TDVI: Inteligencia Artificial

Análisis de componentes principales

UTDT - LTD

Estructura de la clase

- Matriz de varianzas y covarianzas
- Análisis de componentes principales

Estructura de la clase

- Matriz de varianzas y covarianzas
- Análisis de componentes principales

Matriz de varianzas y covarianzas

Supongamos que tenemos estos dos conjuntos de datos:

```
dinero_1 = np.array([950, 960, 970, 990, 930])
dinero_2 = np.array([950, 1850, 50, 800, 1150])
```

¿Cuánto vale la media del dinero en cada aula?

```
dinero_1.mean()  # 960
dinero_2.mean()  # 960
```

¿Qué se está perdiendo la medida?

Para captar la dispersión de los datos se suele utilizar la varianza:

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

```
((dinero_1 - dinero_1.mean())**2).mean()  # 376   | dinero_1.var()
((dinero_2 - dinero_2.mean())**2).mean()  # 336400
```

Matriz de varianzas y covarianzas

La covarianza mide la relación lineal entre dos variables. La misma indica qué tanto ocurre que cuando una variable (x_1) está encima de su respectiva media, otra variable (x_2) se encuentra también por encima de su respectiva media

$$cov(X_j, X_k) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})$$

¿En qué unidad se encuentra expresada la covarianza?

A la matriz que contiene las varianzas y covarianzas entre todas las variables de un conjunto de datos se la llama matriz de covarianzas

IMPORTANTE: A la suma de los elementos de la diagonal la llamaremos varianza total

Matriz de varianzas y covarianzas

La covarianza se ve afectada por la unidad de medida de las variables, esto algo molesto

Se suele utilizar el coeficiente de correlación de Pearson (varía entre -1 y 1)

$$\rho_{jk} = \frac{cov(X_j, X_k)}{\sigma_{X_j}\sigma_{X_k}}$$

A la matriz que contiene las correlaciones, se la llama matriz de correlaciones

La matriz de correlaciones es igual a la matriz de covarianzas de las variables estandarizadas (*z*-scores)

Matriz de varianzas y covarianzas

Probémoslo en Python. Veamos el bloque 1 de código

Estructura de la clase

- Matriz de varianzas y covarianzas
- Análisis de componentes principales

Análisis de componentes principales

El análisis de componentes principales (PCA) es una técnica de aprendizaje no supervisado

PCA nos permite resumir gran parte de la variabilidad de un conjunto de datos en un conjunto de menor de variables (igualmente estas nuevas variables son combinaciones lineales de las anteriores)

Análisis de componentes principales

PCA busca una representación de baja dimensión de los datos originales tal que se conserve la mayor variabilidad/varianza posible

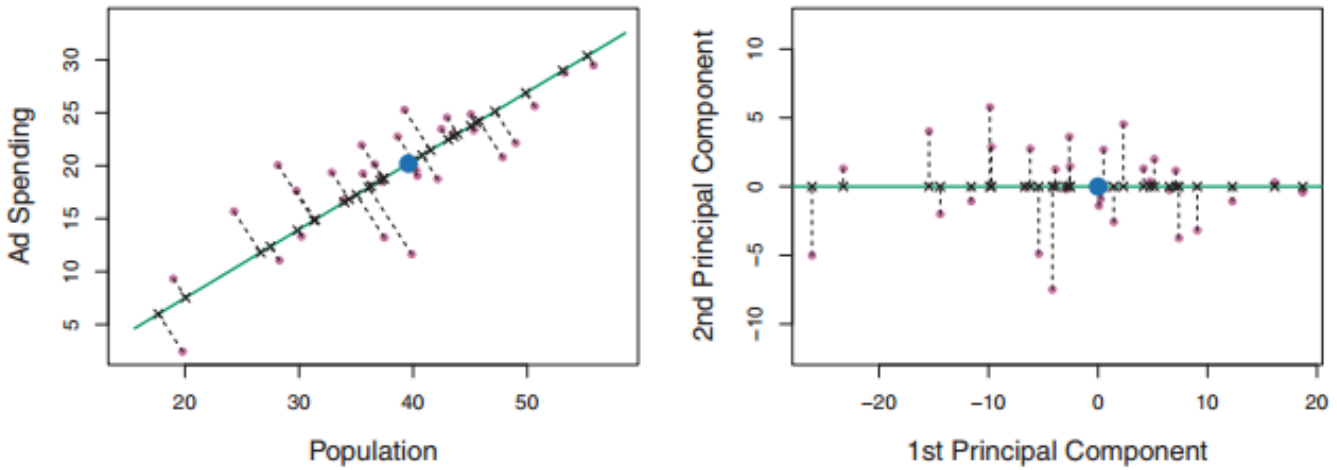


FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all n of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents* $(\overline{\texttt{pop}}, \overline{\texttt{ad}})$. Right: *The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.*

Análisis de componentes principales

El primer componente principal (Z_1) de un conjunto de variables $X_1, X_2, \dots, X_p$ será la combinación lineal "normalizada" de las variables originales ($z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}$ tal que $\sum_{j=1}^{p} \phi_{j1}^2 = 1$)

Asumiendo, sin pérdida de generalidad, que las variables originales tienen media cero, entonces la varianza de Z_1 vendrá dada por:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2$$

Es objetivo será encontrar los valores de ϕ_1 que hacen que Z_1 tenga la mayor varianza posible. El primer componente principal surge de resolver el siguiente problema de optimización:

$$\underset{\phi_{11},\dots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 \right\} \text{ subject to } \sum_{j=1}^{p}\phi_{j1}^2 = 1.$$

Análisis de componentes principales

Una vez encontrado el primer componente, el segundo componente (Z_2) vendrá dado por otra combinación lineal de las variables originales que tenga la mayor varianza posible y que a la vez no esté correlacionada con Z_1

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \cdots + \phi_{p2}x_{ip}$$

El tercero se busca de la misma manera, sólo que no debe estar correlacionado ni con Z_1 ni con Z_2

Este proceso se puede repetir hasta explicar toda la varianza de los datos de los datos originales, lo que generalmente ocurre cuando la cantidad de componentes es igual a p (el número original de variables)

Análisis de componentes principales

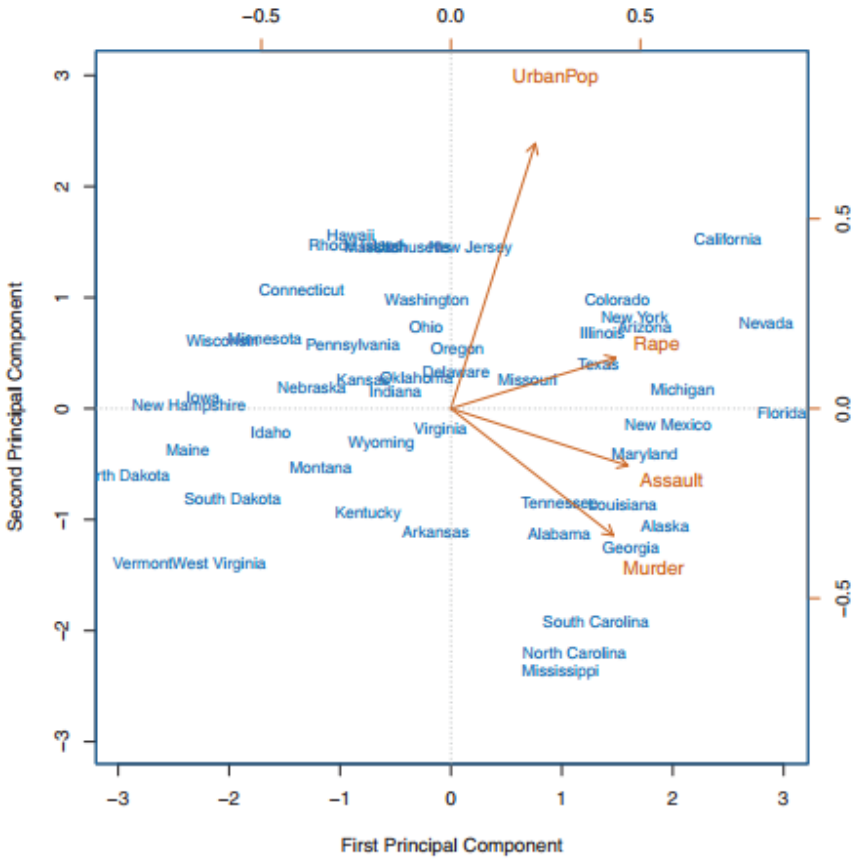


FIGURE 10.1. *The first two principal components for the* `USArrests` *data. The blue state names represent the scores for the first two principal components. The orange arrows indicate the first two principal component loading vectors (with axes on the top and right). For example, the loading for* `Rape` *on the first component is* 0.54*, and its loading on the second principal component* 0.17 *(the word* `Rape` *is centered at the point* $(0.54, 0.17)$*). This figure is known as a biplot, because it displays both the principal component scores and the principal component loadings.*

En base a la posición de las observaciones se pueden detectar observaciones similares entre sí

En base a la dirección de las flechas se pueden identificar cuánto influye cada variable original en cada componente (y entender mejor la correlación entre variables)

Análisis de componentes principales

Probémoslo en Python. Veamos el bloque 2 de código

Análisis de componentes principales

Otra forma de pensar componentes principales: proporciona superficies lineales de baja dimensión que son las más cercanas a las observaciones originales

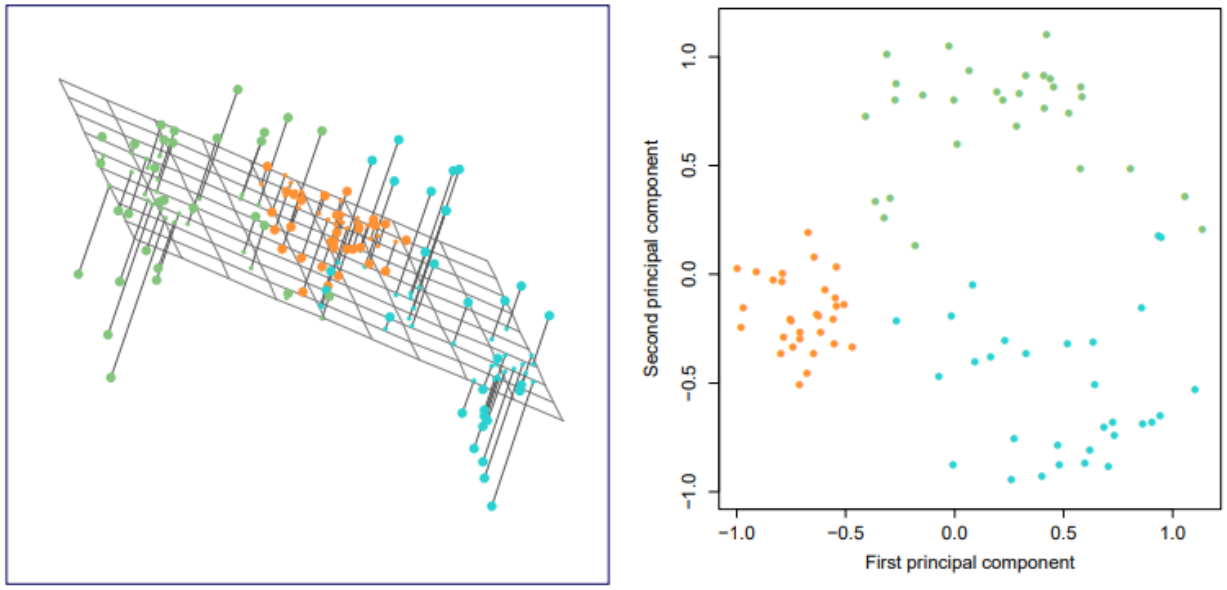


FIGURE 12.2. *Ninety observations simulated in three dimensions. The observations are displayed in color for ease of visualization.* Left: *the first two principal component directions span the plane that best fits the data. The plane is positioned to minimize the sum of squared distances to each point.* Right: *the first two principal component score vectors give the coordinates of the projection of the 90 observations onto the plane.*

Análisis de componentes principales

Importa las escalas de las variables, por este motivo se suele estandarizar a las variables antes de llevar adelante PCA

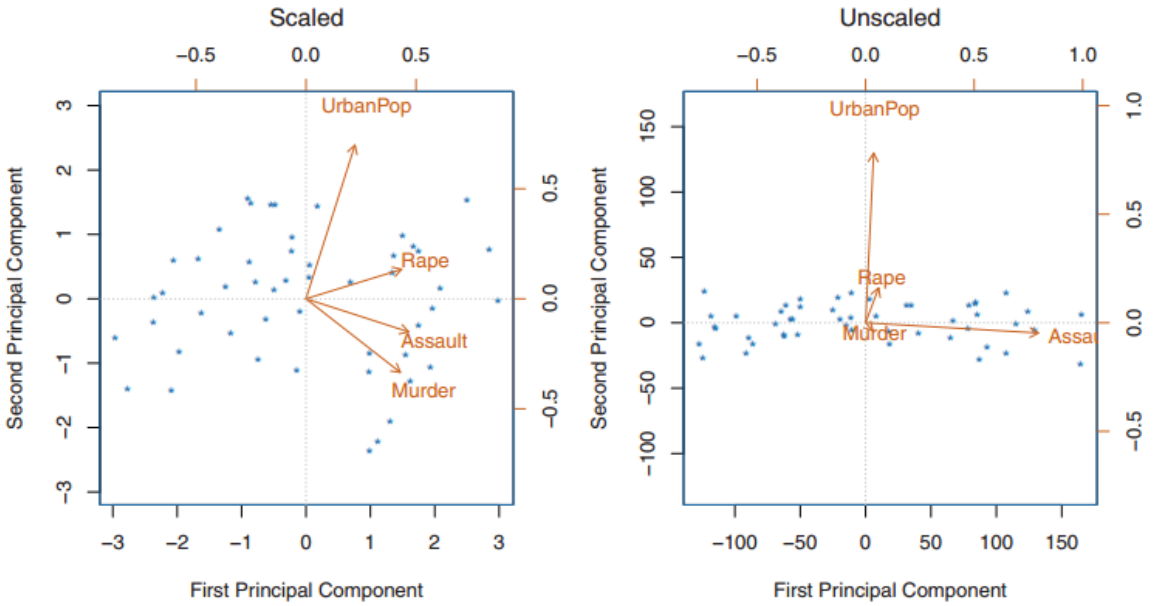


FIGURE 10.3. *Two principal component biplots for the* `USArrests` *data.* Left: *the same as Figure 10.1, with the variables scaled to have unit standard deviations.* Right: *principal components using unscaled data.* `Assault` *has by far the largest loading on the first principal component because it has the highest variance among the four variables. In general, scaling the variables to have standard deviation one is recommended.*

Análisis de componentes principales

Resulta importante analizar cuánta información/varianza se captura al representar los datos originales en base a *m* componentes principales

Para esto sirve analizar la proporción de la varianza explicada por cada componente

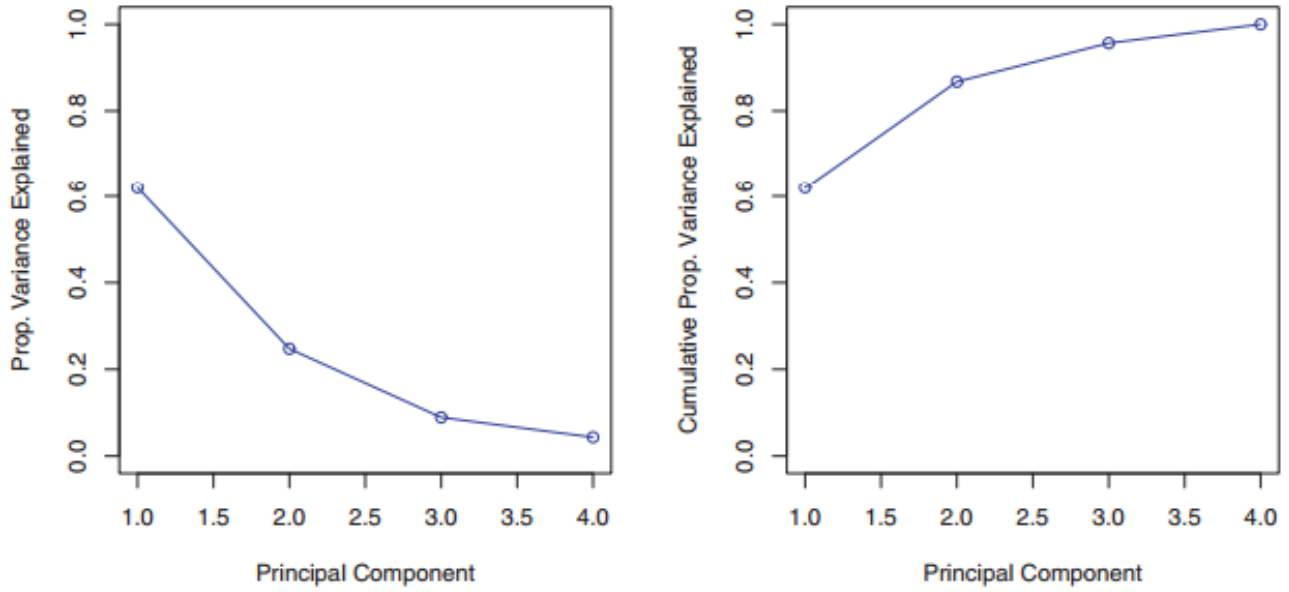
Varianza total original: $\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$ (se asume que X_j tiene media 0)

Varianza explicada por el componente *m*: $\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2$

Proporción de la varianza explicada por *m*: $\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$

Análisis de componentes principales

Resulta útil graficar la proporción de la varianza explicada por los componentes principales (scree plot)



Análisis de componentes principales

Probémoslo en Python. Veamos el bloque 3 de código

Análisis de componentes principales

Consideraciones:

- Cuántos componentes considerar depende del problema que se trate. No hay un criterio objetivo claro para definir el número

- Muchas veces se analizan gráficos como los presentados en la slides anteriores (scree plots) y se busca un codo en la figura

- A veces se afirma que es recomendable dejar la menor cantidad de componentes tal que se explique al menos un 80% de la variabilidad de los datos

- Caso especial: si se usa el análisis de componentes principales como una etapa de pre-procesamiento de los datos en un problema de aprendizaje supervisado, la cantidad de componentes puede entenderse como un hiperparámetro más a optimizar

Análisis de componentes principales

Tarea

Prueben llevar adelante un análisis de componentes principales en base a los datos “USJudgeRatings”.

```
from statsmodels.datasets import get_rdataset
USJudgeRatings = get_rdataset("USJudgeRatings").data

# Sigan ustedes!
```

Bibliografía

ISLP. Sección 12.2