Module 9 Assignment on Unsupervised Methods: PC and Clustering

Jingwen Zhong // Graduate Student

5/2/2021

Module Assignment Questions

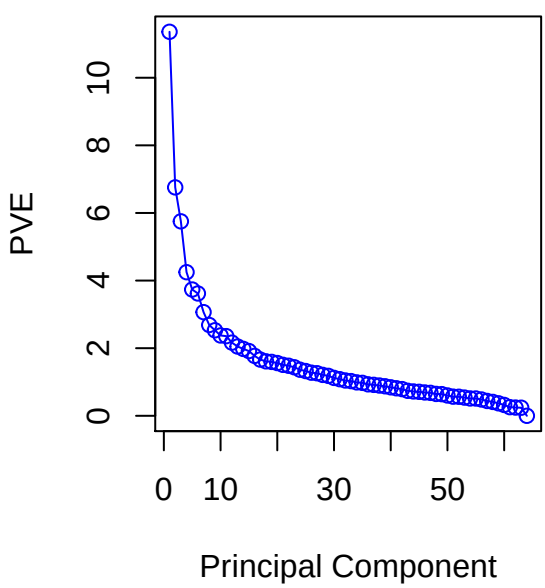
Applications

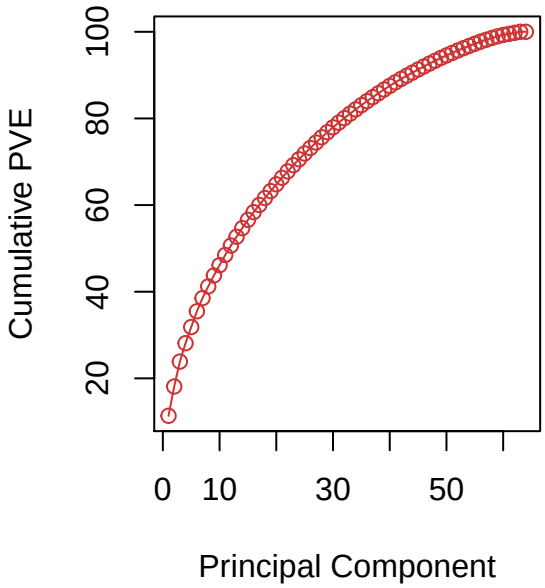
You will perform four unsupervised methods on a high dimensional data: `PCA`, `K-Means` clustering, `hierarchical` clustering, and one of `DBSCAN` and `GMM` clusterings. The data is the `NCI60` cancer cell line microarray data set, which consists of 6,830 gene expression measurements on 64 cancer cell lines. Each cell line is labeled with a cancer type: there is 14 imbalanced types. In performing unsupervised methods, we don't use labels. But after performing the clustering, we can check to see the extent to which these cancer types agree with the results of these unsupervised techniques. You will do this as well.

Do scaling before performing any unsupervised methods. Then, apply each method by justifying how you use and by including informative plots and summaries: each method has parameters and hyperparameters, consider these. Show how you decide optimal numbers of clusters. There is no unique answer key: any decision should be justified as long as you reflect our lab discussions and details. Don't include irrelevant and uncommented results. Make write-ups and outputs readable and compact. Include only necessary codes and outputs in minimalist format.

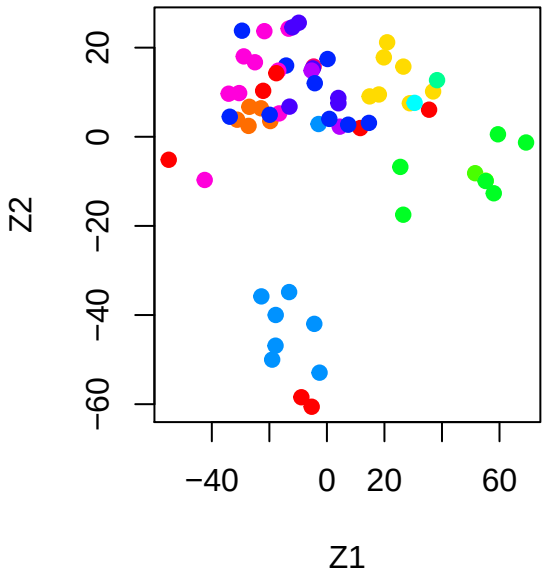
Fit the four models (1-4) below by including narratives in terms of data reduction and clustering, and answer the questions (5, 6):

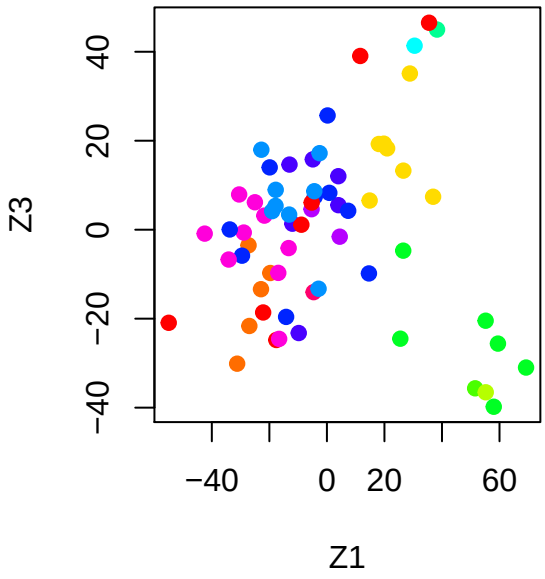
```
## nci.labs
##      BREAST         CNS       COLON K562A-repro K562B-repro    LEUKEMIA
##           7           5           7           1           1           6
## MCF7A-repro MCF7D-repro    MELANOMA       NSCLC     OVARIAN    PROSTATE
##           1           1           8           9           6           2
##       RENAL     UNKNOWN
##           9           1
```

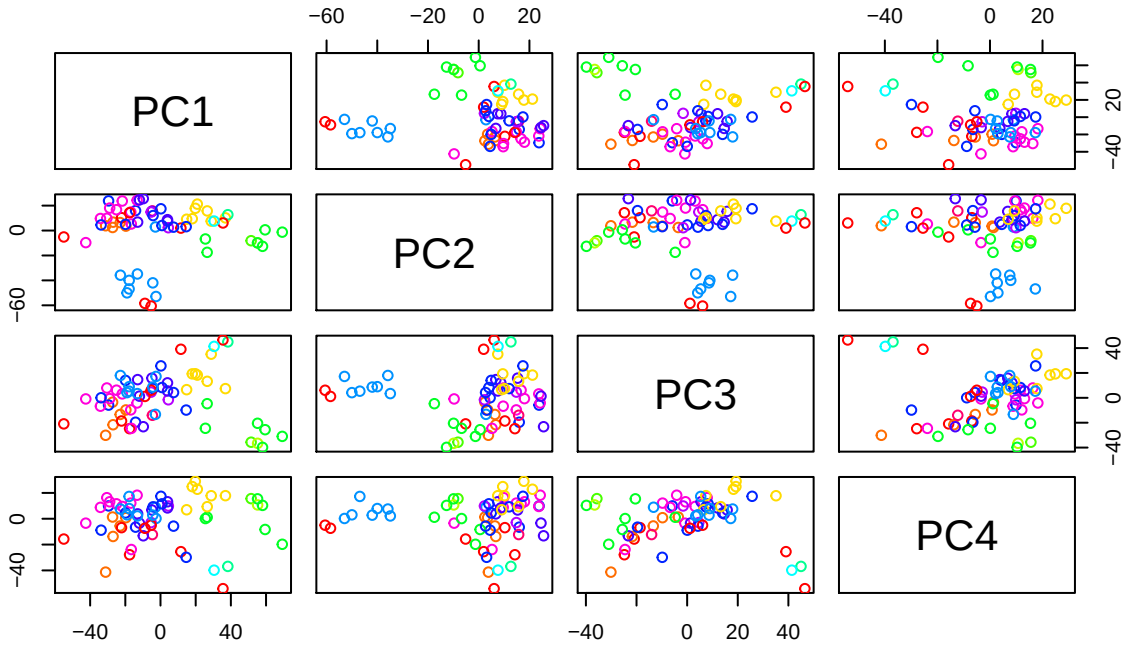

PVE
0
2
4
6
8
10
0
10
30
50
Principal Component


Cumulative PVE
20
40
60
80
100
0
10
30
50
Principal Component

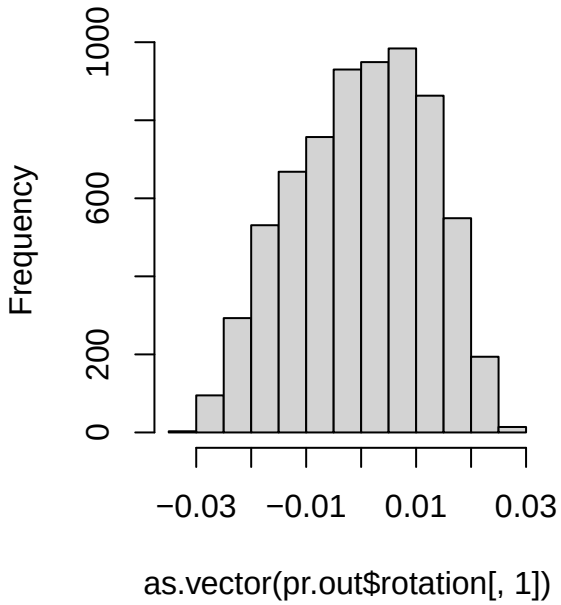
1. PCA


Z2
−60
−40
−20
0
20
−40
0
20
60
Z1


Z3
−40
−20
0
20
40
−40
0
20
60
Z1



istogram of as.vector(pr.out$rotatio



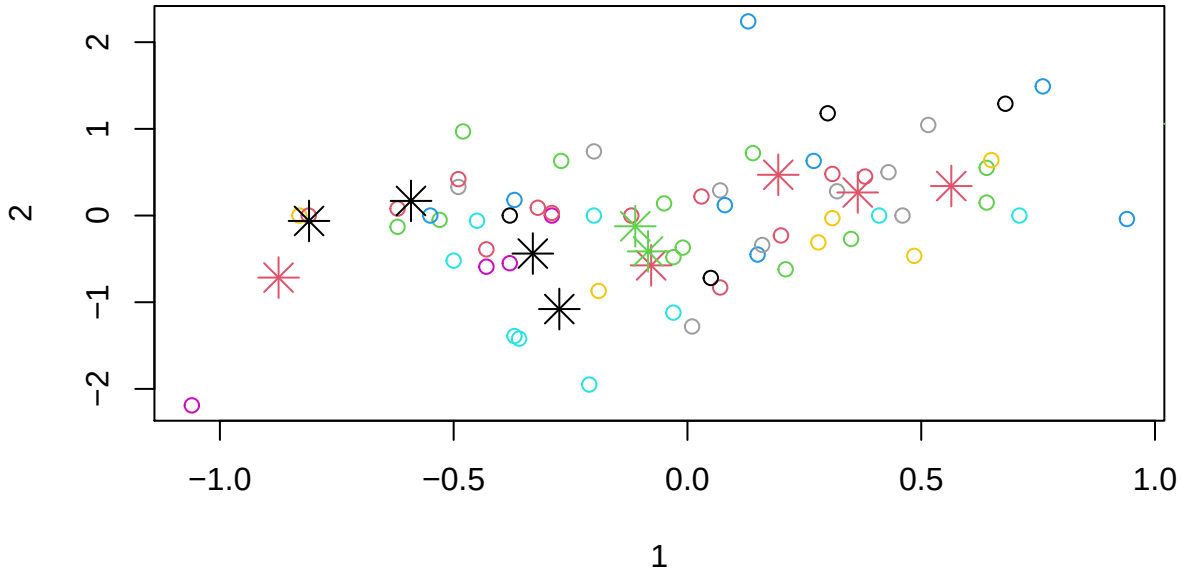
I decide 32 PCs is enough, because it can explain more than 80% variations. I use Hierarchical along with PCA to get the accuracy.

For the hyperparameter k, I use loop from 1 to 20 to check the k with the highest accuracy. I didn't choose method for hclust.

```
## accuracy = 0.203125 and k =  9
```

2. K-Means

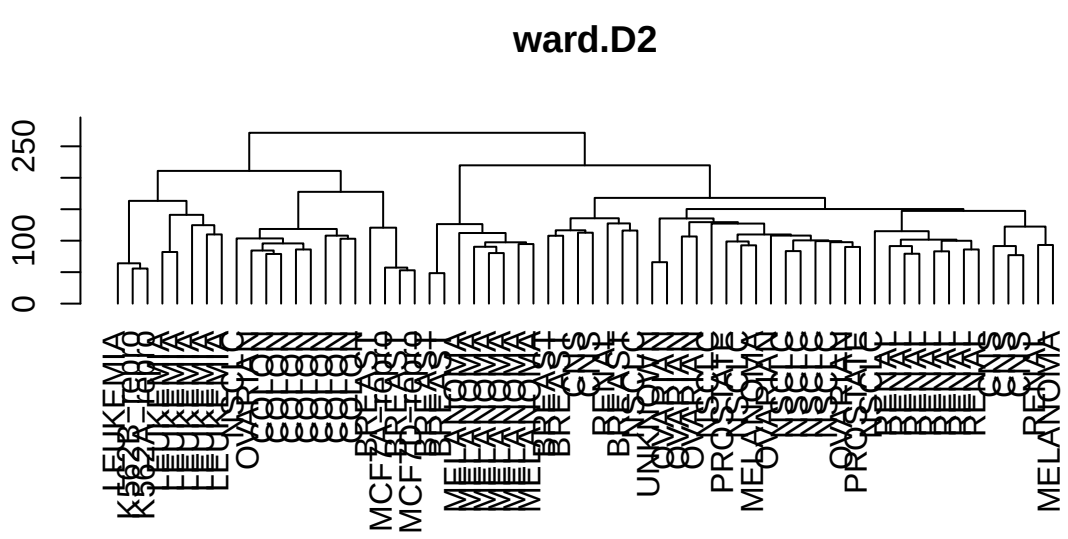
```
## accuracy = 0.234375 and k =  15
```



For the hyperparameter k, I use loop from 1 to 20 to check the k with the highest accuracy.

3. Hierarchical

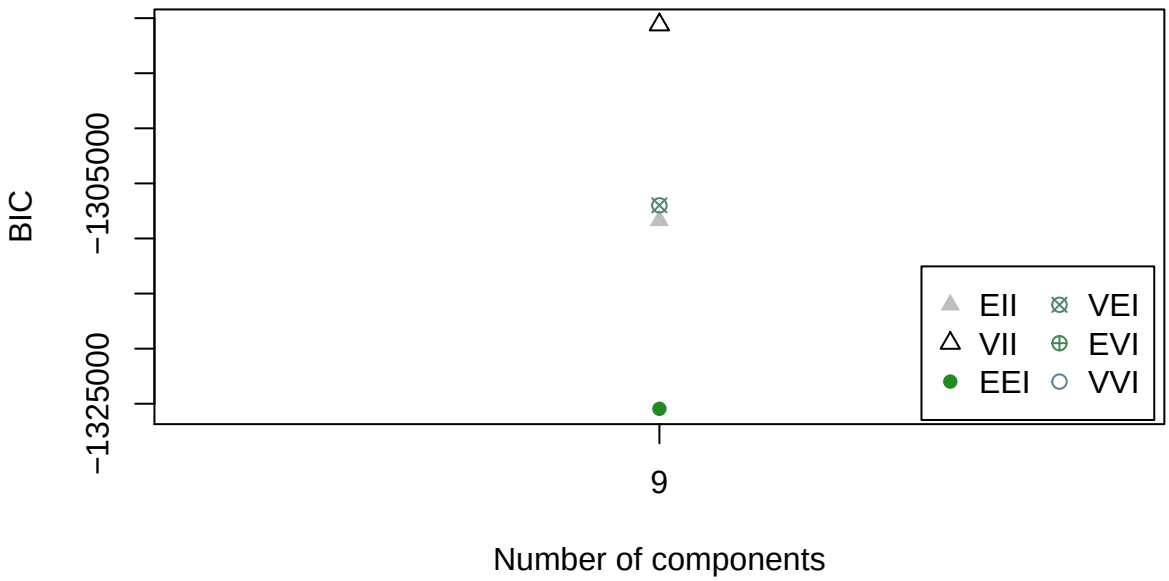
```
## accuracy = 0.21875 and k =  9
```



For the hyperparameter k, I use loop from 1 to 20 to check the k with the highest accuracy. I choose ward.D2 method for hclust.

4. GMM

```
## accuracy = 0.234375 and k =  9
```



For the hyperparameter G, I didn't use a loop to get it, because it said no storage when I loop through 1 to 20, but by hand checking I choose 9 to be my cluster and it has a high accuracy.

5. Do the comparison of the four methods above in a table by fitted clusters and true clusters: did clustering methods discover correct clusters? Include an accuracy table or a graph that compares the results obtained from the four methods. Comment.

Table 1: Accuracy table

	Accuracy
PCA with Hierarchical	0.203125
Kmeans	0.234375
Hierarchical	0.218750
GMM	0.234375

The clustering methods did a very bad job, the accuracy is only round 0.22, and it didn't discover correct clusters.

---

6. What insights/contextual conclusions did you get about the data from the PCA application? Explain. (this may overlap with your PCA narratives, here, be more contextual on the results of PCA in determining clusters of cancer types.)

more than 80% of the variance is explained by the first 32 principal components, and there is an elbow after the 7th component in PVE, but since the baseline of cumulative PVE is 80%, I still use 32 principle components.

from the graph of PC1, PC2, PC3 with true cancer types, it shows that the same colors do try to cluster together.

---

7. BONUS. Use any `manifold` method to cluster the data in terms of cancer types. Then check with the true labels. Does it discover? Explain and include graphs. BONUS:

Write comments, questions: ...

I hereby write and submit my solutions without violating the academic honesty and integrity. If not, I accept the consequences.

List the fiends you worked with (name, last name): Chenglu Xia

Disclose the resources or persons if you get any help: ...

How long did the assignment solutions take?: 10hrs

References

...