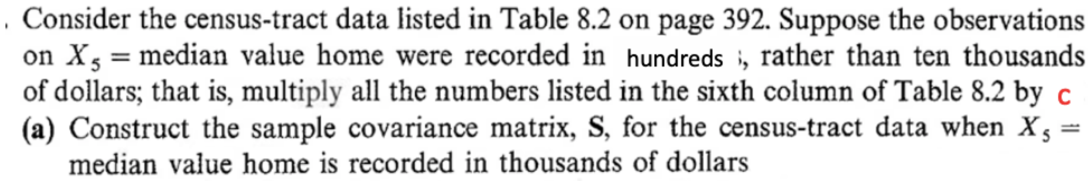
資料科學計算 HW1

R10946013 劉謦瑄

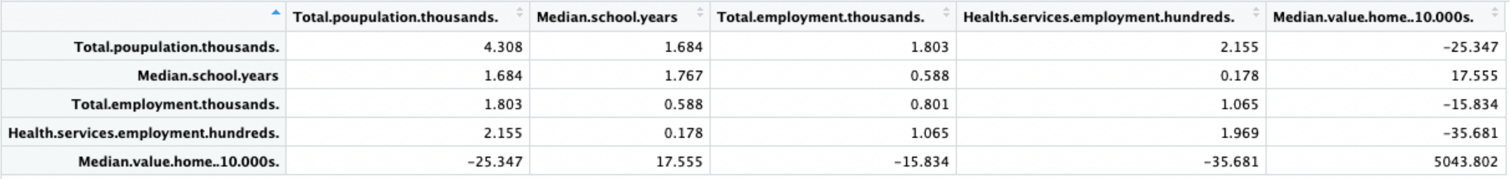
Due Date: 2021/11/03



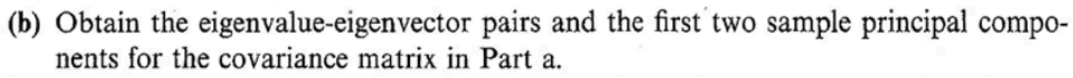
. Consider the census-tract data listed in Table 8.2 on page 392. Suppose the observations on X_5 = median value home were recorded in hundreds ;, rather than ten thousands of dollars; that is, multiply all the numbers listed in the sixth column of Table 8.2 by c

(a) Construct the sample covariance matrix, **S**, for the census-tract data when X_5 = median value home is recorded in thousands of dollars

→ **Sample variance matrix**:

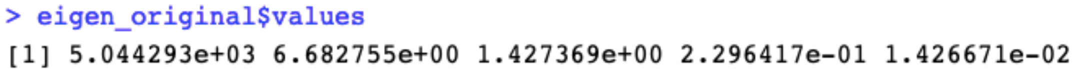


	Total.poupulation.thousands.	Median.school.years	Total.employment.thousands.	Health.services.employment.hundreds.	Median.value.home..10.000s.
Total.poupulation.thousands.	4.308	1.684	1.803	2.155	-25.347
Median.school.years	1.684	1.767	0.588	0.178	17.555
Total.employment.thousands.	1.803	0.588	0.801	1.065	-15.834
Health.services.employment.hundreds.	2.155	0.178	1.065	1.969	-35.681
Median.value.home..10.000s.	-25.347	17.555	-15.834	-35.681	5043.802



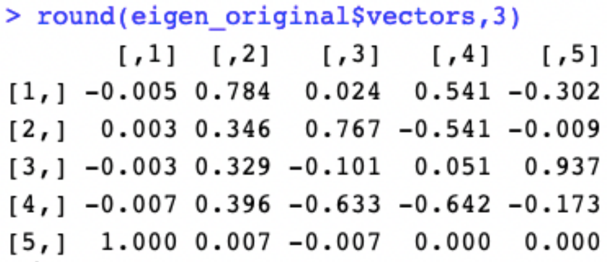
(b) Obtain the eigenvalue-eigenvector pairs and the first two sample principal components for the covariance matrix in Part a.

→ eigenvalue:

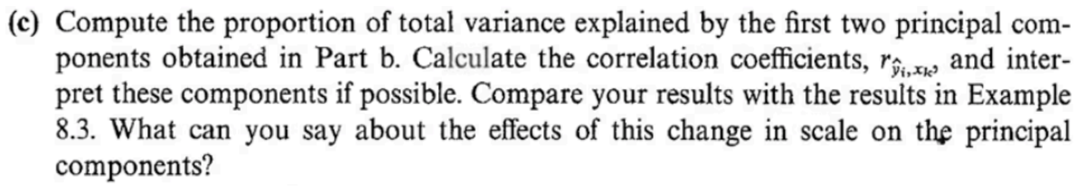


```
> eigen_original$values
[1] 5.044293e+03 6.682755e+00 1.427369e+00 2.296417e-01 1.426671e-02
```

→ eigenvector:

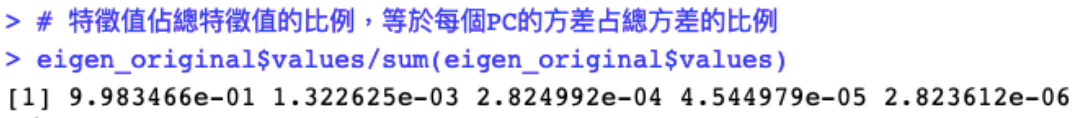


```
> round(eigen_original$vectors,3)
       [,1]  [,2]   [,3]   [,4]   [,5]
[1,] -0.005 0.784  0.024  0.541 -0.302
[2,]  0.003 0.346  0.767 -0.541 -0.009
[3,] -0.003 0.329 -0.101  0.051  0.937
[4,] -0.007 0.396 -0.633 -0.642 -0.173
[5,]  1.000 0.007 -0.007  0.000  0.000
```



(c) Compute the proportion of total variance explained by the first two principal 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?

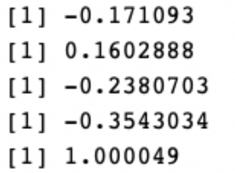
→ proportion of PCs and first two sample PCs:



```
> # 特徵值佔總特徵值的比例，等於每個PC的方差占總方差的比例
> eigen_original$values/sum(eigen_original$values)
[1] 9.983466e-01 1.322625e-03 2.824992e-04 4.544979e-05 2.823612e-06
```

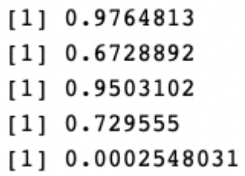
First PC: 0.9983466 / Second PC: 0.001322625

→ correlation coefficient of PC1:



```
[1] -0.171093
[1] 0.1602888
[1] -0.2380703
[1] -0.3543034
[1] 1.000049
```

→ correlation coefficient of PC2:



```
[1] 0.9764813
[1] 0.6728892
[1] 0.9503102
[1] 0.729555
[1] 0.0002548031
```

說明：

由更改單為 hundred 後的原始資料來看，X5 的數字分佈明顯與其他 X1~X4 相距甚遠，因此在未標準化的情況下，從我的 eigenvalue 以及 proportion of PCs 的計算結果來看，PC1 就已經可以解釋了約 99%的樣本數。

而從 correlation coefficient 的結果來看，r 值越趨近於 1 代表此 PC 與呈現正相關，越趨近於-1 則代表呈現負相關。因此，可以看出 PC1 和 X5 呈現了完全正相關的狀態，卻與其他的 characteristic 呈現不相關或負相關的狀態。

由實驗結果可得知，在進行 PCA 之前，進行標準化的重要性。