymbol{a}_1^\top \boldsymbol{\Sigma}\boldsymbol{a}_1$, lo que permite mostrar (i).

Para verificar (ii), como $\boldsymbol{t}_1, \boldsymbol{t}_2, \dots, \boldsymbol{t}_p$ son mutuamente ortogonales, ellos forman una base para $\mathbb{R}^p$. Tenemos que $\boldsymbol{a}_j$ puede ser expresado como

$$\boldsymbol{a}_j = c_j \boldsymbol{t}_j + c_{j+1}\boldsymbol{t}_{j+1} + \cdots + c_p \boldsymbol{t}_p,$$

además $\boldsymbol{a}_j \boldsymbol{a}_j = 1 = \sum_{r=j}^p c_r^2 \boldsymbol{t}_r^\top \boldsymbol{t}_r = \sum_{r=j}^p c_r^2$, y

$$\begin{aligned}
\operatorname{var}(\boldsymbol{a}_j^\top \boldsymbol{x}) &= \boldsymbol{a}_j^\top \Big(\sum_{r=j}^p c_r \boldsymbol{\Sigma}\boldsymbol{t}_r\Big) = \boldsymbol{a}_j^\top \sum_{r=j}^p c_r \lambda_r \boldsymbol{t}_r = \Big(\sum_{r=j}^p c_r \boldsymbol{t}_r\Big)^\top \Big(\sum_{r=j}^p c_r \lambda_r \boldsymbol{t}_r\Big) \\
&= \sum_{r=j}^p c_r^2 \lambda_r \geq \lambda_j \sum_{r=j}^p c_r^2 = \lambda_j,
\end{aligned}$$

con la igualdad si y sólo si $\boldsymbol{a}_j = \boldsymbol{t}_j$. De este modo, $\operatorname{var}(\boldsymbol{a}_j^\top \boldsymbol{x})$ es maximizada en $\boldsymbol{a}_j = \boldsymbol{t}_j$.

Finalmente, (iii) sigue, pues:

$$\sum_{j=1}^p \operatorname{var}(y_j) = \operatorname{tr}\boldsymbol{\Lambda} = \operatorname{tr}\boldsymbol{\Lambda}\boldsymbol{T}^\top\boldsymbol{T} = \operatorname{tr}\boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^\top = \operatorname{tr}\boldsymbol{\Sigma}$$

Interpretación:

- (i) y (ii) indica que la componente y_1 es la combinación lineal normalizada de los elementos de $\boldsymbol{x} - \boldsymbol{\mu}$ con varianza máxima λ_1.

- Ahora $\boldsymbol{a}^\top \boldsymbol{x}$ es no correlacionado con y_1, esto es

$$\begin{aligned} \operatorname{Cov}(\boldsymbol{a}^\top \boldsymbol{x}, \boldsymbol{t}_1^\top (\boldsymbol{x} - \boldsymbol{\mu})) &= \operatorname{Cov}(\boldsymbol{a}^\top \boldsymbol{x}, \boldsymbol{t}_1^\top \boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{t} \\ &= \lambda \boldsymbol{a}^\top \boldsymbol{t}_1 = 0, \end{aligned}$$

  si y sólo si $\boldsymbol{a} \perp \boldsymbol{t}_1$. De este modo, desde (ii) y_2 es la combinación lineal no correlacionada con y_1 con varianza máxima $(= \lambda_2)$.

- En general y_j es la combinación lineal no correlacionada con $y_1, \dots, y_{j-1}$ con varianza máxima λ_j.

Interpretación:

- (iii) provee una técnica simple para decidir cual k debe ser escogido. La razón
$$r_k = \frac{\lambda_j}{\sum_{r=1}^{p} \lambda_r} = \frac{\operatorname{var} y_j}{\operatorname{tr} \boldsymbol{\Sigma}},$$
mide la contribución de y_j a $\operatorname{tr} \boldsymbol{\Sigma}$ la variación total de $\boldsymbol{x}$. De este modo, podemos incorporar componentes sucesivos $y_1, y_2, \dots$ y detener en y_k cuando r_k sea cercano a la unidad.

- Si se estandariza los x_j y se trabaja con $x_j^* = (x_j - \mu_j)/\sigma_{jj}^{1/2}$. Entonces $\operatorname{Cov}(\boldsymbol{x}^*) = \boldsymbol{R}$ la matriz de correlación. Se puede extraer un nuevo conjunto de componentes principales,
$$\boldsymbol{y}^* = \boldsymbol{L}^\top \boldsymbol{x}^*,$$
donde $\boldsymbol{L}$ es ortogonal y $\boldsymbol{L}^\top \boldsymbol{R} \boldsymbol{L}$ es diagonal. Sin embargo $\boldsymbol{y}^*$ en general difiere de $\boldsymbol{y}$.

Componentes principales muestrales

En la práctica $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ no son conocidos y deben ser estimados desde $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. Sea

$$\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}}, \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top,$$

y suponga $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdot \geq \widehat{\lambda}_p$ y $\widehat{\boldsymbol{T}} = (\widehat{\boldsymbol{t}}_1, \dots, \widehat{\boldsymbol{t}}_p)$ valores propios y matriz ortogonal de vectores propios de $\widehat{\boldsymbol{\Sigma}}$.

Podemos definir el vector de componentes principales (scores) para cada observación

$$\boldsymbol{y}_i = \widehat{\boldsymbol{T}}^\top (\boldsymbol{x}_i - \overline{\boldsymbol{x}}), \qquad i = 1, \dots, n,$$

obteniendo la matriz de datos

$$\boldsymbol{Y}^\top = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n) = \widehat{\boldsymbol{T}}^\top (\boldsymbol{x}_1 - \overline{\boldsymbol{x}}, \dots, \boldsymbol{x}_n - \overline{\boldsymbol{x}}).$$

Componentes principales muestrales

Frecuentemente se prefiere usar $\boldsymbol{S}$ el estimador insesgado de $\boldsymbol{\Sigma}$ en lugar de $\widehat{\boldsymbol{\Sigma}}$ para definir las componentes principales. En este caso

$$\boldsymbol{S}\,\widehat{\boldsymbol{t}}_j = \frac{n}{n-1}\widehat{\boldsymbol{\Sigma}}\,\widehat{\boldsymbol{t}}_j = \Big(\frac{n}{n-1}\widehat{\lambda}_j\Big)\widehat{\boldsymbol{t}}_j,$$

y los valores propios de $\boldsymbol{S}$ son $n\widehat{\lambda}_j/(n-1)$.

Si se utiliza $\widehat{\lambda}_j/\operatorname{tr}\widehat{\boldsymbol{\Sigma}}$ para determinar la magnitud relativa de un valor propio, entonces el factor de escala $n/(n-1)$ calcela y ambos enfoques (usando $\widehat{\boldsymbol{\Sigma}}$ o $\boldsymbol{S}$) son idénticos.

Resultado 2:

Sea $\boldsymbol{A} \in \mathbb{R}^{p\times k}$ y considere $\boldsymbol{y}_i(k) = \boldsymbol{g}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$, $i = 1, \dots, n$, para cualquier función $\boldsymbol{g} : \mathbb{R}^p \to \mathbb{R}^k$ tal que $\overline{\boldsymbol{y}}(k) = \boldsymbol{0}$. Si f es estrictamente creciente e invariante bajo transformaciones ortogonales. Entonces,

$$f\Big\{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{y}_i(k))(\boldsymbol{x}_i - \overline{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{y}_i(k))^\top\Big\},$$

es minimizada con respecto a $\boldsymbol{A}$ y $\boldsymbol{g}$ cuando

$$\boldsymbol{A} = (\widehat{\boldsymbol{t}}_1, \dots, \widehat{\boldsymbol{t}}_k) = \widehat{\boldsymbol{T}}_1, \qquad \boldsymbol{g}(\boldsymbol{x} - \overline{\boldsymbol{x}}) = \widehat{\boldsymbol{T}}_1^\top(\boldsymbol{x} - \overline{\boldsymbol{x}}).$$

Previo (Seber, 1984; pp. 177):

Sea $\boldsymbol{T} = (\boldsymbol{T}_1, \boldsymbol{T}_2)$ donde $\boldsymbol{T}_1 = (\boldsymbol{t}_1, \dots, \boldsymbol{t}_k)$ definida desde $\boldsymbol{T}^\top\boldsymbol{\Sigma}\boldsymbol{T} = \boldsymbol{\Lambda}$. Entonces

$$f(\boldsymbol{\Delta}) = f(\mathsf{Cov}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{y}(k))),$$

es minimizada cuando $\boldsymbol{A}\boldsymbol{y}(k) = \boldsymbol{T}_1\boldsymbol{T}_1^\top\boldsymbol{x}$.

Demostración:

Sea $\boldsymbol{v}$ un vector aleatorio tomando el valor $\boldsymbol{x}_i$ $(i = 1, \dots, n)$ con probabilidad $\frac{1}{n}$, $\boldsymbol{w} = \boldsymbol{v} - \overline{\boldsymbol{x}}$ y $\boldsymbol{y}(k) = \boldsymbol{g}(\boldsymbol{w})$. Entonces $\boldsymbol{\eta} = \mathsf{E}(\boldsymbol{v}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}}$, $\mathsf{E}(\boldsymbol{w}) = \boldsymbol{0}$, $\mathsf{E}(\boldsymbol{y}(k)) = \overline{\boldsymbol{y}}(k) = \boldsymbol{0}$ y

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{w}) = \mathsf{Cov}(\boldsymbol{v}) = \mathsf{E}\{(\boldsymbol{v} - \boldsymbol{\eta})(\boldsymbol{v} - \boldsymbol{\eta})^\top\} &= \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\eta})(\boldsymbol{x}_i - \boldsymbol{\eta})^\top \\ = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top = \widehat{\boldsymbol{\Sigma}},\end{aligned}$$

y de ahí que

$$\begin{aligned}&f\Big\{ \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{y}_i(k))(\boldsymbol{x}_i - \overline{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{y}_i(k))^\top \Big\} \\ &\qquad = f\big\{ \mathsf{E}[(\boldsymbol{w} - \boldsymbol{A}\boldsymbol{y}(k))(\boldsymbol{w} - \boldsymbol{A}\boldsymbol{y}(k))^\top] \big\} = f\big(\mathsf{Cov}(\boldsymbol{w} - \boldsymbol{A}\boldsymbol{y}(k)) \big),\end{aligned}$$

que es minimizada para todos los vectores $\boldsymbol{y}(k)$ tales que $\mathsf{E}(\boldsymbol{y}(k)) = \boldsymbol{0}$ cuando $\boldsymbol{A}\boldsymbol{y}(k) = \widehat{\boldsymbol{T}}_1 \widehat{\boldsymbol{T}}_1^\top \boldsymbol{w}$ esta ecuación se satisface si hacemos $\boldsymbol{A} = \widehat{\boldsymbol{T}}_1$ y $\boldsymbol{g}(\boldsymbol{w}) = \widehat{\boldsymbol{T}}_1{}^\top \boldsymbol{w}$.

Calculando las componentes principales muestrales

En la práctica se recomienda obtener las componentes principales usando la descomposición valor singular (SVD) de la matriz de datos centrados,

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{z}_1^\top \\ \vdots \\ \boldsymbol{z}_n^\top \end{pmatrix},$$

con $\boldsymbol{z}_i = \boldsymbol{x}_i - \overline{\boldsymbol{x}}$, para $i = 1, \dots, n$. Es fácil notar que:

$$\boldsymbol{S} = \frac{1}{n-1} \boldsymbol{Z}^\top \boldsymbol{Z}.$$

De este modo, obtener la SVD de $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ con $\boldsymbol{U} \in \mathbb{R}^{n \times p}$ tal que $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_p$, $\boldsymbol{D} = \operatorname{diag}(d_1, \dots, d_p)$ de valores singulares $(d_1 \geq d_2 \geq \cdots \geq d_p)$. Es fácil notar que

$$d_1^2/(n-1), \dots, d_p^2/(n-1),$$

son los valores propios de $\boldsymbol{S}$, mientras que las columnas de $\boldsymbol{V}$ corresponden a los vectores propios de $\boldsymbol{S}$. De este modo, los scores son calculados como sigue:

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{V}.$$