Research Project: Gene Expression Data

Hamza Akhtar

1 Problem Statement

Consider the NCI60 data set, which consists of 6,830 gene expression measurements for each of 64 cancer cell lines. Each cell line is labelled with a cancer type. Without using the information on the cancer type, we are interested to find out whether there are groups, or clusters, among the cell lines based on their gene expression measurements. You may like reduce the dimension of the data first before embarking on clustering. Try to provide some visualization for your findings. You can also check to see the extent to which these cancer types agree with the results of these unsupervised techniques.

2 Approach

All 64 cell lines are associated with a cancer type. Our aim is to find out whether are are groups formed within this data set that link the cell lines without using the information on the cancer type directly. We can use the cancer type at certain points to reflect upon our observations and gain an contextual insight into the groups and clusters formed through visualisations. The data set has a dimension of 6830 x 64, so in order to obtain visualisations, we will have to reduce the dimension to make this practical.
A method that will initially allow us to do this called "Principal Component Analysis" (PCA). First we will examine the types of cancer that span our data set:

2.1 Cancer types

	No. of Cancer Cell Lines
BREAST	7
CNS	5
COLON	7
K562A-repro	1
K562B-repro	1
LEUKEMIA	6
MCF7A-repro	1
MCF7D-repro	1
MELANOMA	8
NSCLC	9
OVARIAN	6
PROSTATE	2
RENAL	9
UNKNOWN	1

Table 1: Cancer Types

There are 14 types of cancer in the data set. We will use this information later on in the analysis.

3 Principal Component Analysis

We first plot and compare some of the first few principal component scores. This will allow us to visualise the data as we can scale the large dimension of the data set down to a more practical form and in doing so emphasises variance, bringing out patterns. Now the PC's go from PC1 to PC64 with PC1 having the largest score. We can begin by comparing PC1 and PC2.

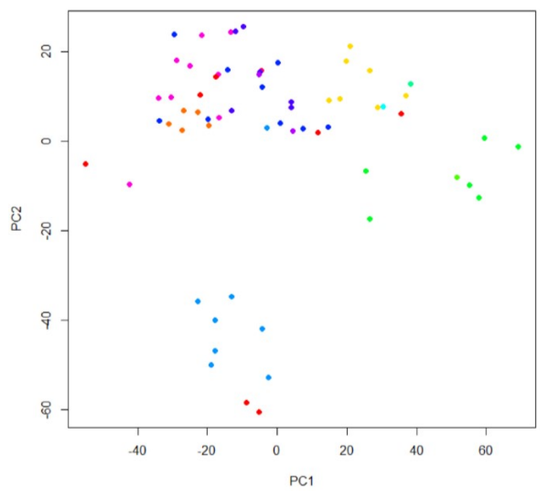

Figure 1: PC1 vs PC2

Figure 1 has plotted the 14 different types of cancer using different colours and we can see that most of the colours are located near to eachother. The exception in this case are the red dots, but all of the other colours seem to be in the same area.

We can now compare PC2 and PC3.

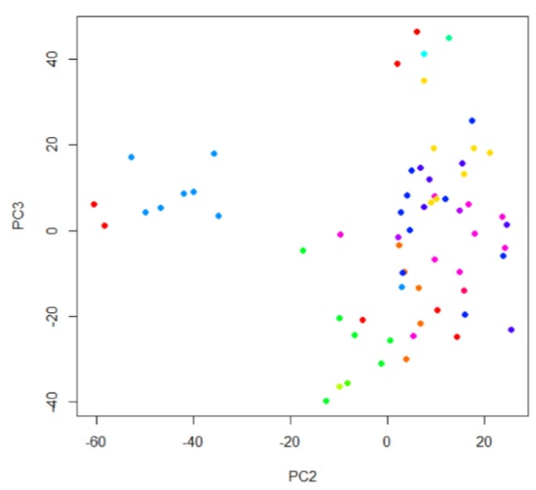

Figure 2: PC2 vs PC3

Similarly, **Figure 2** above shows us that again, apart from the red dots, the different coloured points seem to be in the same type of area. From these two figures, we can see that there is indeed some sort of pattern between the gene expression measurements and the cancer types.

We can produce a summary of the proportion of variance for each principal component score which informs us the proportion of the total variance associated with all 64 PCs. Ultimately this may tell us about the importance of the PCs and if there's a cut-off point for importance.

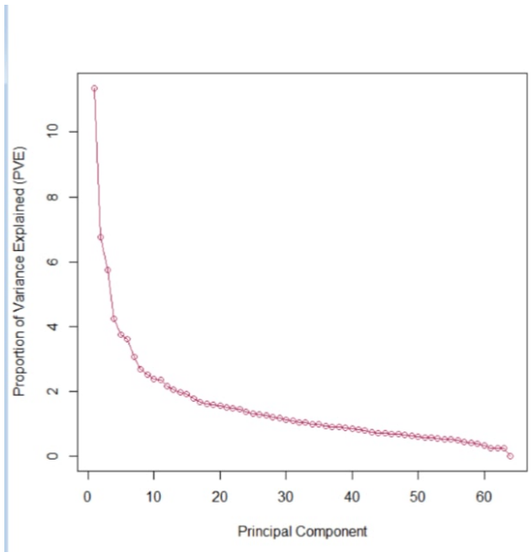

Figure 3: Proportion of Variance Explained (PVE)

We can see that after the 7th principal component, there is a swift decrease in the PVE for the PCs. This implies that this is the cut-off point as there would be little benefit in analysing after the 7th PC.

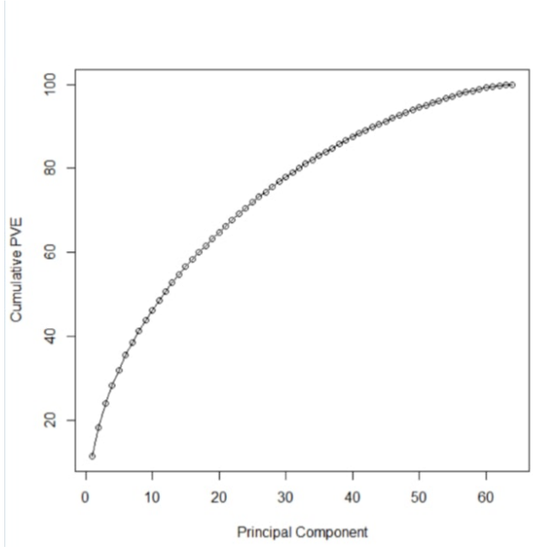

Figure 4: Cumulative PVE

Figure 4 shows that indeed, the total variance explained by all of the PCs is indeed equivalent to 100

4 Clustering

4.1 K-Means Clustering

From the "Principal Component Analysis" section, we have concluded that there are indeed some patterns in the dataset. In this section we will use clustering techniques to analyse how the patterns translate into distinct grouping in relation to the types of cancer. We use hierarchical clustering techniques such as single, average and complete linkage, also using Euclidean distance as the dissimilarity measure to see if there are indeed groups and clusters.

With single linkage clustering, all branches seem to tend back to one source, where observations are attached to one large cluster one by one. Considering the context of our problem and our objective, we prefer to use average and complete linkage as they provide more wholesome visualisations for this sort of problem.

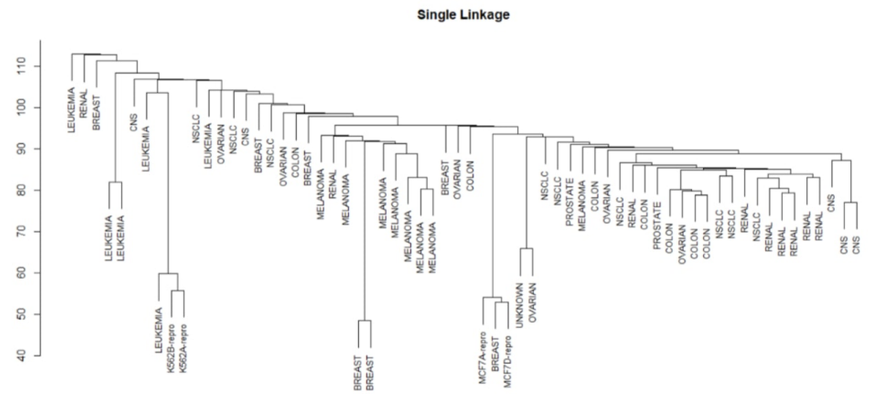

Figure 5: Single Linkage Clustering

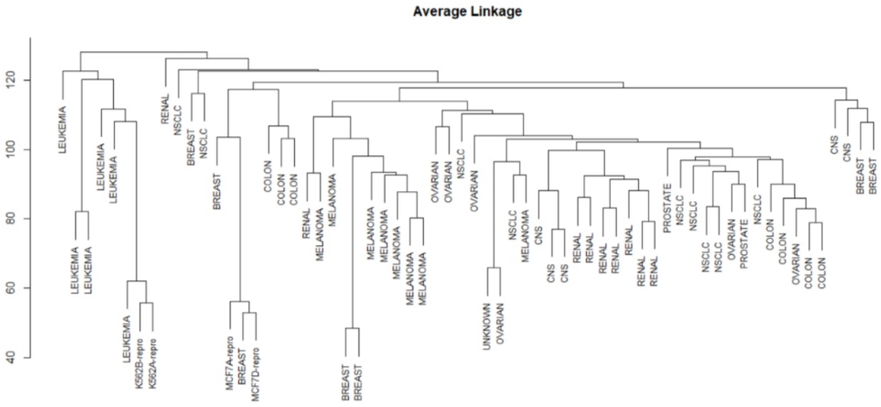

Figure 6: Average Linkage Clustering

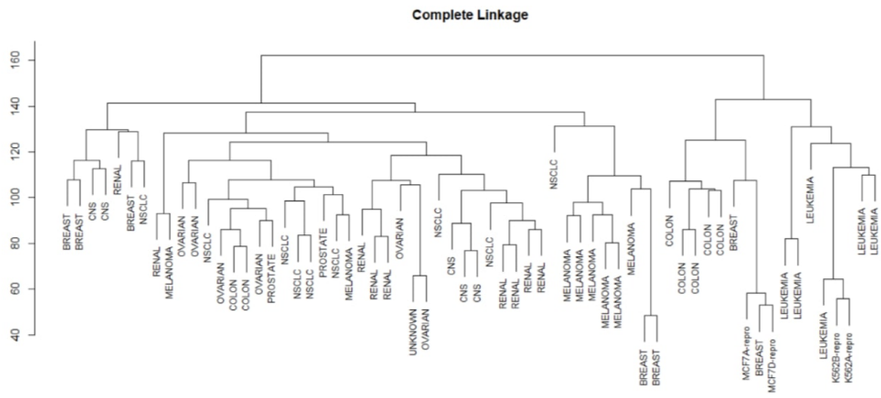

Figure 7: Complete Linkage Clustering

We will use complete linkage to analyse the data.

Cluster	BR	CNS	CLN	K562A	K562B	LEUK	MCF7A	MCF7D
1	5	5	2	0	0	0	0	0
2	0	0	0	1	1	6	0	0
3	2	0	5	0	0	0	1	1

Cluster	MLNMA	NSCLC	OVRN	PRST	RNL	UNK
1	8	9	6	2	9	1
2	0	0	0	0	0	0
3	0	0	0	0	0	0

Table 2: K-Means clustering k=3

For a 3 cluster analysis, we can clearly see that 12 of the cancer types are located in one cluster. The Breast cancer and Colon cancer observations are spread over multiple clusters, clusters 1 and 3. There is a very very clear pattern here.

Cluster	BR	CNS	CLN	K562A	K562B	LEUK	MCF7A	MCF7D
1	2	3	2	0	0	0	0	0
2	3	2	0	0	0	0	0	0
3	0	0	0	1	1	6	0	0
4	2	0	5	0	0	0	1	1

Cluster	MLNMA	NSCLC	OVRN	PRST	RNL	UNK
1	8	8	6	2	8	1
2	0	1	0	0	1	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0

Table 3: K-Means clustering k=4

For a 4 cluster analysis, 9 of the cancer types are all in one cluster. For example, Melanoma and Ovarian observations are all in cluster 1. Breast cancer cell lines are spread out over 3 clusters.

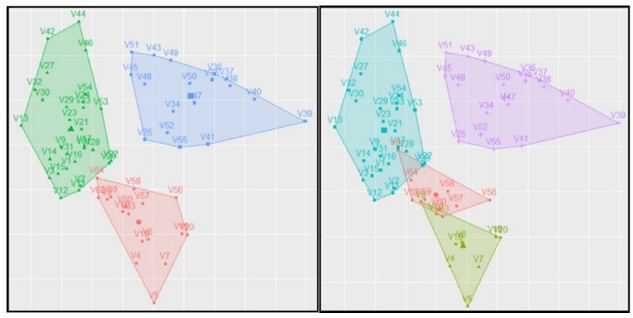

Figure 8: K=3

Figure 9: K=4

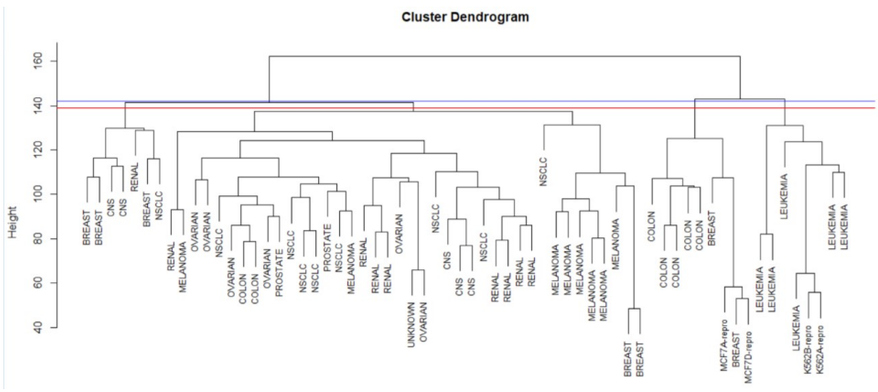

Figure 10: K=3 Clusters in blue and K=4 Clusters in Red

4.2 Hierarchical Clustering

We will now analyse Hierarchical Clustering using a decreased amount of principal components. Earlier we observed that there was an "elbow" in **Figure 3** after the 7th principal component. This implied that only the first 7 principal components are worth analysing. The dissimilarity measure is now based on the first 7 PCs. In this section, we will attempt to find trends and describe them using clustering using this information.

Cluster	BR	CNS	CLN	K562A	K562B	LEUK	MCF7A	MCF7D
1	5	5	7	0	0	2	0	0
2	0	0	0	1	1	4	0	0
3	2	0	0	0	0	0	1	1

Cluster	MLNMA	NSCLC	OVRN	PRST	RNL	UNK
1	8	9	6	2	9	1
2	0	0	0	0	0	0
3	0	0	0	0	0	0

Table 4: Hierarchical Clustering with K=3

Using the Hierarchical clustering method where we only consider the first 7 PCs, we can see that the clusters produced do in fact vary. Cluster 3 from the hier. method is kind of similar to cluster 3 in k-means clustering, but not too similar. For k-means with k=3, there were 12 cancer types only spread over one cluster and in hier. case, there are also 12 spread over 1 cluster and none over 3.We can clearly see that the clusters in both methods are slightly similar however, overall there is a very clear difference in their contents.
Again we see that the four clusters obtained using hier. clustering and k-means clustering are somewhat different. Cluster 3 in both methods are identical however the other clusters differ significantly.

Cluster	BR	CNS	CLN	K562A	K562B	LEUK	MCF7A	MCF7D
1	0	2	7	0	0	2	0	0
2	5	3	0	0	0	0	0	0
3	0	0	0	1	1	4	0	0
4	2	0	0	0	0	0	1	1

Cluster	MLNMA	NSCLC	OVRN	PRST	RNL	UNK
1	1	8	5	2	7	0
2	7	1	1	0	2	1
3	0	0	0	0	0	0
4	0	0	0	0	0	0

Table 5: Hierarchical Clustering with K=4

Cluster 2 in K-means clustering is identical to cluster 3 in hierarchical clustering. However, the other clusters differ lots. The k-means method had 9 of the cancer types spread over 1 cluster, but hier. clustering has only 7 types spread over 1 cluster. Breast cancer cells in both methods were almost identically distributed over 3 clusters.

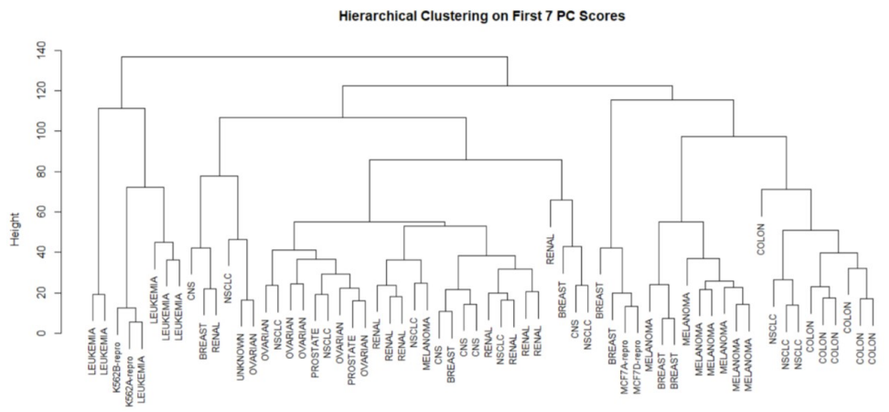

Figure 11: Hierarchical Clustering Dendrogram

Comparing **Figure 11** to **Figure 7** which was the complete linkage dendrogram, we can see that there is a very clear visual difference.
Sometimes it is better to perform clustering on the first few important principal components rather than performing it on the whole data set.

5 Conclusion

The dimension reduction medium that we used allowed us to gain an insight into the groups and clusters that we have in our NCI60 dataset. Ideally I would expect the the optimal cluster amount to be 14 assuming that when we reduce the dimension, the cell lines are evenly clustered so that they don't overlap. This is very unlikely as of course data will overlap and we see from **Figure 1** that this is not the case. Generally there are indeed trends and groups in the dataset that prove that there are link between the gene expression measurements and types of cancer.