Guía de ejercicios Capítulo II

Análisis de Componentes Principales

# Objetivos de la Unidad

|  |  |  |
| --- | --- | --- |
| **#** | **Descripción** | **Preguntas** |
| 1. | Describir los fundamentos matemáticos de modelo *Análisis de Componentes Principales* (PCA). | Todas |
| 2. | Aplicar el modelo PCA en un conjunto de datos perteneciente a un problema específico. | Todas |

# Preguntas

P1. Comportamiento de consumo alimenticio

Un estudio del consumo de diferentes familias francesas intenta una caracterización mediante análisis de componentes principales. Las familias (sujetos) se caracterizan por su condición socioeconómica, clasificando el tipo de trabajo de los jefes de hogar, y por sus tamaños considerando el número de hijos. Los jefes de hogar tienen las siguientes características: (MA#) trabajador manual, (EM#) empleado, (PF#) profesional. Donde el símbolo # puede tomar los valores 2, 3, 4 o 5 y representa el número de hijos de las familias. Se mide el consumo promedio de los siguientes productos (atributos o variables originales *x1*): Pan, Verduras, Frutas, Carnes, Aves, Lácteos y Vinos.

La matriz de correlación de los productos es la siguiente:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Pan** | **Verduras** | **Frutas** | **Carnes** | **Aves** | **Lácteos** | **Vinos** |
| **Pan** | 1,00 |  |  |  |  |  |  |
| **Verduras** | 0,59 | 1,00 |  |  |  |  |  |
| **Frutas** | 0,20 | 0,87 | 1,00 |  |  |  |  |
| **Carnes** | 0,32 | 0,89 | 0,96 | 1,00 |  |  |  |
| **Aves** | 0,25 | 0,83 | 0,93 | 0,98 | 1,00 |  |  |
| **Lácteos** | 0,86 | 0,66 | 0,33 | 0,37 | 0,23 | 1,00 |  |
| **Vinos** | 0,30 | -0,35 | -0,49 | -0,44 | -0,40 | -0.40 | 1,00 |

Después de realizar un análisis de componentes principales, los valores propios para las dos primeras componentes son *1=4,339* y *2=1,829*, los vectores propios se muestran en la siguiente tabla.

|  |  |  |
| --- | --- | --- |
|  | **Componente 1** | **Componente 2** |
| Pan | -0,497 | 0,841 |
| **Verduras** | -0,972 | 0,131 |
| **Frutas** | -0,931 | -0,277 |
| **Carnes** | -0,963 | -0,19 |
| **Aves** | -0,912 | -0,265 |
| **Lácteos** | -0,584 | 0,707 |
| **Vinos** | 0,425 | 0,649 |

La distribución de los grupos familiares en el espacio de las dos primeras componentes principales se muestra en la siguiente figura:

### Comp. 1

### Comp. 2

PF4

EM5

PF5

MA5

MA3

PF3

PF2

MA4

MA2

EM3

EM2

EM4

1. Determine el porcentaje de información que se pierde en el estudio con solo las dos primeras componentes.
2. Intente una caracterización de las componentes inspeccionando sólo la distribución de las familias en el plano de las dos componentes.
3. Dibuje en forma aproximada los productos consumidos (atributos o variables originales) en el espacio de las componentes.
4. Asocie cada componente a los productos consumidos.
5. Intente relacionar los productos consumidos por las familias y sus condiciones socioeconómicas y/o tamaños.

P2. Caracterización billetes falsos

Para caracterizar billetes falsos los bancos suizos realizaron un análisis que consistía en tomar medidas de los billetes. Para el análisis tenía tres grupos diferente de billetes. Originales de papel, originales de plástico y billetes falsos. Cada billete fue caracterizado por las siguientes variables:

LON : Longitud del billete.

LD : Largo de la Diagonal del billete.

AI : Ancho Izquierdo del billete.

AD : Ancho Derecho del billete.

AMI : Ancho Margen Inferior del billete

AMS : Ancho Margen Superior del billete.

A continuación, se realizó un análisis de componentes principales con los siguientes resultados:

- Valores propios: 2,58; 1,34; 0,76; 0,56; 0,50; 0,26.

- Vectores propios para las dos primeras componentes:

|  |  |  |
| --- | --- | --- |
|  | **Componente 1** | **Componente 2** |
| LON | 0,395 | 0,799 |
| LD | 0,207 | 0,345 |
| AI | 0,445 | -0,263 |
| AD | 0,411 | -0,375 |
| AMI | 0,347 | -0,072 |
| AMS | 0,560 | -0,163 |

La distribución de los diferentes tipos de billetes se muestra en la siguiente figura (originales de papel en círculos, originales de plástico en cuadrados y falsos en triángulos):

**CP1**

**CP2**

Se realizó un trabajo similar con monedas midiendo 6 variables de tamaño de éstas y los valores propios del análisis fueron los siguientes: 1,96; 1,54; 1,09; 0,73; 0,40; 0,28.

1. Determine el porcentaje de validez del análisis.
2. Interprete cada una de las componentes
3. Identifique las principales características de los billetes originales.
4. Determine si existen diferencias entre las falsificaciones.
5. Para el análisis de las monedas, ¿se logrará tener una precisión similar a la de los billetes?

P3. Evaluación desempeño académico

Se dispone de 15 alumnos a los cuales se les mide las calificaciones de primer año de enseñanza media, en las siguientes asignaturas: Castellano (C), Matemática (M), Física (Fi), Inglés (I), Filosofía (F), Historia (H), Química (Q) y Gimnasia (G).

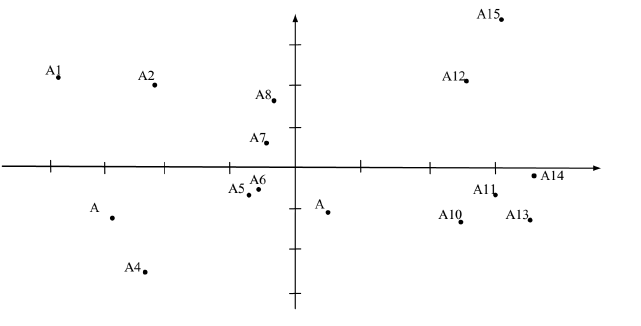
Se posee además, una tabla que relaciona los alumnos con los colegios dónde realizaron la enseñanza básica. Para un análisis de componentes principales se tienen los siguientes resultados:

Valores propios: 3,71; 2,86; 0,954

Vectores propios:

|  |  |  |  |
| --- | --- | --- | --- |
| **Asignatura** | **Componente 1** | **Componente 2** | **Componente 3** |
| Castellano | 0,500 | 0,085 | -0,028 |
| Matemática | -0,113 | 0,555 | 0,133 |
| Física | -0,052 | 0,575 | 0,076 |
| Inglés | 0,499 | 0,037 | -0,005 |
| Física | 0,450 | 0,122 | -0,303 |
| Historia | 0,493 | 0,064 | -0,011 |
| Química | -0,073 | 0,574 | -0,021 |
| Gimnasia | 0,187 | -0,069 | 0,940 |

Al aplicar la matriz de transformación (para las primeras dos componentes) a los datos originales se tiene la siguiente distribución de sujetos en el plano principal:



1. Determine la validez del análisis.
2. Ubique las asignaturas en el plano principal de las dos primeras componentes.
3. Caracterice las dos primeras componentes.
4. Usando la matriz de transformación, analice y caracterice la tercera componente.
5. Utilizando la tabla que relaciona alumnos con colegios y suponiendo que la habilidad de cada estudiante está dada por su formación básica, identifique las fortalezas de los colegios de enseñanza.

P4. Caracterización billetes falsos

En un estado venezolano se utilizó ACP para analizar la situación del sector lechero. Las variables utilizadas fueron: SUP: superficie total de la hacienda; VACA: número total de vacas; SANI: índice sanitario; INST: índice de las instalaciones; MAQ: índice de maquinarias; PROM: Promedio de la lecha/vaca. Los resultados del ACP fueron los siguientes:

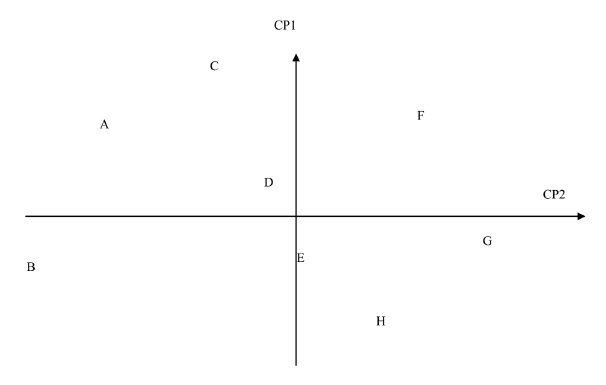
Valores propios: 1,794; 1,341.

Vectores propios ponderados por la raíz de cada valor propio:

|  |  |  |
| --- | --- | --- |
|  | **Componente 1** | **Componente 2** |
| SUP | 0,79 | 0,10 |
| VACA | 0,76 | 0,40 |
| SANI | 0,44 | -0,48 |
| INST | 0,32 | -0,48 |
| MAQ | 0,53 | -0,26 |
| PROM | -0,01 | -0,80 |

Determine una caracterización de las componentes principales, al examinar la relación con las variables originales.

Al graficar las haciendas en el plano de las dos primeras componentes principales se pueden caracterizar 8 tipos de haciendas con problemas particulares, como se muestra en la figura. Utilice esta información para identificar los tipos de haciendas. Determine además el porcentaje de validez del estudio.



P5. Caracterización servicios hospitalarios

Un estudio del Hospital de Andalucía (España) analizó 22.846 ingresos para intentar caracterizar los diferentes servicios (o unidades) del hospital. Los servicios analizados fueron los siguientes: Medicina Interna, Ginecología, Pediatría, Cirugía, Traumatología, Urología, Digestivo, Otorrinolaringología, Cardiología, Neurología, Hematología, Oftalmología y Psiquiatría. Para caracterizar los servicios se midieron 7 variables:

NI : Número de ingresos.

MO : Índice de mortalidad.

RE : Número de reingresos al servicio (por el mismo diagnóstico).

NE : Número de consultas externas al servicio.

ICM : índice promedio de la complejidad de los pacientes admitidos.

ES : Número de estancias por servicio (promedio de días cama usados en cada servicio).

Índice de funcionalidad (representa la eficiencia del servicio)

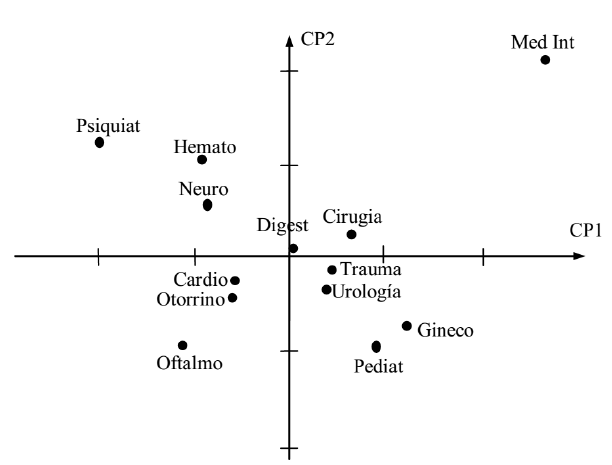
El resultado del análisis de componentes principales muestra los siguientes resultados para las dos primeras componentes:

Valor propio 1: 2,558

Valor propio 2: 1,829

Vector propio 1: [0,860; 0,421; -0,406; -0,250; -0,562; 0,820; 0,663]

Vector propio 2: [-0,066; 0,747; 0,670; 0,388; 0,635; 0,508; 0,078]



1. Determine la validez del análisis.
2. Ubique las variables medidas en el plano de las dos primeras componentes.
3. Interprete cada componente.
4. Clasifique los servicios según la interpretación de cada componente.
5. Seleccione los servicios con mayor carga de trabajo cualitativa y cuantitativa.
6. Determine los servicios más eficientes.

# Respuestas

P1. Comportamiento de consumo

Texto.

P2. Caracterización billetes falsos

Texto.