Estadística Aplicada III ITAM
11 de marzo de 2024 2024-I

Tarea 4. Fecha de entrega: 27 de marzo de 2024.

Lecturas

- Capítulos 7 y 8 del libro de Johnson & Wichern (disponibles en Piazza)
- Nota: "Generating market risk scenarios using principal components analysis: methodological and practical considerations"..
- Principal Component Analysis Application to Statistical Process Control.

Problemas

1. Si dos variables X y Y tienen covarianza $\mathbf{S} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, entonces mostrar que si $c \neq 0$ entonces la primera componente principal está dada por:

$$\sqrt{\frac{c^2}{c^2 + (V_1 - d)^2}} X + \frac{c}{|c|} \sqrt{\frac{(V_1 - d)^2}{c^2 + (V_1 - d)^2}} Y,$$

   donde V_1 es la varianza explicada por la primera componente principal. ¿Cuál es el valor de V_1?

2. Considerar los datos en el archivo `T8-5.DAT` correspondientes a un tramo censal. Suponer que las observaciones en la variable $X_5 =$ valor de la mediana de hogares fue registrada en unidades de diez miles más que de cientos de miles de dólares (es decir, multipliquen todos los datos listados en la quinta columna por 10).

   a. Comparar las estimaciones con los datos en diez miles y cientos de miles (son dos matrices de covarianzas) para las componentes principales en cada caso.
   b. Tratar de obtener la interpretación de las dos primeras componentes principales en cada caso.
   c. Describir cuáles son los efectos en el cambio de escala.

3. Considerar los datos sobre toros en el conjunto de datos `T1 − 10.DAT` sobre toros. Estos datos contienen las características medidas de 76 toros jóvenes (menores a dos años) vendidos en una subasta. Los datos que se incluyen corresponden a las siguientes variables:

   - `Raza`: 1= Angus, 5= Hereford, 8= Simental
   - `PVenta`: precio de venta
   - `YrHgt`: medición al hombro al año (pulgadas)
   - `FtFreBody`: Cuerpo libre de grasa (libras)
   - `PrctFFB`: Porcentaje del cuerpo libre de grasa
   - `Frame`: Cornamenta. Escala de 1 (pequeña) a 8 (grande)
   - `BkFat`: Grasa trasera (en pulgadas)

- `SaleHt`: medición al hombro en el momento de venta (pulgadas)
- `SaleWt`: peso de venta (libras)

Utilizando las 7 últimas variables dadas, hacer un análisis de componentes principales usando la matriz de covarianzas de los datos y la matriz de correlación. El análisis debe incluir lo siguiente:

a. Determinar el número apropiado de componentes que resumen adecuadamente la variabilidad de los datos originales.
b. Interpretación de las componentes principales.
c. ¿Será posible desarrollar un índice 'Tamaño de cuerpo' o 'configuración de cuerpo' basado en las 7 variables consideradas? Expliquen
d. Hacer una gráfica de las dos primeras componentes. ¿Hay outliers? Si los hay, hacer una sustitución de la matriz de covarianzas con una matriz de covarianzas estimada de manera robusta.
e. Evalúen si los datos originales son normales. Si no lo son, buscar las transformaciones que los acerquen a normalidad. Repetir el análisis con los datos transformados y probar la significancia de la varianza de las componentes principales con el resultado de Anderson.

4. Consideren la matriz de correlaciones siguiente. Los datos originales corresponden a las mediciones de 8 variables de química sanguínea de 72 pacientes en un estudio clínico. (Jolliffe, 2002). La matriz de correlaciones de las variables `rblood`, `plate`, `wblood`, `neut`, `lymph`, `bilir`, `sodium` y `potass`, en ese orden, es la siguiente:

```
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]
[1,]  1.000  0.290  0.202 -0.055 -0.105 -0.252 -0.229  0.058
[2,]  0.290  1.000  0.415  0.285 -0.376 -0.349 -0.164 -0.129
[3,]  0.202  0.415  1.000  0.419 -0.521 -0.441 -0.145 -0.076
[4,] -0.055  0.285  0.419  1.000 -0.877 -0.076  0.023 -0.131
[5,] -0.105 -0.376 -0.521 -0.877  1.000  0.206  0.034  0.151
[6,] -0.252 -0.349 -0.441 -0.076  0.206  1.000  0.192  0.077
[7,] -0.229 -0.164 -0.145  0.023  0.034  0.192  1.000  0.423
[8,]  0.058 -0.129 -0.076 -0.131  0.151  0.077  0.423  1.000
```

y las desviaciones estándar, que tienen considerales diferencias, son:

```
rblood  plate wblood   neut  lymph  bilir sodium potass
 0.371 41.253  1.935  0.077  0.071  4.037  2.732  0.297
```

- Aplicar componentes principales a la matriz de covarianzas y a la matriz de correlaciones. Explicar las diferencias.
- Basado en la observación anterior, sobre qué debería hacerse el análisis?

5. Encontrar las componentes principales de la siguiente matriz de correlación calculada de las mediciones de 7 características físicas en 3,000 convictos criminales: Las variables son: 1. largo de la cabeza, 2. ancho de la cabeza, 3. ancho del la cara, 4. longitud del dedo pulgar izquierdo, 5. longitud del antebrazo izquierdo, 6. longitud del pie izquierdo, 7. Altura.

$$\begin{pmatrix} 1 & & & & & & \\ 0.402 & 1 & & & & & \\ 0.396 & 0.618 & 1 & & & & \\ 0.301 & 0.150 & 0.321 & 1 & & & \\ 0.305 & 0.135 & 0.289 & 0.846 & 1 & & \\ 0.339 & 0.206 & 0.363 & 0.759 & 0.797 & 1 & \\ 0.340 & 0.183 & 0.345 & 0.661 & 0.800 & 0.736 & 1 \end{pmatrix}$$

Esta es la base de la antropometría de Alphonse Bertillon (1853-1914)