Estadística Multivariada

Cuestionario: Análisis por Correspondencias, Análisis Factorial y Clústers

Yair Castillo Emilio Valencia Alec Torres

Mayo 2024

1 Análisis por correspondencias

1.1 Análisis por Correspondencias (AC)

El análisis por correspondencias es una técnica multivariada que se utiliza para analizar tablas de contingencia. Su objetivo principal es transformar una tabla de frecuencias en una representación gráfica, donde se pueden observar las relaciones entre filas y columnas.

1.2 Pasos del Análisis por Correspondencias

1. **Construcción de la tabla de contingencia:** Se empieza con una tabla que contiene las frecuencias de ocurrencia de diferentes categorías.
2. **Cálculo de perfiles:** Se calculan las proporciones de cada categoría.
3. **Cálculo de distancias:** Se calculan las distancias entre las proporciones.
4. **Descomposición en eigenvalores:** Se descompone la matriz de distancias para obtener las coordenadas principales.
5. **Visualización:** Se representan las filas y columnas en un espacio de menor dimensión (generalmente 2D) para facilitar la interpretación.

1.3 Tabla Comparativa

	Análisis por Correspondencias (AC)	**Análisis por Componentes Principales (PCA)**
Tipo de datos	Categóricos	Numéricos
Objetivo principal	Visualizar relaciones entre categorías en tablas de contingencia	Reducir la dimensionalidad manteniendo la varianza
Método matemático	Descomposición en eigenvalores de una matriz de contingencia	Descomposición en eigenvalores de la matriz de covarianza
Visualización	Mapas de correspondencia (2D)	Gráficos de componentes principales
Aplicación típica	Análisis de encuestas, datos de categorización	Datos cuantitativos en ciencia, ingeniería y economía

Table 1: Comparativa entre Análisis por Correspondencias y Análisis por Componentes Principales

2 MCA de iris

El MCA nos ayudó a transfvormar los datos categóricos de iris en un espacio dimensión reducida donde se pueden visualizar y analizar patrones. La separación de los grupos de especies en el gráfico indica que el MCA ha capturado efectivamente las diferencias y similitudes entre las flores de iris basadas en sus características. Esta visualización puede ser útil para entender mejor las relaciones entre las diferentes categorías y cómo ayudan a diferenciar entre las especies.

3 Análisis Factorial

El análisis factorial es una técnica estadística multivariante que se usa para identificar estructuras en un conjunto de variables observadas. Se basa en la idea de que hay factores ocultos que afectan las variables que medimos. Estos factores ayudan a reducir la dimensionalidad de los datos, simplificando su interpretación.

Concepto	**Descripción**
Reducción de Dimensionalidad	Permite condensar un gran número de variables en unos pocos factores, facilitando la interpretación y visualización de los datos.
Identificación de Factores Latentes	Ayuda a descubrir los factores que influyen en las observaciones.
Simplificación y Eficiencia	Simplifica el modelo estadístico, mejora la eficiencia en análisis posteriores y reduce el ruido en los datos.

Table 2: Significado del Análisis Factorial

3.1 Tipos de Análisis Factorial

Tipo	**Descripción**
Análisis de Componentes Principales (PCA)	Explica la mayor cantidad de varianza posible en los datos a través de componentes lineales no correlacionados. Utilizado principalmente para reducción de dimensionalidad.
Análisis Factorial Exploratorio (EFA)	Descubre la estructura básica de los datos sin asumir una forma específica de antemano. Útil para explorar relaciones entre variables.
Análisis Factorial Confirmatorio (CFA)	Prueba hipótesis específicas sobre la estructura de factores latentes previamente identificados. Requiere una teoría previa o un modelo a validar.

Table 3: Tipos de Análisis Factorial

3.2 Proceso de Análisis Factorial

1. **Selección de Variables**: Escoger las variables que serán incluidas en el análisis. Deben estar relacionadas con el fenómeno que se estudia.

2. **Matriz de Correlación**: Crear una matriz de correlación para ver las relaciones entre las variables. Una alta correlación sugiere que las variables pueden compartir un factor común.

3. **Extracción de Factores**: Utilizar métodos como la varianza máxima (varimax) o rotación ortogonal para extraer los factores.

4. **Rotación de Factores**: Aplicar técnicas de rotación para facilitar la interpretación de los factores. La rotación varimax es una opción común que produce factores no correlacionados.

5. **Interpretación de Factores**: Asignar un significado a cada factor basándose en las variables que más contribuyen a él.

3.3 Aplicaciones

- **Psicología**: Para identificar rasgos de personalidad o constructos psicológicos.
- **Marketing**: Para segmentar mercados y entender el comportamiento del consumidor.
- **Educación**: Para evaluar diferentes dimensiones del rendimiento académico.

4 Ejercicios del libro de Rencher

13.1 $\text{var}(y_i) = \text{var}(y_i - \mu_i) = \text{var}(\lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots + \lambda_{im} f_m + \epsilon_i)$

$$= \sum_{j=1}^{m} \lambda_{ij}^2 \,\text{var}(f_j) + \text{var}(\epsilon_i) + \sum_{j \neq k} \lambda_{ij}\lambda_{ik} \,\text{cov}(f_j, f_k)$$

$$+ \sum_{j=1}^{m} \lambda_{ij} \,\text{cov}(f_j, \epsilon_i)$$

$$= \sum_{j=1}^{m} \lambda_{ij}^2 + \psi_i.$$

La última igualdad sigue lo siguiente: $\text{var}(f_j) = 1$, $\text{var}(\epsilon_i) = \psi_i$, $\text{cov}(f_j, f_k) = 0$, and $\text{cov}(f_j, \epsilon_i) = 0$.

13.2 $\text{cov}(\mathbf{y}, \mathbf{f}) = \text{cov}(\boldsymbol{\Lambda}\mathbf{f} + \epsilon, \mathbf{f})$ [por (13.3)]

$$= \text{cov}(\boldsymbol{\Lambda}\mathbf{f}, \mathbf{f}) \quad \text{[por (13.10)]}$$

$$= \mathbb{E}[(\boldsymbol{\Lambda}\mathbf{f} - \mathbb{E}(\boldsymbol{\Lambda}\mathbf{f}))(\mathbf{f} - \mathbb{E}(\mathbf{f}))'] \quad \text{[por (3.31)]}$$

$$= \mathbb{E}[(\boldsymbol{\Lambda}\mathbf{f} - \boldsymbol{\Lambda}\mathbb{E}(\mathbf{f}))(\mathbf{f} - \mathbb{E}(\mathbf{f}))']$$

$$= \boldsymbol{\Lambda}\mathbb{E}[(\mathbf{f} - \mathbb{E}(\mathbf{f}))(\mathbf{f} - \mathbb{E}(\mathbf{f}))']$$

$$= \boldsymbol{\Lambda}\text{cov}(\mathbf{f}) = \boldsymbol{\Lambda} \quad \text{[por (13.7)]}$$

13.3

$$E(\mathbf{f}^*) = E(T'\mathbf{f}) = T'E(\mathbf{f}) = T'0 = 0,$$

$$\text{cov}(\mathbf{f}^*) = \text{cov}(T'\mathbf{f}) = T'\text{cov}(\mathbf{f})T = T'IT = I$$

13.5

$$\sum_{i=1}^{p} \sum_{j=1}^{m} \hat{\lambda}_{ij}^2 = \sum_{i=1}^{p} \left[\sum_{j=1}^{m} \hat{\lambda}_{ij}^2 \right] = \sum_{i=1}^{p} \hat{h}_i^2 \quad \text{[por (13.28)]}$$

Intercambiando el orden de la suma tenemos:

$$\sum_{i=1}^{p} \sum_{j=1}^{m} \hat{\lambda}_{ij}^2 = \sum_{j=1}^{m} \sum_{i=1}^{p} \hat{\lambda}_{ij}^2 = \sum_{j=1}^{m} \theta_j \quad \text{[por (13.29)]}.$$

5 K-means

5.1 Matriz de disimilaridades

$$\begin{bmatrix} 0 & 1 & 2.2361 & 2 & 3 & 4.1231 \\ 1 & 0 & 1.4142 & 2.2361 & 3.1623 & 4.4721 \\ 2.2361 & 1.4142 & 0 & 3.6056 & 4.4721 & 5.8310 \\ 2 & 2.2361 & 3.6056 & 0 & 1 & 2.2361 \\ 3 & 3.1623 & 4.4721 & 1 & 0 & 1.4142 \\ 4.1231 & 4.4721 & 5.8310 & 2.2361 & 1.4142 & 0 \end{bmatrix}$$

5.2 Implementar K-means

Centroides finales:

$$\begin{bmatrix} -0.33333333 & 1.0 \end{bmatrix} y \begin{bmatrix} 3.0 & -0.33333333 \end{bmatrix}$$

Asignar Punto:

El punto (1,1) pertenece al centroide 1

6 Demostración de media muestral

$$\hat{\mu} = \frac{d}{d\mu} \sum_{i=1}^{n} |(x_i - \mu)|^2 = \sum_{i=1}^{n} -2(x_i - \mu) = -2 \left[\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu \right] = -2 \left[\sum_{i=1}^{n} x_i - n\mu \right] = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i = n\mu$$

$$\Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

7 Análisis de clústers

1. **Verdadero** El análisis de clústers se utiliza para agrupar las observaciones según sus características. La idea es encontrar clústers de observaciones que son similares entre sí en términos de las características dadas.

2. **Verdadero** Aunque es menos común que agrupar observaciones, es posible realizar un análisis de clústers en las características basándose en las observaciones. Esto se conoce como "clustering de características" o "co-clustering".

3. **Falso** El análisis por clústers es parte del aprendizaje no supervisado.

8 K-means 2.0

Centroides finales		
	Coordenada 1	Coordenada 2
Centroide 1	0.66	4.33
Centroide 2	5.5	2

Table 4: Centroides finales del análisis de clústers

Punto	Etiqueta
1	0
2	0
3	1
4	1
5	0

Table 5: Etiquetas de los puntos del análisis de clústers

9 K-means 3.0

Centroides finales		
	Coordenada 1	Coordenada 2
Centroide 1	1.5	3.5
Centroide 2	7.0	4.33
Centroide 3	3.66	9.0

Table 6: Centroides finales del análisis de clústers

Punto	Etiqueta
1	2
2	0
3	1
4	2
5	1
6	1
7	0
8	2

Table 7: Etiquetas de los puntos del análisis de clústers

10 Datos financieros

Table 8: Movimientos Finales, Rendimientos y Clusters

Ticker	Movimientos Finales	Rendimientos	Cluster
AAPL	163.560280	0.301509	1
ALSEA.MX	55.592428	0.069857	1
AMZN	127.119995	0.213405	1
BIMBOA.MX	72.001814	0.174854	1
BRK-B	208.350006	0.119742	2
CEMEXCPO.MX	13.429999	0.058931	1
FEMSAUBD.MX	112.488464	0.071421	1
GFNORTEO.MX	127.158035	0.190989	1
GMEXICOB.MX	79.339310	0.178164	1
GOOGL	100.605000	0.203079	1
JNJ	75.407791	0.065541	1
KIMBERA.MX	17.774832	0.096980	1
LIVEPOLC-1.MX	87.730934	0.043765	1
MSFT	302.243355	0.295411	2
NVDA	472.500286	0.503506	0
PE&OLES.MX	270.050034	0.019000	2
TLEVISACPO.MX	67.918924	-0.212091	1
TSLA	398.038669	0.596429	0
V	152.498680	0.177168	1
WMT	31.175812	0.118097	1

Índice de Sharpe óptimo: 1.1278952473791954
Rendimiento óptimo: 0.231569764251138
hahhaha Riesgo óptimo: 0.2053114106023756