Computation in Data science - PartII - HW1

B06702064 會計五 林聖硯

(a) Consider the census-tract data listed in Table 8.5. Suppose the observations on X5 = median value home were recorded in hundreds, rather than ten thousands, of dollars; that is, multiply all the numbers listed in the sixth column of the table by 100.

我們需要將第五行的X全部乘上100，才會得到正確的資料矩陣。

可以發現，在這個covariance matrix內，第五列(row)和第五行(column)的元素之非對角線元素相比於原本資料矩陣之covariance matrix，乘上了100倍；而對角線之元素乘上了10,000倍。這是因為在算其他特徵向量與該特徵之covariance時，第五行的資料相較於原資料乘上了100倍；而在算對角線元素(自己的變異數時)，100倍的影響則被放大了2次，變成10,000倍。

乘上100前(原本的資料矩陣之covariance matrix)

	Total.poupulation.thousands.	Median.school.years	Total.employment.thousands.
Total.poupulation.thousands.	4.308	1.684	1.803
Median.school.years	1.684	1.767	0.588
Total.employment.thousands.	1.803	0.588	0.801
Health.services.employment.hundreds.	2.155	0.178	1.065
Median.value.home..10.000s.	-0.253	0.176	-0.158

	Health.services.employment.hundreds.	Median.value.home..10.000s.
Total.poupulation.thousands.	2.155	-0.253
Median.school.years	0.178	0.176
Total.employment.thousands.	1.065	-0.158
Health.services.employment.hundreds.	1.969	-0.357
Median.value.home..10.000s.	-0.357	0.504

乘上100後

	Total.poupulation.thousands.	Median.school.years	Total.employment.thousands.
Total.poupulation.thousands.	4.308	1.684	1.803
Median.school.years	1.684	1.767	0.588
Total.employment.thousands.	1.803	0.588	0.801
Health.services.employment.hundreds.	2.155	0.178	1.065
Median.value.home..10.000s.	-25.347	17.555	-15.834

	Health.services.employment.hundreds.	Median.value.home..10.000s.
Total.poupulation.thousands.	2.155	-25.347
Median.school.years	0.178	17.555
Total.employment.thousands.	1.065	-15.834
Health.services.employment.hundreds.	1.969	-35.681
Median.value.home..10.000s.	-35.681	5043.802

(b) Obtain the eigenvalue-eigenvector pairs and the first two sample principal components for the covariance matrix in Part a.

The eigenvalue of PC1 and PC2 are $[\lambda_1, \lambda_1] = [5044.293, 6.682755]$

The eigenvalue of PC1 and PC2 are
$\hat{e_1'} = [-0.005, 0.003, -0.003, -0.007, 1.000], \hat{e_2'} = [0.784, 0.346, 0.329, 0.396, 0.007]$

```
> eigen_100$values
[1] 5.044293e+03 6.682755e+00 1.427369e+00 2.296417e-01 1.426671e-02
> round(eigen_100$vectors, 3)
       [,1]  [,2]   [,3]   [,4]   [,5]
[1,] -0.005 0.784  0.024  0.541 -0.302
[2,]  0.003 0.346  0.767 -0.541 -0.009
[3,] -0.003 0.329 -0.101  0.051  0.937
[4,] -0.007 0.396 -0.633 -0.642 -0.173
[5,]  1.000 0.007 -0.007  0.000  0.000
```

(c) Compute the proportion of total variance explained by the first two principlal components obtained in Part b. Calculate the correlation coefficients \(r_{\hat{y_i}, x_k}\); and interpret these components if possible. Compare your results with the results in Example 8.3. What can you say about the effects of this change in scale on the principal components?

PC1能夠解釋所有資料的變異比例為\(99.83466\%\)，PC2能夠解釋所有資料的變異比例為\(0.1322625\%\)。

```
> eigen_100$values[1] / sum(eigen_100$values)
[1] 0.9983466
> eigen_100$values[2] / sum(eigen_100$values)
[1] 0.001322625
```

每一個主成分對不同feature的correlation，可以看成是這個主成分被某個feature的組成比例(若correlation越高，代表這個主成分來自某個feature的比例越高。比較以下的圖後可看出，在scaling之前，第一個主成分與前四個feature關聯性都很強，但在scaling後，幾乎被第五個feature拉走(它們的correlation為1.0000)。也就是說，對資料做scaling，會很劇烈地影響到主成分與feature之間的關係，影響最後PCA的解釋。

original correlation matrix

```
                                        [,1]       [,2]          [,3]       [,4]          [,5]
Total.poupulation.thousands.       0.9909529 -0.04547131  0.0008309198  0.12466571  0.0180144632
Median.school.years                0.6052948 -0.76755723 -0.0767390991 -0.19693353  0.0001359553
Total.employment.thousands.        0.9840134  0.12395516  0.0101799361  0.02701916 -0.1235297055
Health.services.employment.hundreds. 0.7991523  0.55180790  0.0981029790 -0.21715814  0.0167633410
Median.value.home..10.000s.       -0.2014558 -0.49372623  0.8450726763  0.03380566 -0.0017337840
```

correlation matrix after multiplying column 5 with 100

```
> t(corr_100)
                                        [,1]         [,2]          [,3]          [,4]          [,5]
Total.poupulation.thousands.       -0.1721973 0.9768468232  0.0137883546  1.246583e-01 -1.796762e-02
Median.school.years                 0.1858521 0.6725551211  0.6892372238 -1.954445e-01 -6.240595e-05
Total.employment.thousands.        -0.2494069 0.9503990410 -0.1351435191  2.713045e-02  1.235913e-01
Health.services.employment.hundreds. -0.3582376 0.7298375498 -0.5392263655 -2.193339e-01 -1.679417e-02
Median.value.home..10.000s.         1.0000000 0.0002385117 -0.0001268801 -1.090766e-05  2.017942e-05
```