Főkomponens analízis

Ispány Márton és Jeszenszky Péter

2016. január 7.

Adatmátrix

Definíció (adatmátrix)

Adatmátrixnak nevezünk egy m sorból és n oszlopból álló $\mathbf{D}$ mátrixot, amelyben a sorok adatobjektumokat, az oszlopok pedig attribútumokat reprezentálnak.
Ha $\mathbf{D} = (d_{ij})_{m\times n}$, akkor jelölje $\mathbf{d}_{i*}$ a mátrix i-edik sorát, $\mathbf{d}_{*j}$ pedig a mátrix j-edik oszlopát, ahol $i = 1, \ldots, m$ és $j = 1, \ldots, n$.

Megjegyzés

A továbbiakban $\mathbf{D} \in \mathbb{R}^{m\times n}$ adatmátrixokat tekintünk. Ekkor értelemszerű, hogy $\mathbf{d}_{i*} \in \mathbb{R}^n$ és $\mathbf{d}_{*j} \in \mathbb{R}^m$.

Sajátérték, sajátvektor

Definíció (sajátérték, sajátvektor)

Egy $\mathbf{A} \in \mathbb{R}^{n \times n}$ mátrix esetén $\lambda \in \mathbb{R}$ **sajátérték** és $\mathbf{x} \in \mathbb{R}^n$ **sajátvektor**, ha

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x} \quad \text{és} \quad \mathbf{x} \neq \mathbf{0}.$$

Kovariancia

Definíció (kovariancia)

Legyen $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ és $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, ekkor

$$\operatorname{cov}(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})(y_i - \overline{\mathbf{y}})$$

ahol

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{és} \quad \overline{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^{n} y_i .$$

Kovariancia mátrix

Definíció

Egy $\mathbf{D} \in \mathbb{R}^{m \times n}$ adatmátrix **kovariancia mátrixá**nak nevezzük az $\mathbf{S} = (s_{ij})_{n \times n} \in \mathbb{R}^{n \times n}$ mátrixot, ahol

$$s_{ij} = \operatorname{cov}(\mathbf{d}_{*i}, \mathbf{d}_{*j}), \quad i, j = 1, \ldots, n.$$

Főkomponens analízis (principal component analysis, PCA)

Dimenzió redukciós eljárás, amelynek során olyan transzformációt keresünk egy adatmátrixhoz, amelynek alkalmazása az alábbi tulajdonságokkal rendelkező transzformált adatmátrixot eredményezi:

- Minden attribútumpár esetén 0 a kovariancia.
- Az attribútumok annak sorrendjében rendezettek, hogy milyen mértékben járulnak hozzá a szóráshoz: az első attribútum járul hozzá legnagyobb mértékben a szóráshoz, az utolsó a legkevésbé.

PCA végrehajtása: kiindulás

Legyen adott a $\mathbf{D} \in \mathbb{R}^{m \times n}$ adatmátrix, amelyre teljesül, hogy minden oszlopban 0 az attribútumértékek átlaga!
Ha ez nem teljesül, akkor minden attribútumértékből vonjuk ki az oszlopának átlagát:

$$d_{ij} \leftarrow d_{ij} - \overline{\mathbf{d}_{*j}},$$

ahol $i = 1, \ldots, m$, $j = 1, \ldots, n$.

PCA végrehajtása: az eljárás lépései

1. Határozzuk meg a $\mathbf{D}$ adatmátrix $\mathbf{S}$ kovariancia mátrixát! Ha minden oszlopban 0 az attribútumértékek átlaga, akkor $\mathbf{S} = \mathbf{D}^T\mathbf{D}/(n-1)$.
2. Határozzuk meg az $\mathbf{S}$ kovariancia mátrix sajátértékeit és sajátvektorait! Legyenek $\lambda_1, \ldots, \lambda_n$ a sajátértékek, amelyeket rendezzük úgy, hogy $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_{n-1} \geq \lambda_n$ teljesüljön. Legyenek $\mathbf{x}_1, \ldots, \mathbf{x}_n$ a sajátértékekhez tartozó egységnyi hosszú sajátvektorok.
3. Képezzünk az alábbi $\mathbf{X}$ mátrixot, amelynek oszlopvektorai az $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sajátvektorok:

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n].$$

4. Transzformáljuk az adatmátrixot az alábbi módon:

$$\mathbf{D}' = \mathbf{DX}.$$

A PCA tulajdonságai

A $\mathbf{D}' = \mathbf{DX}$ transzformált adatmátrixra teljesülnek az alábbiak:

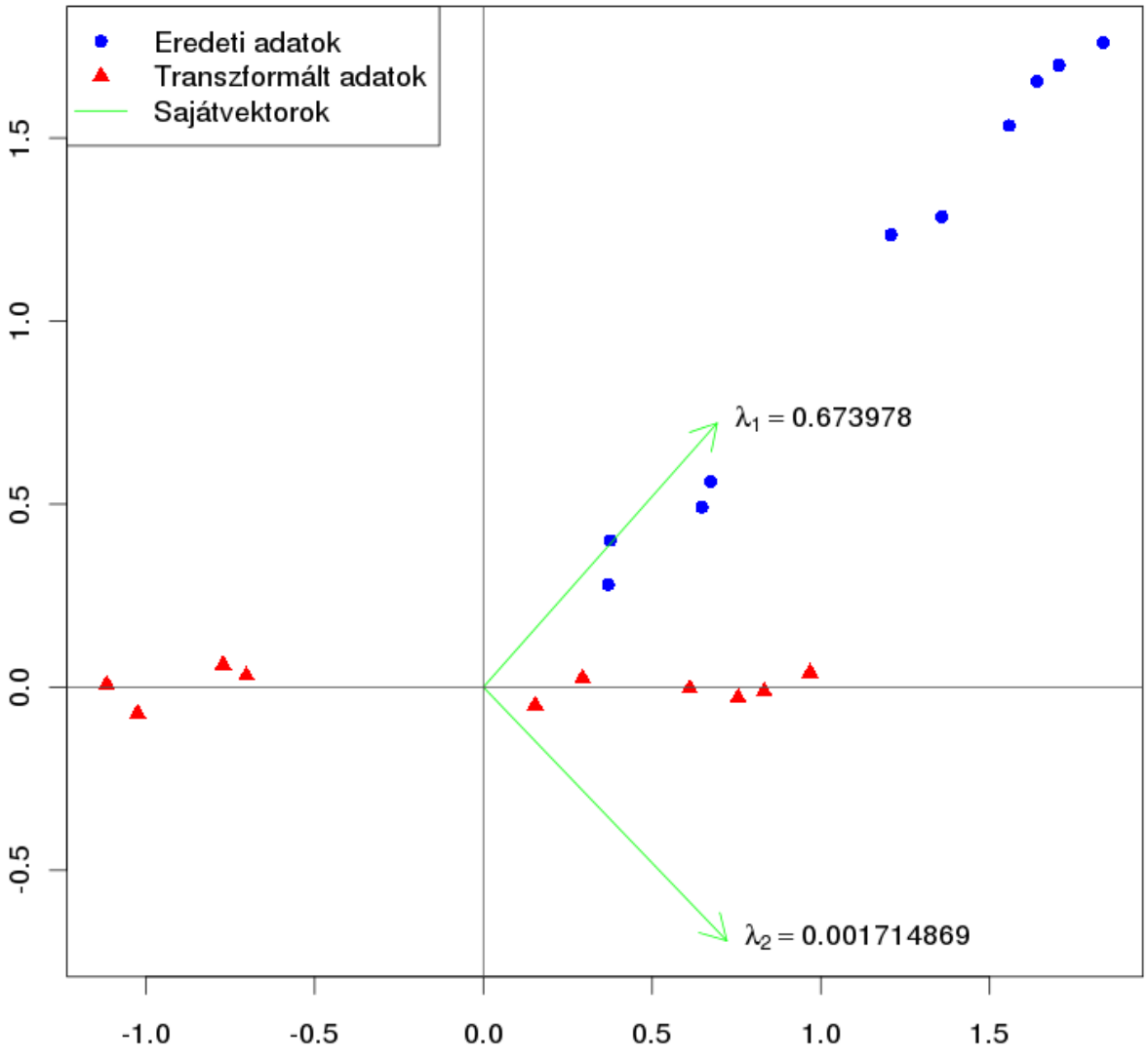
- Minden attribútum a régi attribútumok lineáris kombinációja. (A lineáris kombinációkban a súlyok a sajátvektorok komponensei.)
- A j-edik attribútum varianciája pontosan λ_j.
- Megegyezik a régi és új attribútumok varianciáinak összege.

Az új attribútumokat nevezzük **főkomponensek**nek (az első új attribútum az első főkomponens, stb.).

Megjegyzés

A legnagyobb sajátértékhez tartozó sajátvektor adja meg azt az irányt, amelyben a legnagyobb az adatok varianciája. A második legnagyobb sajátértékhez tartozó sajátvektor merőleges az előzőre, másrészt ebben az irányban a legnagyobb a megmaradó variancia, stb.

PCA szemléltetése



PCA végrehajtása R-ben lépésről-lépésre

Az alábbi módon végezhető el a főkomponens analízis például a

$$\mathbf{D} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}$$

adatmátrixra:

```
> D <- cbind(c(3, 2), c(2, 3))
> D <- scale(D, center=TRUE, scale=FALSE)
> S <- cov(D)
> X <- eigen(S)$vectors
> DX <- D %*% X
```

Eredményül a $\mathbf{DX} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 \end{pmatrix}$ transzformált adatmátrixot kapjuk.

(Ne feledjük, hogy az R numerikusan számol!)

PCA végrehajtása R-ben

Megjegyzés

A `prcomp()` függvény szolgál főkomponens analízis végrehajtására, tehát az előbbi sorok helyett elegendő ennyi:

```
> D <- cbind(c(3, 2), c(2, 3))
> DX <- prcomp(D)$x
```

A `prcomp()` függvény visszatérési értéke egy olyan lista, amelynek komponensei az alábbiak:

- `sdev`: a főkomponensek szórását – azaz a sajátértékek négyzetgyökeit – tartalmazó vektor,
- `rotation`: a sajátvektorokból képzett transzformációs mátrix,
- `x`: a transzformált adatmátrix.