

Solución Examen MA34B 2006-02

1. Justificación (materia necesaria para resolver el ejercicio)

Primero veamos algunas cosas que hay que saber para resolver. Dada la matriz de datos X (ver enunciado), se tendrá

$$V = \frac{1}{j} X^t X$$

Donde j es la cantidad de observaciones, en este caso $j = 20$. V := Matriz de correlaciones de X , en caso de que X esté centralizada y normalizada. Centralizada

quiere decir $\sum_{j=1}^{20} X_{ij} = 0$ Donde X_i = variable medida. Por otro lado normalizada

quiere decir que cada valor que toma la variable medida está comparado con la varianza de dicha variable. Propiedades de V : Simétrica y semi-definida positiva, esto es $V^t = V$ y $\sum u_i = i$ donde i es la cantidad de variables (en este caso $i = 3$, notar que el análisis de componentes se hizo sobre 3 variables aún cuando se tengan observadas cuatro variables, esto quiere decir simplemente que la cuarta variable no se consideró para el análisis de componentes) y u_i son los valores propios de la matriz V . Además, al ser semi definida positiva se tendrá $u_i \geq 0$, esto es los valores propios asociados a los vectores propios de la matriz V son todos mayores o iguales que cero. Luego se tendrá la siguiente regla a tener en cuenta en el ACP: $\text{Traza} V = i$ Esto es la suma de las componentes diagonales de la matriz V es igual a la cantidad de variables en el ACP.

Nota: en el curso tradicionalmente nunca se ha pedido, ni calcular valores ni vectores propios ni calcular traspuestas ni inversas de matrices, todos esos valores son dados.

Se buscan los valores propios y vectores propios de modo de poder escribir la matriz de valores observados (X) en una base nueva que no será más que la base formada por los vectores propios de V . Luego X , en nuestro ejemplo, se podrá escribir como:

$$X = \vec{c}_1 \vec{u}_1 + \vec{c}_2 \vec{u}_2 + \vec{c}_3 \vec{u}_3$$

Donde la matriz $C = \vec{c}_1 \vec{c}_2 \vec{c}_3$ es la matriz de vectores propios, también llamados componentes principales. Y los vectores \vec{u}_i son los valores que toman las 3 variables para una observación dada. Notar que así estamos haciendo un análisis sobre las observaciones, es decir, estamos comparando observaciones contra observaciones (la distancia entre un vector de 3 dimensiones respecto a otro). También puede hacerse un análisis sobre las variables, es decir comparar el vector que contiene las 20 observaciones respecto al vector que contiene las 20 observaciones para una de las otras dos variables. En este caso el enunciado dice que se hizo un análisis sobre las variables. Luego X se podrá escribir como:

$$X = \vec{d}_1 \vec{u}_1 + \vec{d}_2 \vec{u}_2 + \vec{d}_3 \vec{u}_3 \text{ Donde } d_i \text{ corresponde a los vectores propios de la matriz } V' = \frac{1}{j} X X^t. \text{ Luego se puede demostrar la siguiente relación importante}$$

para el ACP (ver apuntes profesora Nancy Lacourly): $C_i = \sqrt{l_i} d_i$. A estas alturas deberían notar un cambio de notación respecto a las u respecto al apunte de la profesora Nancy, se usó esta notación para reforzar el análisis de los datos obtenidos por sobre el álgebra. Por último, cada valor propio asociado a cada componente principal mide la varianza de dicha componente y $\text{Traza} V$ mide la

varianza total en la representación. Así la calidad de la representación de C_i viene dada por $100 \frac{l_i}{\text{Traza}V}$ donde l_1 corresponde al vector propio asociado a C_i . Sabiendo estas cosas se está listo para resolver el ejercicio, veamos.

Solución:

Viendo la tabla de variabilidad representada por cada componente se tiene que la calidad es del 100% usando las dos componentes. Se muestran dos componentes pues ellas dos solas representan el 100% de variabilidad o varianza. El tercer valor propio = 0 ($1.6671 + 1.3329 = 3 = i$). Los ejes que se muestran son los tres ejes iniciales que se tenían para representar los datos representados ahora en la nueva base, esto permite clarificar la relación entre ellos y las observaciones.

2. Justificación: La distancia del origen a la proyección de X_j mide la calidad de la representación. Esto es $0 \leq d(\vec{X_j}) \leq 1$ dicha distancia se calcula mediante:

$$d(\vec{X_j}) = (\text{cor}(X_j, C_1), \text{cor}(X_j, C_2))$$

Como sólo tenemos d_j y los valores propios usamos el hecho de que:

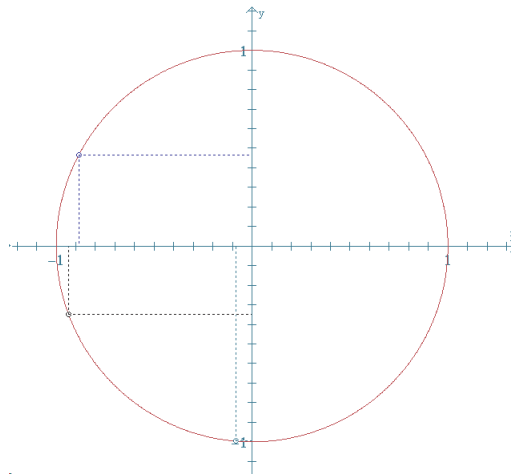
$$C_j = \sqrt{l_j} d_j$$

Solución:

$$d(X_1) = (\sqrt{1.6671} * -0.685, \sqrt{1.3329} * 0.403) = (-0.885, 0.465) \Rightarrow \|d(X_1)\| = 0.999$$

$$d(X_2) = (\sqrt{1.6671} * -0.065, \sqrt{1.3329} * -0.863) = (-0.0839, -0.996) \Rightarrow \|d(X_2)\| = 0.999$$

$$d(X_3) = (\sqrt{1.6671} * -0.725, \sqrt{1.3329} * -0.304) = (-0.936, -0.35) \Rightarrow \|d(X_3)\| = 0.999$$



3.- Justificación: Para realizar la regresión se necesita que las variables no sean dependientes.

Solución:

$Cov(x_i, x_j) \neq 0 \forall i \neq j$ Pero no se da esta condición viendo la tabla de correlaciones, luego las variables son dependientes y no se podrá escribir x_4 en función de las otras tres variables (se puede ver el gráfico también).

4.- Primero hay que normalizar y estandarizar los datos. Para ello se resta el valor de la variable respecto al promedio y se compara con la varianza de dicha variable. Con ello se tiene:

$$\frac{x'_1 - \bar{x}_1}{V(x_1)} = -0.03$$

$$\frac{x'_2 - \bar{x}_2}{V(x_2)} = 0.1$$

$$\frac{x'_3 - \bar{x}_3}{V(x_3)} = -3.7$$

Luego para ver este nuevo punto en el gráfico de componentes principales se debe hacer:

$$(-0.03, 0.1, -3.7) * \begin{bmatrix} -0.685 \\ -0.065 \\ -0.725 \end{bmatrix} = 2.7$$

$$(-0.03, 0.1, -3.7) * \begin{bmatrix} 0.403 \\ -0.863 \\ -0.304 \end{bmatrix} = 1.02$$

5.- Tenemos las nuevas cp's de una nueva variable. Para ubicar en el círculo se usa la distancia, luego

$$d(X_4) = (\sqrt{1.6671} * 0.15, \sqrt{1.3329} * 0.44) = (0.19, 0.51) \Rightarrow \|d(X_3)\| = 0.544$$