Exercise set #1. MVA: Basic multivariate calculations in R

Laura Santulario Verdú
Eudald Romo Grau

February 10, 2018

1 Calculate the mean vector of the 6 numerical values.

Birth	Death	Infant	LifeEM	LifeEF	GNP
29.2299	10.83608	54.90103	61.48557	66.15113	5380

Figure 1: Mean of each of the numerical values of the provided data.

2 Compute the centered data matrix.

	Birth	Death	Infant	LifeEM	LifeEF	GNP
Albania	-4.529897	-5.1360825	-24.10103	8.114433	9.348866	-4780
Bulgaria	-16.729897	1.0639175	-40.50103	6.814433	8.548866	-3130
Czechoslovakia	-15.829897	0.8639175	-43.60103	10.314433	11.548866	-2400
Former_E._Germany	-17.229897	1.5639175	-47.30103	8.314433	9.748866	-5479
Hungary	-17.629897	2.5639175	-40.10103	3.914433	7.648866	-2600
Poland	-14.929897	-0.6360825	-38.90103	5.714433	9.548866	-3690

Figure 2: Mean of each of the numerical values of the provided data.

3 Compute the variance-covariance matrix of the data.

	Birth	Death	Infant	LifeEM	LifeEF	GNP
Birth	183.51295	30.61006	534.7950	-112.87675	-133.34521	-64112.21
Death	30.61006	21.59921	139.9259	-32.77874	-35.44691	-11324.63
Infant	534.79497	139.92590	2115.3178	-414.32926	-483.56687	-209273.69
LifeEM	-112.87675	-32.77874	-414.3293	92.46687	103.98164	46881.30
LifeEF	-133.34521	-35.44691	-483.5669	103.98164	121.11862	54232.77
GNP	-64112.20833	-11324.62604	-209273.6875	46881.30417	54232.76771	63413346.77

Figure 3: Mean of each of the numerical values of the provided data.

4 Compute the correlation matrix. Which variables show strong linear association?

	Birth	Death	Infant	LifeEM	LifeEF	GNP
Birth	1.0000000	0.4861966	0.8583534	-0.8665189	-0.8944140	-0.5943155
Death	0.4861966	1.0000000	0.6546232	-0.7334666	-0.6930331	-0.3059952
Infant	0.8583534	0.6546232	1.0000000	-0.9368384	-0.9553516	-0.5713951
LifeEM	-0.8665189	-0.7334666	-0.9368384	1.0000000	0.9825578	0.6122323
LifeEF	-0.8944140	-0.6930331	-0.9553516	0.9825578	1.0000000	0.6188223
GNP	-0.5943155	-0.3059952	-0.5713951	0.6122323	0.6188223	1.0000000

Figure 4: Mean of each of the numerical values of the provided data.

The pair of variables with highest direct linear association is LifeEM-LifeEf and, on a lesser degree, Birth-Infant. The stronger inverse linear associations are between Birth/Infant and LifeEM/LifeEF (Birth-LifeEM, Birth-LifeEF, Infant-LifeEM, Infant-LifeEF).

5 Compute the matrix of standardized variables.

	Birth	Death	Infant	LifeEM	LifeEF	GNP
1	-0.3343913	-1.1051293	-0.5240199	0.8438497	0.8494806	-0.6002575
2	-1.2349799	0.2289228	-0.8805992	0.7086579	0.7767890	-0.3930556
3	-1.1685431	0.1858889	-0.9480013	1.0726358	1.0493826	-0.3013845
4	-1.2718893	0.3365077	-1.0284491	0.8646484	0.8858264	-0.6880357
5	-1.3014168	0.5516774	-0.8719021	0.4070763	0.6950109	-0.3264999
6	-1.1021062	-0.1368657	-0.8458109	0.5942649	0.8676535	-0.4633787

Figure 5: Mean of each of the numerical values of the provided data.

6 Compute the variance-covariance matrix of standardized variables. What do you observe?

The Variance-Covariance matrix of standardized data corresponds with the correlation matrix, as it can be seen in Fig. 4 and 6

	Birth	Death	Infant	LifeEM	LifeEF	GNP
Birth	1.0000000	0.4861966	0.8583534	-0.8665189	-0.8944140	-0.5943155
Death	0.4861966	1.0000000	0.6546232	-0.7334666	-0.6930331	-0.3059952
Infant	0.8583534	0.6546232	1.0000000	-0.9368384	-0.9553516	-0.5713951
LifeEM	-0.8665189	-0.7334666	-0.9368384	1.0000000	0.9825578	0.6122323
LifeEF	-0.8944140	-0.6930331	-0.9553516	0.9825578	1.0000000	0.6188223
GNP	-0.5943155	-0.3059952	-0.5713951	0.6122323	0.6188223	1.0000000

Figure 6: Mean of each of the numerical values of the provided data.

7 Compute the matrix of Euclidean distances between the individuals, using the original data matrix. Which variable(s) contribute(s) most to the Euclidean distances? What could you do to reduce its/their influence?

The variables with highest variance contribute the most to the Euclidean distance. As can be seen in Fig 7, GNP has the highest standard deviation (hence highest variance) and it can be seen

Birth	Death	Infant	LifeEM	LifeEF	GNP
13.5467	4.647495	45.99258	9.61597	11.00539	7963.25

Figure 7: Standard deviation of each of the numerical values of the provided data.

	Albania	Bulgaria	Czechoslovakia	Former_E._Germany	Hungary	Poland
Albania	0.0000	1650.1390	2380.1163	699.5324	2180.1164	1090.1620
Bulgaria	1650.1390	0.0000	730.0217	2349.0107	530.0117	560.0097
Czechoslovakia	2380.1163	730.0217	0.0000	3079.0038	200.1863	1290.0195
Former_E._Germany	699.5324	2349.0107	3079.0038	0.0000	2879.0133	1789.0245
Hungary	2180.1164	530.0117	200.1863	2879.0133	0.0000	1090.0118
Poland	1090.1620	560.0097	1290.0195	1789.0245	1090.0118	0.0000

Figure 8: Mean of each of the numerical values of the provided data.

8 Compute the matrix of Mahalanobis distances between the individuals.

	Albania	Bulgaria	Czechoslovakia	Former_E._Germany	Hungary	Poland
Albania	0.000000	2.7571232	2.477508	3.0079587	3.276377	2.292311
Bulgaria	2.757123	0.0000000	1.325117	0.8044926	1.249750	1.529519
Czechoslovakia	2.477508	1.3251171	0.000000	1.0734243	2.321823	2.226411
Former_E._Germany	3.007959	0.8044926	1.073424	0.0000000	1.800156	2.073467
Hungary	3.276377	1.2497504	2.321823	1.8001558	0.000000	1.169334
Poland	2.292311	1.5295188	2.226411	2.0734671	1.169334	0.000000

Figure 9: Mean of each of the numerical values of the provided data.

9 Does the pair of observations with the largest Euclidean distance also have the largest Mahalanobis distance? Why so or why not?

No, the pair of observations wiht largest Euclidean distance is Cambodia and Switzerland (rows 56 and 37), but the pair with largest Mahalanobis distance is Mexico and Peru (rows 23 and 20).

Mahalanobis distance takes into account the variances and covariances of the dataset in order to compute the distance (in fact, it uses the inverse of the covariance matrix as the metric to compute distances). Hence, the effects discussed of variables with high variances (or pairs of variables with high covariance) on the euclidean distance are compensated in the Mahalanobis distance.

10 Is it theoretically possible for the Mahalanobis distance matrix and the Euclidean distance matrix to be identical? Explain your answer.

No. The Euclidean distance matrix is always the identity, and the Mahalanobis distance matrix is the inverse of the sample variance-covariance matrix. In $\mathbb{R}^n$

$$E = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}, M = \begin{bmatrix} s_1 & s_{12} & s_{13} & \dots & s_{1n} \\ s_{21} & s_2 & s_{23} & \dots & s_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ s_{n1} & s_{n2} & s_{n3} & \dots & s_n \end{bmatrix}^{-1}$$

If $E = M$, then the sample variance-covariance matrix would be:

$$S = M^{-1} = E^{-1} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$