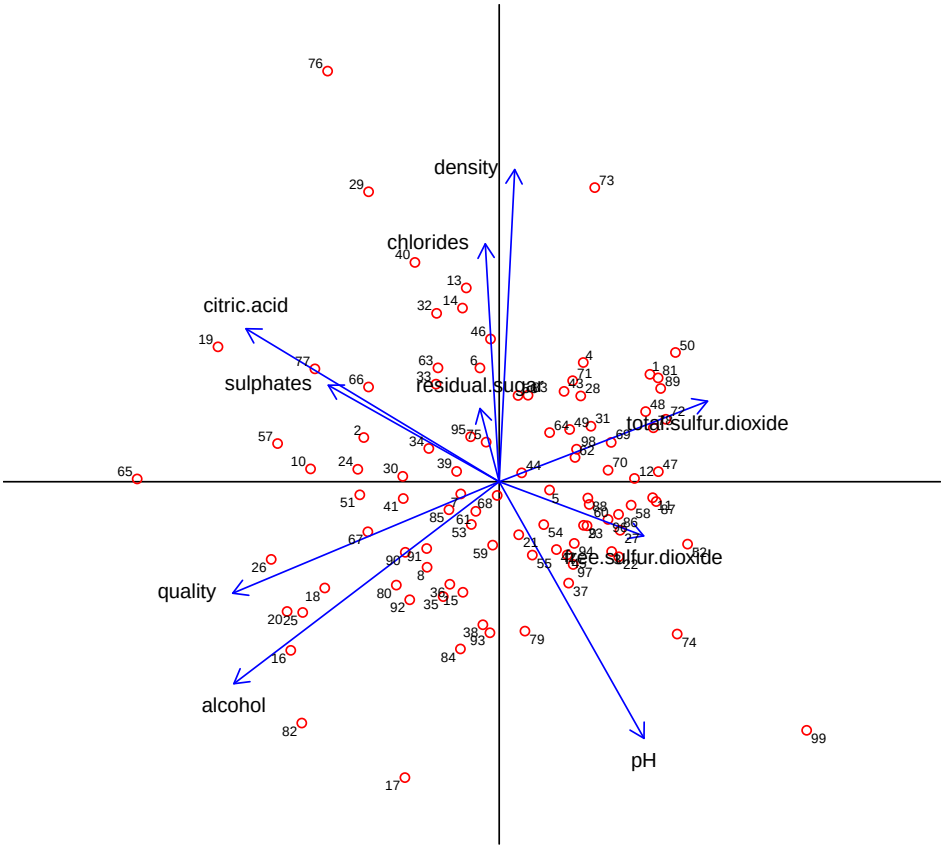
RESIT EXAM MULTIVARIATE ANALYSIS (FIRST PARTIAL)
Master in Statistics and Operations Research (MESIO UPC-UB)
Tuesday 28^{th} of May 2019, Room PC2, 18.00-19.30h
Name and Surname: .....

---

1. (10p) **Principal component analysis**. Physicochemical measurements (citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol) and a taster's quality assessment (quality) have been registered for 99 red wines. The data matrix is analysed by a principal component analysis, and a biplot of the non-standardized first two principal components is shown in the figure below. The eigenvalues obtained in the analysis are: $\lambda_1 = 2.32, \lambda_2 = 1.98, \lambda_3 = 1.73, \lambda_4 = 1.13, \lambda_5 = 0.94, \lambda_6 = 0.61, \lambda_7 = 0.44, \lambda_8 = 0.35, \lambda_9 = 0.33$ and $\lambda_{10} = 0.17$.



(a) (1p) Which variable is, according to the biplot, most tightly correlated with the quality of the wines? .....
.....
.....

(b) (1p) Which variable has, according to the biplot, no linear relationship with wine quality? .....
.....
.....

(c) (1p) Which wine has, according to the biplot, the lowest pH? .....
.....
.....

(d) (1p) Is this a covariance-based or a correlation based analysis? Argue your answer. .....
.....
.....
.....

(e) (1p) Which data matrix is approximated by the biplot in the figure above? .....
.....
.....

(f) (1p) Calculate the goodness-of-fit of the data matrix represented in the two-dimensional biplot.

(g) (1p) How much of the total variance of the data matrix would have been accounted for by a biplot of the last two principal components?

(h) (1p) Use the biplot to give a characterization of wine number 65.

(i) (1p) Try to give an interpretation of the first principal component.

(j) (1p) Would it be useful to draw a unit circle in this biplot? Argue your answer.

2. (5p) **Multivariate analysis and matrix algebra.** Consider a $n \times p$ matrix $\mathbf{X}$ of quantitative variables.

(a) (1p) Show, with a matrix expression, how you would compute the mean vector $\mathbf{m}$ corresponding to this data matrix, which contains the means of the p columns of $\mathbf{X}$. Take care to define $\mathbf{m}$ as a column vector.

(b) (1p) Give a matrix expression that "blows up" $\mathbf{m}$ to a matrix of dimensions $n \times p$, such that $\mathbf{m}$ is repeated in each row of this matrix.

(c) (1p) Obtain an expression for the centred data matrix $\mathbf{X}_c$ by subtracting your previous result from $\mathbf{X}$. Manipulate algebraically, such as to obtain the so-called centring matrix $\mathbf{H}$, that transforms $\mathbf{X}$ into $\mathbf{X}_c$.

(d) (1p) Show that $\mathbf{H}$ satisfies $\mathbf{HH} = \mathbf{H}$.

(e) (1p) What is the result of centring the centred data matrix $\mathbf{X}_c$? Show this algebraically.

3. (5p) **Singular value decomposition**. Consider a quantitative $n \times p$ data matrix $\mathbf{X}$.

(a) (2p) Describe the singular value decomposition of $\mathbf{X}$, indicating the properties of the matrices obtained in the decomposition.

(b) (1p) Indicate how you would obtain an approximation to $\mathbf{X}$ that has rank three, and that is optimal in the least-squares sense. .....

.....

.....

.....

(c) (2p) Show how the total sum-of-squares of $\mathbf{X}$ relates to the singular values.....

.....

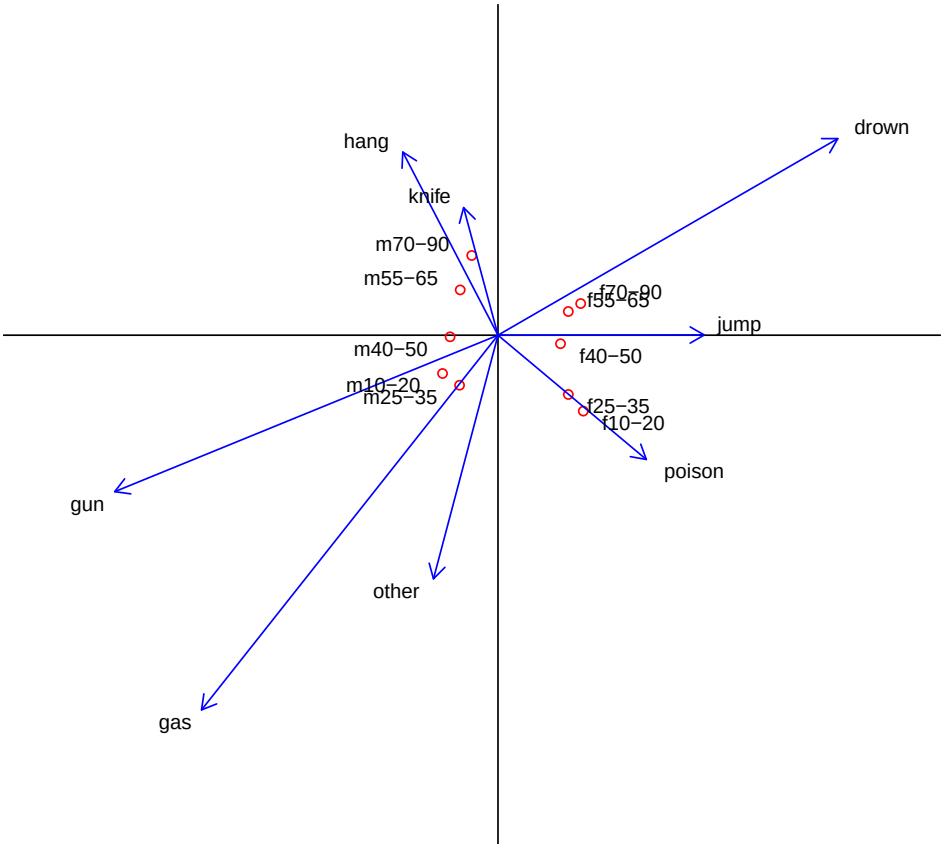
.....

.....

4. (10p) **Correspondence analysis**. In a study on 53182 suicides in Germany, the variables *sex* (m = male or f = female), *method* (poison, gas, hang, drown, gun, knife, jump, or other) and *age interval* (10-20, 25-35, 40-50, 55-65, 70-90) were registered. A contingency table of the data, obtained by the interactive coding of sex and age interval, is given below.

	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

This table was analysed by correspondence analysis, and the following eigenvalues were obtained: $\lambda_1 = 0.0962, \lambda_2 = 0.0597, \lambda_3 = 0.0082, \lambda_4 = 0.0022, \lambda_5 = 0.0014, \lambda_6 = 0.0006, \lambda_7 = 0.0001$. A biplot of the row profiles calculated from this table is shown below.



(a) (1p) Calculate Pearson's chi-square statistic for a test of independence between rows and columns of the contingency table.

(b) (1p) What distribution does Pearson's statistic have, if the null hypothesis of independence is true? Specify exactly the values of the parameter(s) this distribution may have.

(c) (1p) Do you think there is statistically significant association between the rows and columns of this table? Argue your answer.

(d) (1p) Give your interpretation of the first dimension obtained by correspondence analysis.

(e) (1p) Give your interpretation of the second dimension obtained by correspondence analysis.

(f) (1p) For which age-sex category is, according to the biplot, "gun" the least used method?

(g) (1p) How much of the total inertia of the table is accounted for by the two-dimensional biplot?

(h) (1p) The three categorical variables under study might also have been analysed by multiple correspondence analysis, using the indicator matrix. What would have been the total inertia of the data table in that case? Argue your answer.

(i) (1p) Calculate the vector of column masses of the table.

(j) (1p) If the vector of column masses is interpreted as a profile, and projected onto the biplot, which age-sex category would be closest to it?