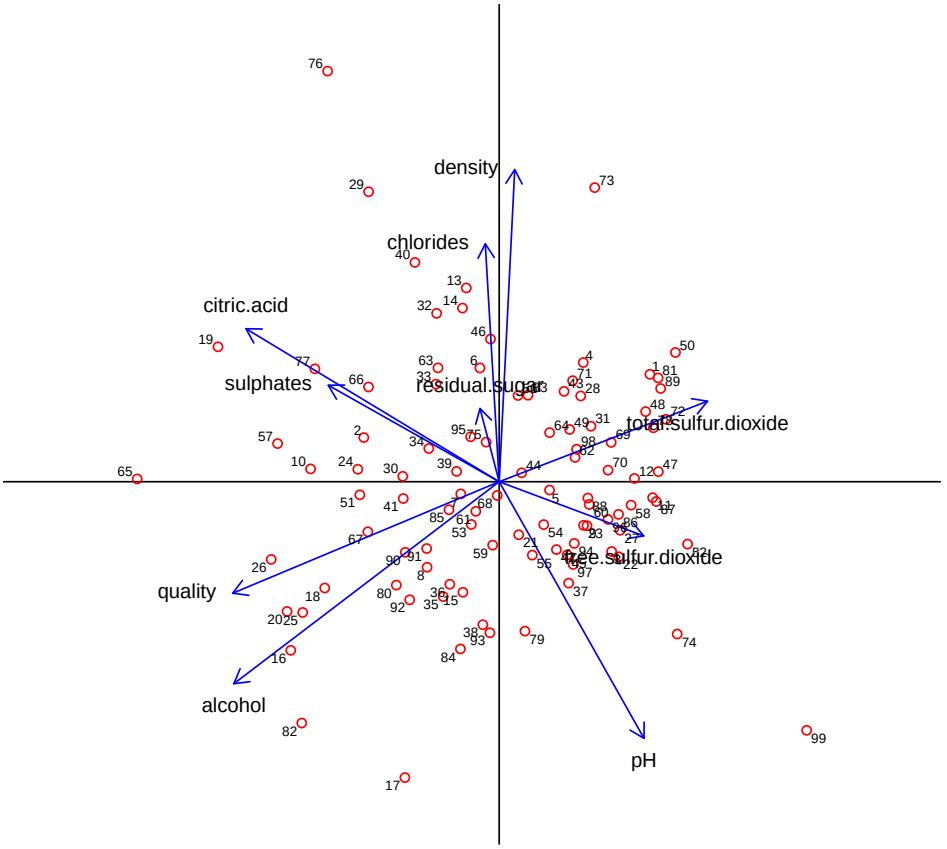
EXAMEN ANÁLISIS MULTIVARIANTE (EXAMEN PARCIAL PRIMERA PARTE)
Máster en Estadística e Investigación Operativa (MESIO UPC-UB)
Martes 28 de mayo de 2019, Aula PC2, 18.00-19.30h
Nombre y apellidos: .....

---

1. (10p) **Análisis de componentes principales**. Se han registrado varias medidas fisicoquímicas (citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol) y una evaluación de la calidad por parte de un catador (quality) de 99 vinos tintos. La matriz de datos se ha analizado mediante un análisis de componentes principales, y se muestra un biplot de los dos primeros componentes principales no estandarizados en la figura más abajo. Los valores propios obtenidos en el análisis son: $\lambda_1 = 2.32, \lambda_2 = 1.98, \lambda_3 = 1.73, \lambda_4 = 1.13, \lambda_5 = 0.94, \lambda_6 = 0.61, \lambda_7 = 0.44, \lambda_8 = 0.35, \lambda_9 = 0.33$ y $\lambda_{10} = 0.17$.



   (a) (1p) Cuál de las variables guarda, según el biplot, mas relación lineal con la calidad del vino? .....
   .....
   .....

   (b) (1p) Cuál de las variables guarda, según el biplot, poca o bien ninguna relación lineal con la calidad del vino?
   .....
   .....

   (c) (1p) Cuál de los vinos tiene, según el biplot, el pH más bajo? .....
   .....
   .....

   (d) (1p) Es este análisis un análisis basado en covarianzas o bien en correlaciones? Argumenta la respuesta. .....
   .....
   .....
   .....

   (e) (1p) Cuál es la matriz de datos que se aproxima en el biplot en la figura arriba? .....
   .....
   .....

(f) (1p) Calcula la bondad del ajuste de la matriz de datos representada en el biplot de dos dimensiones. .....

(g) (1p) Qué porcentaje de la variabilidad de la matriz de datos se hubiera explicado por un biplot de los últimos dos componentes principales? .....

(h) (1p) Usa el biplot para caracterizar el vino con número 65. .....

(i) (1p) Intenta interpretar el primer componente principal. .....

(j) (1p) Hubiera sido útil dibujar un círculo unitario dentro este biplott? Argumenta la respuesta. .....

2. (5p) **Análisis multivariante y el álgebra lineal.** Consideramos una matriz $n \times p$ $\mathbf{X}$ con variables cuantitativas.

(a) (1p) Explica, con una expresión matricial, cómo calcularía el vector de medias $\mathbf{m}$ a partir de la matriz de datos $\mathbf{X}$. Procura definir $\mathbf{m}$ como vector columna. .....

(b) (1p) Proporciona una expresión matricial que genera una matriz de dimensiones $n \times p$ a partir de $\mathbf{m}$ de modo que $\mathbf{m}$ se repita en todas las filas de esta matriz. .....

(c) (1p) Obtén una expresión para la matriz de datos centrados $\mathbf{X}_c$, restando el resultado penúltimo de la matriz $\mathbf{X}$. Demuestra cómo obtener, mediante manipulación algebraica, la matriz de centrado $\mathbf{H}$, con la cual se transforma $\mathbf{X}$ a $\mathbf{X}_c$. .....

(d) (1p) Demuestra que $\mathbf{HH} = \mathbf{H}$. .....

(e) (1p) Cuál es el resultado de centrar la matriz de datos centrados $\mathbf{X}_c$? Demuéstralo con álgebra. .....

3. (5p) **Descomposición en valores singulares**. Consideramos una matriz de datos cuantitativos $\mathbf{X}$ con dimensiones $n \times p$.

(a) (2p) Describe la descomposición en valores singulares de la matriz $\mathbf{X}$, indicando las propiedades de las matrices obtenidas en la descomposición. .....

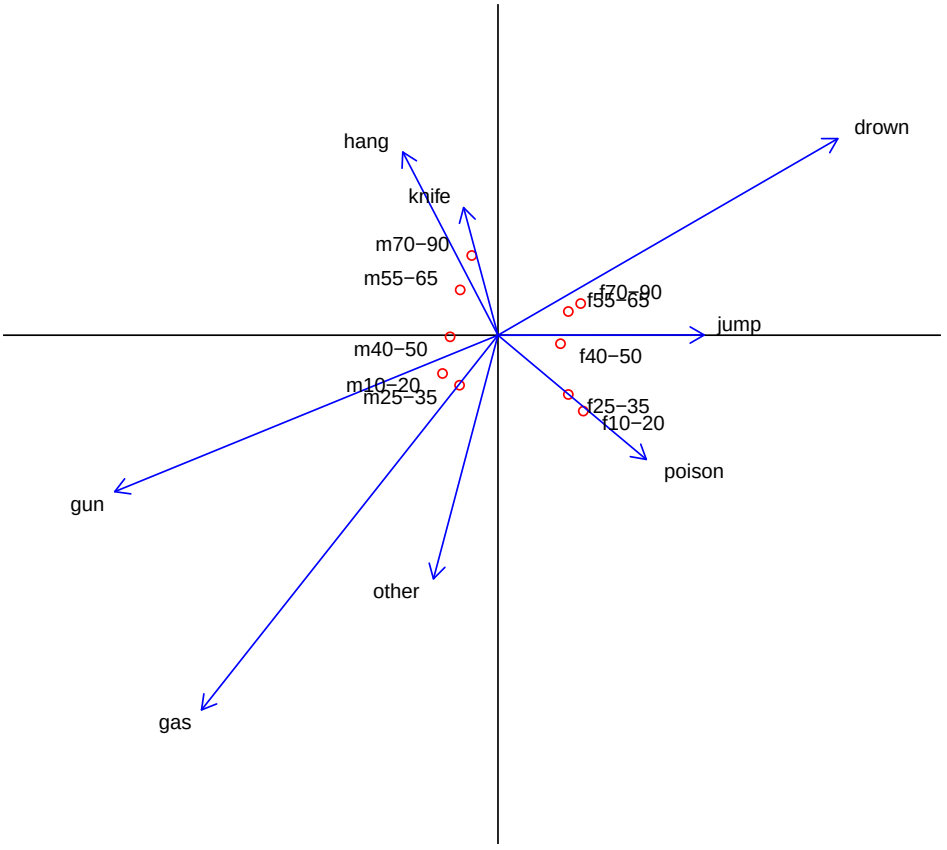
(b) (1p) Indica cómo se obtiene una aproximación a la matriz $\mathbf{X}$ con rango tres, i que sea aproximación óptima en el sentido de los mínimos cuadrados. .....
.....
.....
.....

(c) (2p) Enseña cómo la suma de cuadrados de los elementos de $\mathbf{X}$ se relaciona con los valores singulares. .....
.....
.....
.....

4. (10p) **Análisis de correspondencias**. En un estudio de 53182 suicidios en Alemania, se han registrado las variables *sexo* (m = male (varón) o f = female (mujer)), *método* (poison, gas, hang, drown, gun, knife, jump, o other) y *intervalo de edad* (10-20, 25-35, 40-50, 55-65, 70-90). La tabla de contingencia de los datos, obtenida mediante la codificación interactiva de sexo e intervalo de edad, se muestra a continuación.

	poison	gas	hang	drown	gun	knife	jump	other
m10-20	1160	335	1524	67	512	47	189	464
m25-35	2823	883	2751	213	852	139	366	775
m40-50	2465	625	3936	247	875	183	244	534
m55-65	1531	201	3581	207	477	154	273	294
m70-90	938	45	2948	212	229	105	268	147
f10-20	921	40	212	30	25	11	131	100
f25-35	1672	113	575	139	64	41	276	263
f40-50	2224	91	1481	354	52	80	327	305
f55-65	2283	45	2014	679	29	103	388	296
f70-90	1548	29	1355	501	3	74	383	106

Mediante un análisis de correspondencias de esta tabla se obtuvieron los valores propios a continuación: $\lambda_1 = 0.0962, \lambda_2 = 0.0597, \lambda_3 = 0.0082, \lambda_4 = 0.0022, \lambda_5 = 0.0014, \lambda_6 = 0.0006, \lambda_7 = 0.0001$. El biplot de los perfiles fila calculados a partir de esta tabla se muestra en la figura abajo.



(a) (1p) Calcula el estadístico chi-cuadrado de Pearson para un test de independencia de filas y columnas de la tabla de contingencia.

(b) (1p) Cuál es la distribución del estadístico de Pearson, suponiendo que la hipótesis nula de independencia se cumple? Cabe especificar exactamente el/los valor(es) de los parámetros de esta distribución.

(c) (1p) Existe asociación significativa entre las filas y las columnas de esta tabla? Argumenta la respuesta.

(d) (1p) Cuál es tu interpretación de la primera dimensión obtenida en el análisis de correspondencias?

(e) (1p) Cuál es tu interpretación de la segunda dimensión obtenida en el análisis de correspondencias?

(f) (1p) Cuál es la categoría sexo-edad con, según el biplot, menos uso del método "gun" ?

(g) (1p) Que porcentaje de la inercia total de la tabla queda explicado por el biplot de dos dimensiones?

(h) (1p) Las tres variables categóricas bajo estudio también se podrían analizar mediante el análisis de correspondencias múltiple, usando la matriz de variables indicadoras. ¿Cuál hubiera sido la inercia total de la tabla de datos en este caso? Argumenta la respuesta.

(i) (1p) Calcula el vector de pesos columna de la tabla.

(j) (1p) Si se interpreta el vector de pesos columna como un perfil, y se proyecta este perfil sobre el biplot, que categoría sexo-edad estaría mas cerca a esta proyección?