Métodos estadísticos de exploración

ACP: Análisis de componentes principales

Profesor: Rodrigo Assar C.

Datos

- Matriz de datos:

$$X = \begin{matrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{matrix}$$

- Matriz de covarianzas (empíricas):

$$\Sigma_X = \begin{matrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{matrix}$$

$$\sigma_{ii} = \frac{1}{n}\sum_{k=1}^{n}(X_{ki} - \overline{X}_i)^2 \qquad \sigma_{ij} = \frac{1}{n}\sum_{k=1}^{n}(X_{ki} - \overline{X}_i)(X_{kj} - \overline{X}_j)$$

Las componentes principales

- Nuevas variables: combinaciones lineales de originales. $y_i = a_{1i}v_1 + a_{2i}v_2 + \ldots + a_{pi}v_p$
- Coeficientes elegidos para maximizar la varianza empírica, con norma igual a 1.
- Covarianza entre ejes principales distintos es cero.

$$Var(y_i) = \sum_{k=1}^{p} a_{ki}^2 \sigma_{kk} + \sum_{k=1}^{p} \sum_{j=1, j\neq k}^{p} a_{ki} a_{ji} \sigma_{kj} = \vec{a}_i^t \Sigma_X \vec{a}_i$$

$$Cov(y_j, y_k) = \vec{a}_j^t \Sigma_X \vec{a}_k$$

Elección de coeficientes

- $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p (> 0)$ valores propios de Σ_X.
- Vectores propios asociados: $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_p$ elegidos ortonormales.
- Para i de 1 a p: $y_i = \vec{e}_i^t \vec{v}$
- $Var(y_i) = \lambda_i$
- $Cov(y_j, y_k) = 0$
- Se denominan componentes principales a las proyecciones de los datos en los ejes principales.

Información aportada por componentes

- Proporción de la varianza total explicada por la j-ésima componente:

$$\lambda_j \Big/ \sum_{i=1}^{p} \lambda_i = \lambda_j \Big/ \sum_{i=1}^{p} \mathrm{var}(v_i) = \lambda_j \Big/ \sum_{i=1}^{p} \sigma_{ii}$$

- Proporción explicada por las primeras m componentes:

$$\sum_{j=1}^{m} \lambda_j \Big/ \sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{m} \lambda_j \Big/ \sum_{i=1}^{p} \mathrm{var}(v_i) = \sum_{j=1}^{m} \lambda_j \Big/ \sum_{i=1}^{p} \sigma_{ii}$$

Relación: variables nuevas, originales

$$\mathrm{cov}(y_k, v_j) = \lambda_\kappa e_{jk}$$

$$corr(y_k, v_j) = \sqrt{\lambda_\kappa} e_{jk} \Big/ \sqrt{\sigma_{jj}}$$

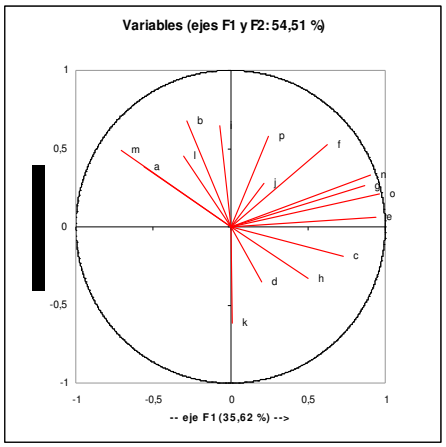
- Usualmente, antes de realizar el ACP se modifica X restando a cada valor en la j-ésima columna el promedio en dicha columna y dividiendo por $\sqrt{\sigma_{jj}}$.

De este modo var(v_j)=1,

$$corr(y_k, v_j) = \sqrt{\lambda_\kappa} e_{jk}$$

- Círculo de correlaciones…

Círculo de correlaciones: Ejemplo



- El eje horizontal corresponde al valor de la correlación con la primera componente principal, el vertical respecto a la segunda componente.
- Se observa por ejemplo que corr(e,F1) es cercana a 1, corr(e,F2) es positiva pero cercana a 0.