Estadística Aplicada III ITAM
12 de agosto de 2020 2020-II

Tarea 1. Fecha de entrega: lunes 24 de agosto de 2020.

Problemas

1. Definan las matrices $\mathbf{A} = \begin{pmatrix} 7 & 5 & 3 \\ 2 & 1 & 8 \end{pmatrix}$ y $\mathbf{B} = \begin{pmatrix} 11 & -7 & 8 \\ 12 & 0 & 9 \end{pmatrix}$. Hacer los cálculos manuales y posteriormente usar `R` o el paquete que ustedes quieran para encontrar las siguientes matrices, si es posible. Si no, decir porqué.

   a) $\mathbf{A}'$
   b) $\mathbf{A} - \mathbf{B}$
   c) $\mathbf{AB}$
   d) $\mathbf{A}'\mathbf{A}$
   e) $\mathbf{AA}'$
   f) $\mathbf{A} + \mathbf{B}$
   g) $\mathbf{A}'\mathbf{B}$
   h) $\mathbf{AB}'$
   i) $17.3\mathbf{A}$
   j) $(1/19)\mathbf{B}$

2. El archivo `T01_06.txt` en la sección de datos en Piazza contiene datos de un estudio sobre esclerosis múltiple. Dos estímulos visuales diferentes ($S1$ y $S2$) producen respuestas en el ojo izquierdo (L) y en el derecho (R) de sujetos, algunos de los cuales tenían esclerosis múltiple (MS). Las columnas registradas en el archivo son:

col 1: $x_1 =$ la edad del sujeto,

col 4: $x_2 = S1L + S1R$, la respuesta total de ambos ojos a estimulos $S1$,

col 3: $x_3 = |S1L - S1R|$, la diferencia entre respuestas de los ojos a estímulos $S1$

col 4: $x_4 = S2L + S2R$,

col 5: $x_5 = |S2L - S2R|$.

col 6: grupo al que pertenece el sujeto

   a. Hacer una gráfica de dispersión de puntos para x_2 y x_4 para el grupo de multiesclerosis. Comentar sobre la apariencia de la gráfica.
   b. Calcular $\bar{\mathbf{x}}$, $\mathbf{S}$, y $\mathbf{R}$ para cada grupo por separado.
   c. Hacer un *scatterplot* de todos los datos y comentar sobre las diferencias de cada par de variables en cada uno de los grupos.

3. En `R` hay un conjunto de datos que se llama `iris3`. Estos datos se usan frecuentemente como ejemplo para las diferentes técnicas del análisis multivariado y se considera clásico. los datos pueden ser llamados con el comando `data` y se puede leer más sobre los datos con:

```
data(iris3)
help(iris3)
```

   Los datos consisten de mediciones de 4 características numéricas de 50 flores de cada una de las tres variedades de la flor iris: *I. Setosa*, *I. versicolor* y *I. virginica*.

a. Sea $\mathbf{X}$ la matriz de los datos para *I. Setosa*. Calcular $\bar{\mathbf{x}}$, la matriz de 4×4 de sumas de cuadrados corregida por la media:

$$\mathbf{A} = \sum_{i=1}^{50} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})',$$

la matriz de covarianza muestral insesgada $\mathbf{S}_X = (1/49)\mathbf{A}$

b. Obtener los eigenvalores y eigenvectores de $\mathbf{S}$.

c. Sea $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4]$ la matriz de 4×4 de eigenvectores de $\mathbf{S}$, y sea $\mathbf{L} = diag\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$. Mostrar numéricamente que, excepto por redondeo, $\mathbf{ULU}' = \mathbf{S}$ y $\mathbf{U}'\mathbf{U} = \mathbf{UU}' = \mathbf{I}_4$.

d. Obtener matriz de gráficas de dispersión de las cuatro variables para cada variedad de iris, todo en la misma gráfica. La gráfica tiene que ser construida de tal forma que sea *clara*. Usen color y todas las opciones que tengan a la mano.

4. Usar los datos del ejercicio anterior.

   a. Crear una nueva matriz $\mathbf{Y}$ de 50×5 cuyas primeras cuatro columnas sean las mismas que $\mathbf{X}$ y cuya última columna es `Petal L.` + `Petal W.` a través de encontrar una matriz C tal que $\mathbf{Y} = \mathbf{XC}$.

   b. Calcular la matriz de covarianzas muestral $\mathbf{S}_Y$ y sus eigenvalores y eigenvectores. Noten que el eigenvalor más pequeño es 0 (excepto posiblemente por redondeo) y el eigenvector correspondiente consiste de pesos $w_1, \ldots w_5$ que definen una combinación lineal de columnas de $\mathbf{Y}$ que tiene varianza 0. Esto ilustra que los eigenvectores que corresponden a pequeños eigenvalores de $\mathbf{S}_Y$ ayudan a descubrir dependencias lineales entre variables.

   c. Mostrar numéricamente que la covarianza muestral $\mathbf{S}_Y$ se puede calcular también como $\mathbf{S}_Y = C'\mathbf{S}_X C$.

5. Elaborar un reporte con los resultados a las preguntas formuladas en el laboratorio que utiliza los datos sobre los bebés.

6. Sea $\mathbf{S}(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{a})(\mathbf{x}_i - \mathbf{a})'$ la matriz de covarianza alrededor de $\mathbf{x} = \mathbf{a}$.

   a) Mostrar que
   $$\mathbf{S}(\mathbf{a}) = \mathbf{S} + (\bar{\mathbf{x}} - \mathbf{a})(\bar{\mathbf{x}} - \mathbf{a})'$$

   b) Mostrar que
   $$\det(\mathbf{S}(\mathbf{a})) = \det(\mathbf{S}) \cdot \{1 + (\bar{\mathbf{x}} - \mathbf{a})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \mathbf{a})\}$$

   y consecuentemente $\text{mín}_{\mathbf{a}} \det(\mathbf{S}(a)) = \det(\mathbf{S})$