Statistical Learning for Data Science project: Exploratory data analysis of the Leukemia gene expression

Baurice Nafack

Course of AY 2021-2022 - Statistical Learning for Data Science,
Prof Matilde Trevisani

1 Problem statement

Researchers in genomics are frequently interested in understanding how their samples are related to one another and how different variables in the data set are correlated. They usually want to answer the following questions [2]: Are our replicates similar to each other? Do the samples from the same treatment group have similar genome-wide signals? Do the patients with similar diseases have similar gene expression profiles?. From these information, one can develop an hypothesis, identify outliers in the sample, or determine which sample groups require more data points. In this perspective, the project goal is to do an exploratory data analysis of the Leukemia gene expression to answer some scientists' questions.

2 Algorithms used to Approach the Problem

The clustering techniques(k-means, k-medoids, and hierarchical clustering) will be applied to find groups of patients that are more similar to each other than the rest of the patients, using the distance or similarity metric between patients' expression profiles. This is a stochastic algorithm that could have different results depending on where the algorithm starts, so we will need to specify a seed for reproducibility.

The dimension reduction technique (PCA, SVD) will be applied to reduce the number of genes expression or variables, to obtain a few principal highly variables that capture most of the variations inside the data instead of thousands. This is useful for clustering, predictive modeling and 2D or 3D visualization of many genes expression to identify modes.

3 Data description

The leukemia dataset consists of gene expression profiles for 72 patients, 47 of who suffer from ALL and 25 from AML leukemia type. This data comes from the landmark Science paper authored by Golub and colleagues in 1999 [1]. It contains 7128 genes with no missing data.

4 Experimental Procedure

The following steps were considered:

- Ordered Dissimilarity Matrix (ODM) was used to assess clustering tendency in the dataset.
- The partition clustering approach (k-means and k-medoids) was employed to find hidden groups within the patients gene expression. The average silhouette value for various k values has been applied to determine the appropriate number of clusters. The validation of cluster consistency within data clusters was endorsed using the within-cluster variation method known as the gap statistic.
- We then used the density plot visualization of the bivariate distribution for the first principal components to have a clearer vision of any clustering distribution within the gene expression of the patients. SVD is applied to plot samples on the reduced dimensions to visualize a clear separation of samples.
- The hierarchical clustering was applied using complete, single, average, ward.D, and ward.D2 linkage to find the relationship between individual data points and clusters. We look at the highest separation between clusters to identify the appropriate level to cut the dendrogram. We obtain a collection of clusters at various levels of the dendrogram cutting point, and we utilize majority voting to choose the appropriate number of clusters.

5 Results and discussion

5.1 Order Dissimilarity Matrix

The order dissimilarity matrix (figure 1(a)) helps us measure how similar gene expressions are to each other and displays the clustering tendency of genes that behave similarly and approximately three subsets are identified. The heatmap (figure 1(b)) shows gene expression values across patients such that each column represents a patient clustered using gene expression. As we can observe in the heatmap, each cluster has a distinct set of expression values. The main clusters almost perfectly distinguish the leukemia types, that will be explore using orther clustering methods.

5.2 Partitional Clustering Method

The number of clusters identified from ODM analysis was used to apply the The partitional clustering method, 23 patients out of 25 having AML Leukemia types were correctly clustered (table 1). The dataset contains only two Leukemia types, so we expected that one of the remaining clusters should have zero patients, but the patient from type ALL was spread into the remaining cluster. It implies there could be another leukemia genes expression type hidden inside the data that was miss label since there exist four main types of leukemia (ALL, AML, CLL, CML). To find the consistent number of cluster within the group, we applied the average silhouette method that helps to identify two main clusters and by using the elbow method, we identify three clusters of gene expression (figure 4 (b) and (d)). The gap statistic shows that $k = 5$ is the best if we take the maximum value as the best. However, after k=4, the statistic has more or less a stable curve. Based on the gap statistic criterion, the optimum number of clusters is 3 (figure 4).

k-means wrongly classify 2 patients from AML type for k=3 and one patient for k=2 (table 1). According to the k-means plot (figure3), We can easily identify a patient from Cluster 2 whose gene expression is significantly different from other gene expression clusters. As a result, the patient belongs to a different type of gene expression that is not represented in our data. Our hypothesis about the existence of a third-class label has been confirmed once more. We then used dimension reduction to conduct additional analysis.

5.3 Dimensionality Reduction Techniques

PCA was applied and using the scree plot to visualize the explained variation by eigenvectors figure 5(a). We identify that the first two top components explain 95% of the leukemia genes expression variation. The density plot visualization of the first two principal components allows us to determine two modes (figure 5(c)). The 2D representation of patients assisted in the visualization of potential clusters. We identified patient 31 from class type AML as the one who was miss classifier due to the proximity of it gene expression to ALL gene expression type(see figure 5(b)). We then apply k-means with $k = 2$ to the first two PCA components. The visualization helps us to identify a patient from cluster 1 who is very far from other data points. Another proof of a third-class label to take into account.

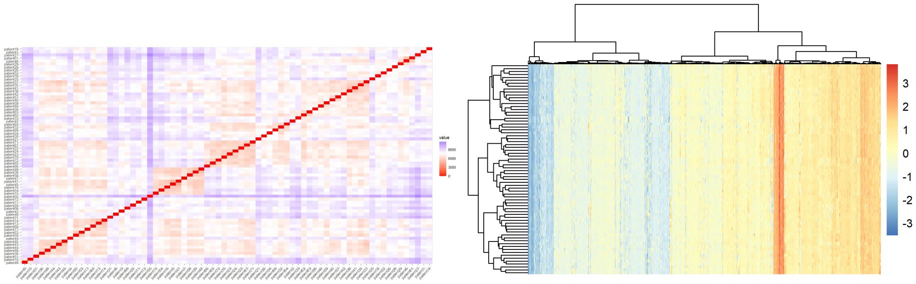
5.4 Hierarchical Clustering Methods

To be consistent with our hypothesis, different linkage clustering method and different level of the cut-off point of a dendrogram was applied to determine the best number of clusters for the hierarchical clustering analysis(figure 2(d)). That number is 2 based on the majority voting principle. That number of the cluster was used to cut the dendrogram, it appears that ward.D and ward.D2 linkage clearly shows patients are cluster in two groups, they also wrongly classifier 1

AML patient. The complete linkage clustering shows an interesting fact, patient 55 is classifier to be alone in its cluster (figure 2(c)). We suppose that it is the patient who was significantly different from other gene expression clusters in the previous methods. This confirms our initial hypothesis.

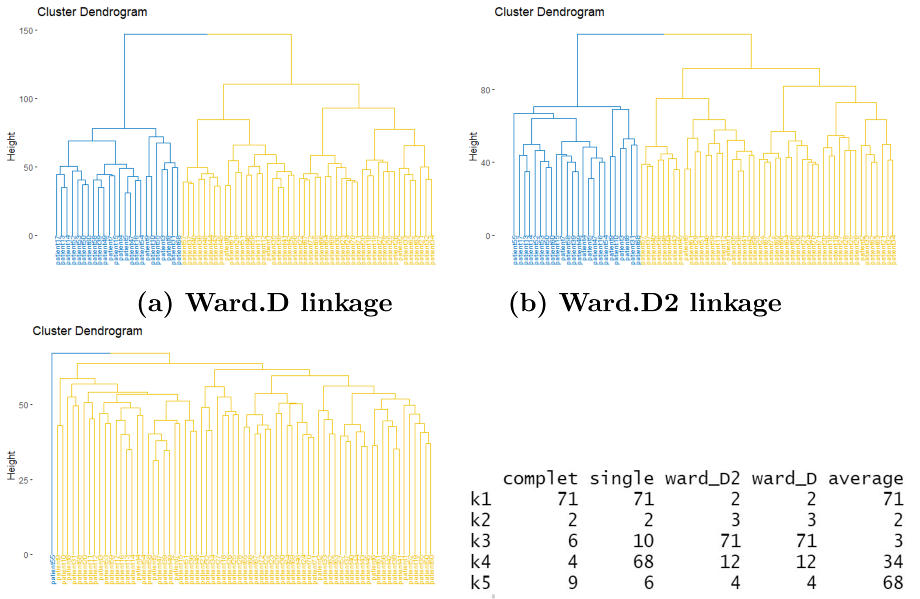
5.5 Conclusion

We knew there were two Leukemia gene expression types in the dataset when we started this analysis. The clustering techniques assisted us in confirming our hypothesis that patient 55 has a very different gene expression comparer to other samples and is also an outlier. It's likely that we should look into additional cancer subtypes. More data from this gene's expression should be collected for further investigation. Using hierarchical clustering, it is easy to see that some replicates in sub-categories are similar to each other, and patients from the same disease type do not always share similar gene expression profiles. Furthermore, it is important to understand why the gene expression of patient 31 is similar to those of AML and ALL leukemia types so he can receive better assistance.



(a) Order Dissimilarity Datrix **(b) Heatmap**

Figure 1: Clustering tendency in the dataset

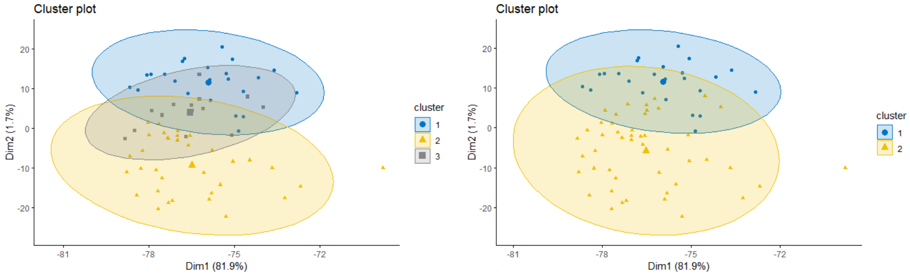


(a) Ward.D linkage **(b) Ward.D2 linkage**

```
   complet single ward_D2 ward_D average
k1      71     71       2      2      71
k2       2      2       3      3       2
k3       6     10      71     71       3
k4       4     68      12     12      34
k5       9      6       4      4      68
```

(c) complete linkage **(d) Clusters number per cutting point level**

Figure 2: The cut-off point of a dendrogram for clustering analysis



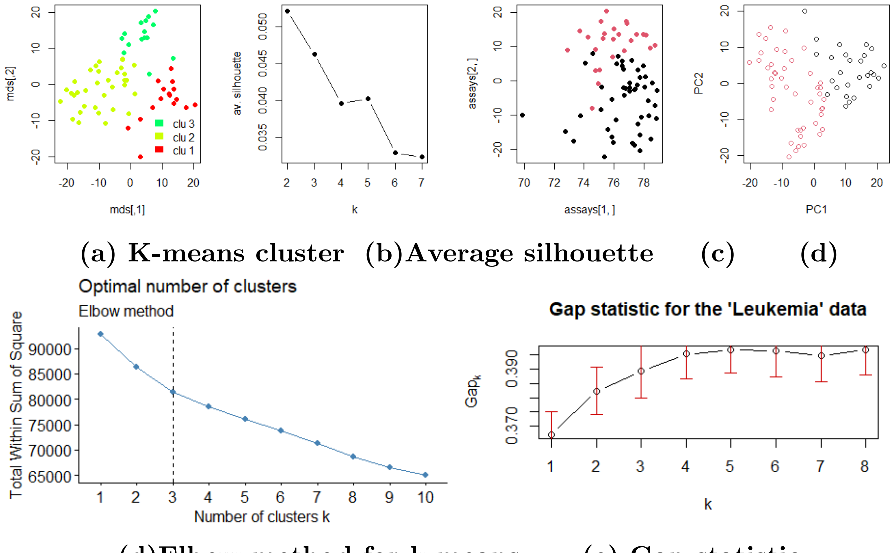
(c) k=3 **(d) k=2**

Figure 3: K-means Cluster

References

[1] Leukemia data. `https://hastie.su.domains/CASI_files/DATA/leukemia.html`. Accessed: 2022-05-22.

[2] Akalin Altuna. "Computational Genomics with R". In *Bioscience, Medicine, Dentistry, Nursing Allied Health*, New York, 2021. Chapman and Hall/CRC. https://compgenomr.github.io/book/.



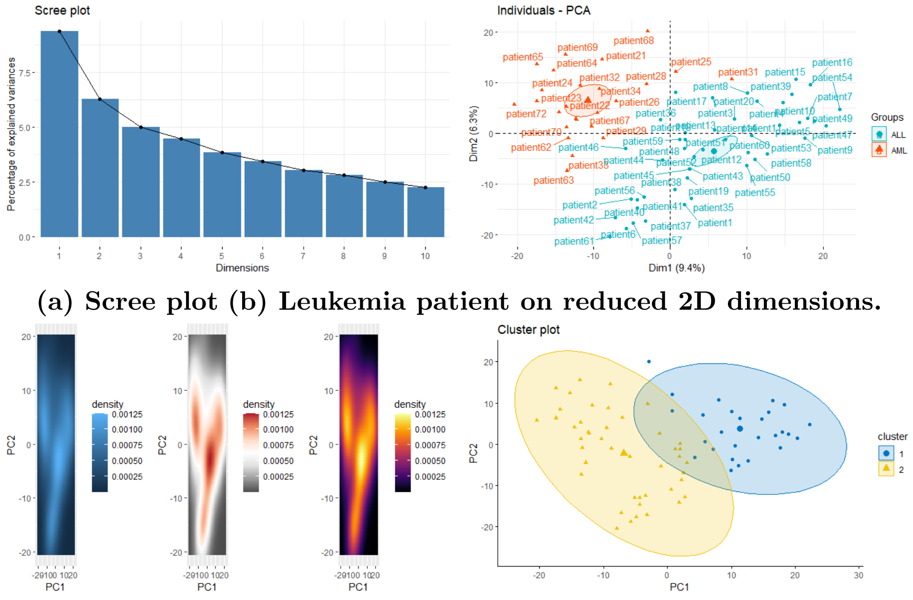
(a) **K-means cluster** (b) **Average silhouette** (c) (d)

(d) **Elbow method for k-means** (e) **Gap statistic**

Figure 4: b) Average silhouette values for k-medoids clustering for k values between 2 and 7. c) and d) Leukemia gene expression values per patient on reduced 2D dimensions

Clustering method	Cluster	ALL	AML
k-means with k=3	1	0	23
	2	33	1
	3	14	1
k-means with k=2	1	0	24
	2	47	1
k-means with PCA	1	24	3
	2	23	22
k-medoids with k=3	1	14	0
	2	26	2
	3	7	23
k-medoids	1	19	23
	2	28	2

Table 1: Partitional Clustering method result.



(a) Scree plot (b) Leukemia patient on reduced 2D dimensions.

(a) Firt 2 PCA components Density plot (d) K-means plot on PCA

Figure 5: Dimension Reduction using PCA