EXAMEN PARCIAL ANÁLISIS MULTIVARIANTE (PRIMERA PARTE)
Máster en Estadística e Investigación Operativa (MESIO UPC-UB)
Jueves 21 de Marzo de 2019, Aula 004, 18.00-19.30h
Nombre y Apellidos: .....

---

1. (10p) **Análisis de componentes principales y el álgebra lineal.** Sea $\mathbf{X}$ la matriz $n \times p$ con las variables cuantitativas originales. Sea $\mathbf{S}$ su matriz de covarianzas muestrales.

   (a) (1p) Indica, con una expresión matricial, cómo se puede obtener la matriz de datos centrados $\mathbf{X}_c$ a partir de la matriz de datos originales. Cabe definir adecuadamente las matrices o vectores auxiliares que se necesitan. .....

   .....

   .....

   .....

   (b) (1p) Describe la descomposición espectral de $\mathbf{S}$, indicando posibles restricciones a las cuales las matrices obtenidas podrian estar sometidos. .....

   .....

   .....

   .....

   (c) (1p) Indica, con un expresión matricial, cómo se calculan los componentes principales a partir del los resultados previos. .....

   .....

   .....

   .....

   (d) (2p) Deriva, mediante álgebra lineal, la matriz de covarianzas de los componentes principales. .....

   .....

   .....

   .....

   (e) (1p) Demuestra que el ACP preserva la varianza total, tal que la varianza total de las variables originales equivale a la varianza total de todos los componentes principales. .....

   .....

   .....

   .....

   (f) (1p) Usa los resultados obtenidos para indicar cómo construir un biplot bi-dimensional de la matriz de datos centrados $\mathbf{X}_c$. Indica claramente las matrices que se usan para graficar las coordenadas de filas y columnas en el biplot. .....

   .....

   .....

   .....

   (g) (2p) Los biplots se pueden escalar de varias maneras para enfatizar mejor ciertos aspectos de los datos. Describe las expresiones matriciales para un biplot alternativo, re-escalando las coordenadas del biplot anterior. Proporciona motivos por el re-escalamiento propuesto. .....

   .....

   .....

   .....

   (h) (1p) Afecta el re-escalamiento del biplot a la bondad de ajuste? Argumenta la respuesta. .....

   .....

   .....

   .....

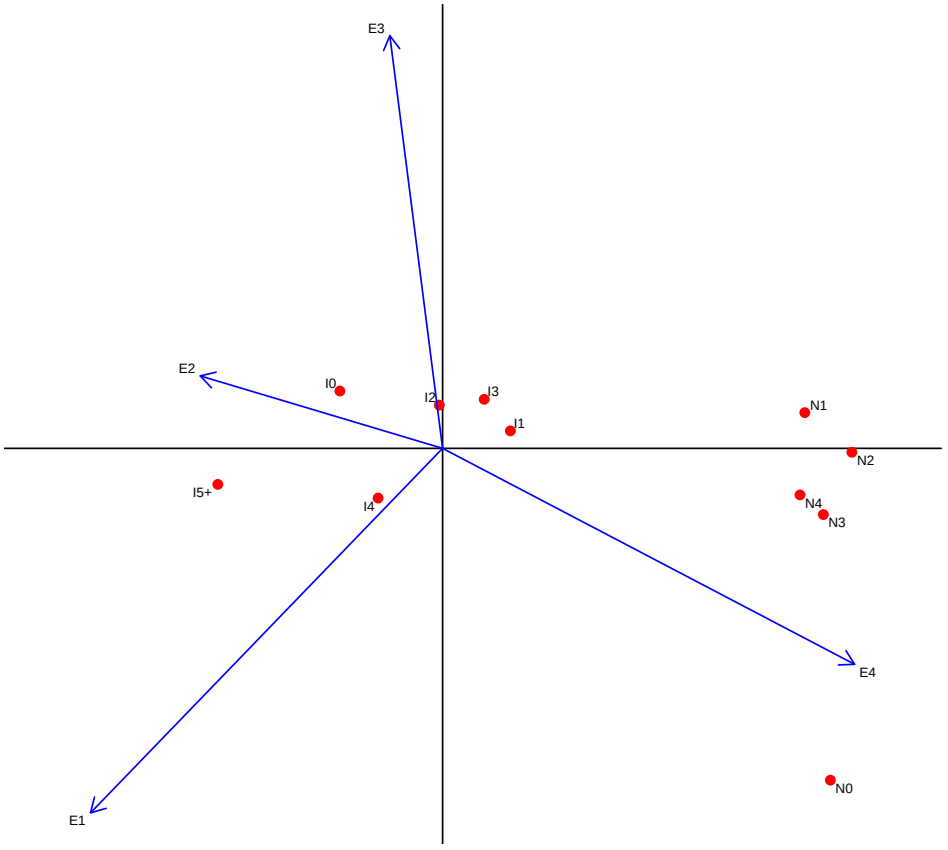
2. (10p) **Análisis de Correspondencias**. Una muestra de 1473 mujeres de Indonesia ha participado en una encuesta. Se recogieron muchas variables, entre otras en nivel de estudios (E1, E2, E3, E4, de bajo a alto), su religión (Islam o no: I o N) y su número de hijos (categorizado como 0, 1, 2, 3, 4 o 5+). Las dos últimas variables fueron usadas en una codificación interactiva. La tabla de datos resultante se ha analizado con una prueba de chi-cuadrado, obteniendo un estadístico chi-cuadrado de 149.72. La misma tabla también se analizó mediante un análisis de correspondencias, y los primeros dos valores propios obtenidos fueron 0.091 y 0.008 respectivamente. El biplot de los perfiles fila se incluye a continuación. Algunos diagnósticos calculados en el análisis de correspondencias también se muestran a continuación.

Rows:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	I0	57	980	22	-172	747	19	96	233	65
2	I1	163	596	36	114	559	23	29	37	17
3	I2	163	933	9	-5	5	0	73	928	107
4	I3	142	850	19	70	356	8	82	494	119
5	I4	114	735	28	-108	461	15	-83	273	98
6	I5	232	1000	325	-377	975	362	-60	25	105
7	N0	10	960	77	651	554	49	-557	406	400
8	N1	29	924	111	608	915	116	60	9	13
9	N2	29	929	139	687	929	148	-7	0	0
10	N3	38	997	155	639	968	171	-111	29	58
11	N4	23	996	81	600	979	91	-78	17	17

Columns:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	E1	104	976	342	-552	891	349	-170	85	374
2	E2	227	837	174	-257	830	165	23	7	15
3	E3	279	797	55	-51	124	8	118	673	480
4	E4	389	1000	428	334	976	478	-52	24	131



Contesta las preguntas a continuación:

(a) (1p) Se puede considerar que las filas y las columnas de esta tabla sean independientes? Argumenta la respuesta. .....
.....
.....
.....

(b) (1p) Cúal es el valor esperado del estadístico de chi-cuadrado si la hipótesis de independencia de filas y columnas fuese cierta? .....
.....
.....
.....

(c) (1p) Cúal és la inercia máxima alcanzable para una tabla cruzada de estas dimensiones? .....
.....
.....

(d) (1p) Qué porcentaje de la inercia total de la tabla cruzada queda representada en un biplot bi-dimensional usando las primeras dos dimensiones del análisis? .....
.....
.....

(e) (1p) A la vista de los resultados obtenidos, cómo se podrian caracterizar las mujeres con el nivel educativo mas alto? .....
.....
.....
.....
.....

(f) (1p) Cómo caracterizarias a las mujeres con cinco o mas hijos? .....
.....
.....
.....
.....

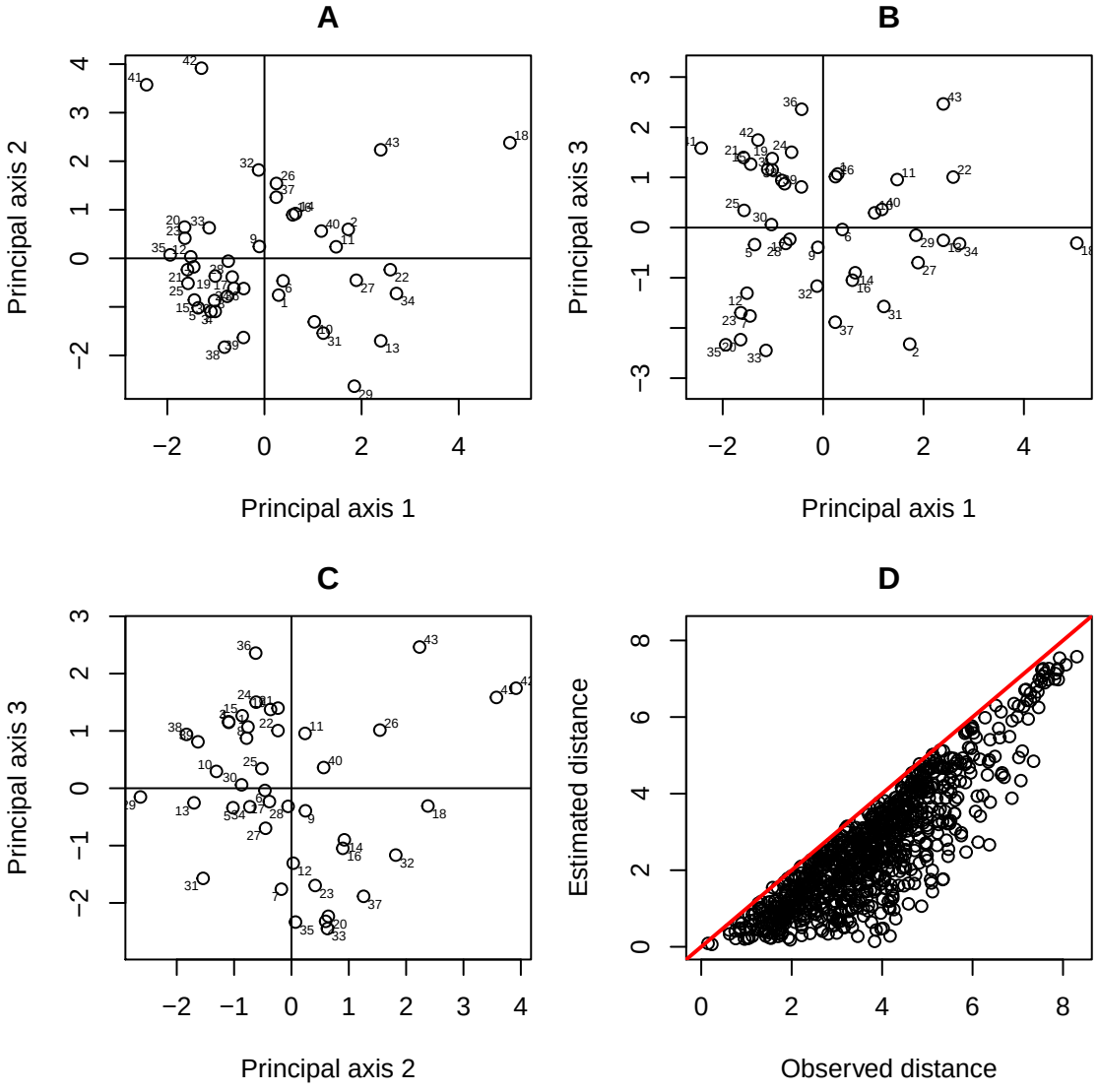
(g) (1p) Cúantas dimensiones se necesitan para representar adecuadamente estos datos? .....
.....
.....
.....
.....

(h) (2p) Cúal es la categoría peor aproximada en el biplot? Se puede mejorar la representación de esta categoria de algún modo? .....
.....
.....
.....
.....

(i) (1p) Coinciden las dimensiones de la tabla cruzada analizada con las esperadas, a la vista de los niveles de las distintas variables categóricas? Argumenta la respuesta. .....
.....
.....
.....
.....

3. (10p) **Escalamiento Multidimensional**. Un investigador en márketing ha recogido datos sobre cereales de desayuno. Las variables registrados son la cantidad de Calorias, Fibra y varios componentes de los cereales (Proteina, Grasa, Sodio, Carbohidratos, Azúcar y Potasio). El investigador quiere estudiar y visualizar la similitud de los distintos cereales para fines de márketing, y decide usar MDS para tal propósito, calculando la matriz de distancias Euclideas de los datos estandarizados. En el análisis se han obtenido los valores propios a continuación. Gráficas de las primeras dimensiones en el análisis se muestran en las figuras A, B y C. Un gráfico diagnóstico de la solución bi-dimensional se muestra en la figura D.

```
 [1]  1.069979e+02  7.790000e+01  7.429086e+01  3.646790e+01  2.088663e+01
 [6]  1.501200e+01  2.541141e+00  1.903558e+00  1.171489e-14  8.724175e-15
[11]  7.082768e-15  7.062333e-15  6.513312e-15  6.244108e-15  6.089274e-15
[16]  5.987974e-15  5.451720e-15  3.988872e-15  2.789861e-15  2.630448e-15
[21]  2.489370e-15  2.435585e-15  2.131313e-15  1.894922e-15  1.212605e-15
[26]  1.087264e-15  1.044636e-15  7.257006e-16  2.628334e-16 -2.611771e-16
[31] -6.457832e-16 -7.341904e-16 -9.142919e-16 -1.788448e-15 -2.949856e-15
[36] -3.736897e-15 -4.552350e-15 -5.569457e-15 -6.227026e-15 -7.636676e-15
[41] -1.029463e-14 -1.044996e-14 -2.006448e-14
```



(a) (1p) Proporciona, en notación matricial, la expresión de la distancia Euclídea d_{ij} entre dos filas de la matriz de datos originales $\mathbf{X}$.

(b) (1p) Le estandarización utilizada por el investigador implica el uso de distancias ponderadas. Proporciona, en notación matricial, una expresión para la distancia Euclídea ponderada d_{ij} entre dos filas de la matriz de datos orginales $\mathbf{X}$. Cabe definir adecuadamente los vectores o matrices de pesos que se necesitan.

(c) (2p) Calcula las bondades de los ajustes para las representaciones de la matriz de distancias en las Figuras A, B y C arriba. Cúal de las tres figuras proporciona la mejor aproximación a la matriz de distancias?

(d) (1p) Cúantas dimensiones hay que extraer del análisis parar poder representar la matriz de distancias de manera exacta, si tener error en la aproximación?

(e) (1p) A la vista de los resultados del análisis, que cereales destacan por ser distintos de la resta?

(f) (1p) Para entender mejor las dimensiones extraïdos por MDS, el investigador calcula correlaciones entre estas dimensiones y las variables originales, obteniendo los resultados:

	Axis-1	Axis-2	Axis-3
Calories	0.18	0.89	-0.18
Protein	0.67	-0.31	-0.34
Fat	0.50	0.41	0.22
Sodium	0.04	0.37	-0.79
Fiber	0.89	-0.18	-0.09
Carbohydrates	-0.38	0.15	-0.84
Sugar	0.06	0.77	0.48
Potassium	0.93	0.00	-0.10

En que sentido difieren los cereales previamente identificados de la resta de cereales?

(g) (2p) Cúales son las conclusiones que uno puede sacar de la figura diagnóstica D?

(h) (1p) Cúal es el cociente de aspecto de las figuras obtenidos por el investigador? Razona la importancia de este cociente.