We now use the help of coding(here we use R) and build two function calculating the number lumping and splitting errors

While learning the kmeans algorithm we came across the fact the kmeans algorithm depends on the initial values we start with. Thus it is always better to run multiple kmeans before we come up with conclusions.

Here we run the algorithm 3 times. Hence number of lumping errors for 3 runs are,

[1] 85
[1] 82
[1] 82

and the number of splitting errors for 3 runs are,

[1] 96
[1] 90
[1] 101

For accuracy, there are 64 Ö 63/2 = 2016 pairs, so the error rate here is pretty good here. Among these classes we are interested in knowing whether few clusters are easier to group and identify and those which are particularly difficult for the

algorithm to cluster.We proceed by forming a matrix with group names as rows and clusters as columns. We need to nd a measure which gives us an idea about the clustering strength,silhoutte plot , entropy are few measures.Here we use entropy to find the clustering.

The greater the value of purity indicates good clustering. The entropy is negative measure, the lower the entropy the better clustering it is.

Now like earlier we calculate the entropy for different classes for 3 times and we get,

##BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA
1.3517840 0.6730117 0.4101163 0.0000000 0.0000000 1.3296613
MCF7A-repro MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE
0.0000000 0.0000000 0.3767702 1.2730283 0.4505612 0.0000000
RENAL UNKNOWN
0.6837389 0.0000000
BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA
1.5498260 0.6730117 0.4101163 0.0000000 0.0000000 0.8675632
MCF7A-repro MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE
0.0000000 0.0000000 0.3767702 1.4270610 1.0114043 0.0000000
RENAL UNKNOWN
0.6837389 0.0000000
BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA
1.3517840 0.0000000 0.4101163 0.0000000 0.0000000 1.0114043
MCF7A-repro MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE
0.0000000 0.0000000 0.3767702 1.2730283 1.0114043 0.0000000
RENAL UNKNOWN
1.2148897 0.0000000

We can see throughout the 3 runs of the algorithm the entropy has changed for all the classes but for some classes though it has changed ,it always have been

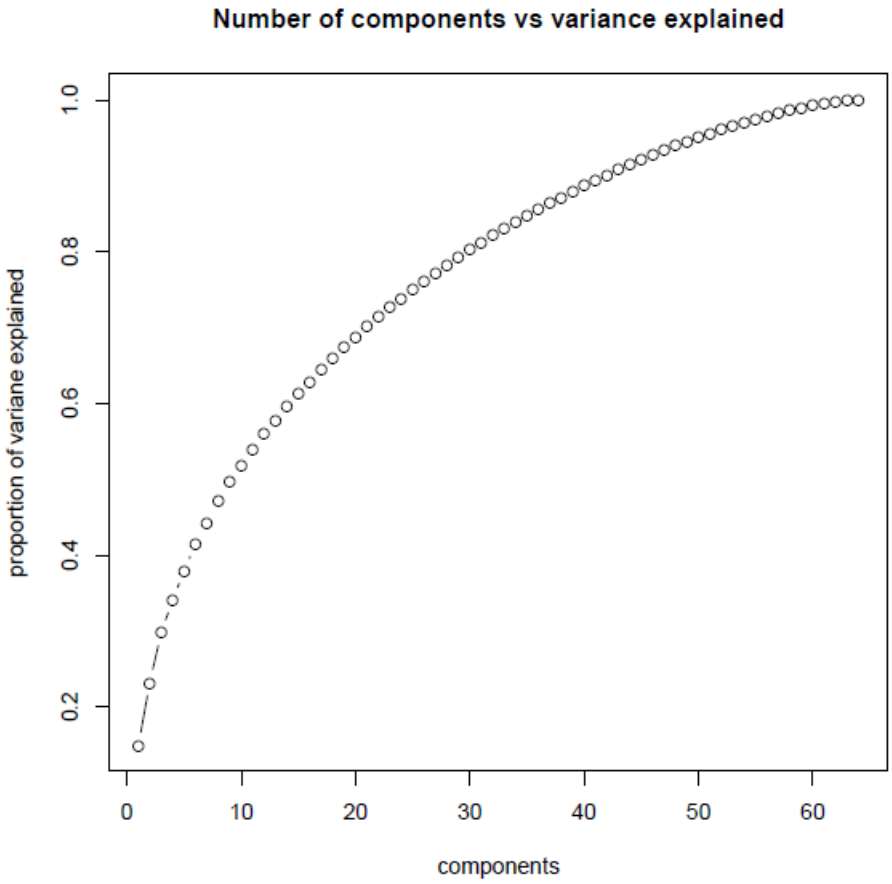
on the higher side.BREAST and NSCLC are classes which have consistently high cluster entropy, indicating that cells of these types tend not to be clustered together and hence making it dicult for the kmeans algorithm to identify this clusters.

```
## CNS CNS CNS RENAL BREAST CNS
## 1 1 1 1 1 1
## CNS BREAST NSCLC NSCLC RENAL RENAL
## 1 1 6 7 3 3
## RENAL RENAL RENAL RENAL RENAL BREAST
## 3 3 3 3 3 13
## NSCLC RENAL UNKNOWN OVARIAN MELANOMA PROSTATE
## 13 6 6 6 6 14
## OVARIAN OVARIAN OVARIAN OVARIAN OVARIAN PROSTATE
## 14 14 14 14 14 14
## NSCLC NSCLC NSCLC LEUKEMIA K562B-repro K562A-repro
## 4 4 4 2 8 8
## LEUKEMIA LEUKEMIA LEUKEMIA LEUKEMIA LEUKEMIA COLON
## 8 2 9 9 2 14
## COLON COLON COLON COLON COLON COLON
## 11 11 11 11 11 11
## MCF7A-repro BREAST MCF7D-repro BREAST NSCLC NSCLC
## 10 10 10 10 14 5
## NSCLC MELANOMA BREAST BREAST MELANOMA MELANOMA
## 5 12 12 12 12 12
## MELANOMA MELANOMA MELANOMA MELANOMA
## 12 12 12 12
```

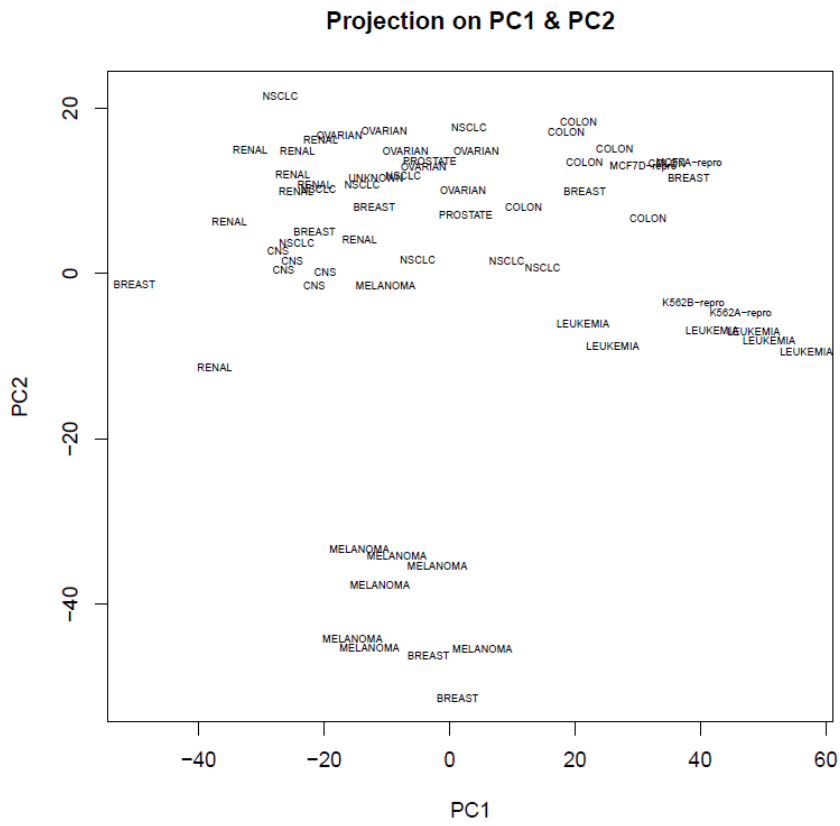
we can see cell 1 & 2 are always put into same cluster which belongs to CNS and likewise cell 4 & 5 are always put into same cluster while cell 4 belongs to RENAL and 5 belongs to BREAST.

Till now we have considered all the variables and perform our analysis but per- forming PCA we can reduce the dimensions to a much lower dimensional for ease of computation. R has inbuilt function prcomp which we use to perform the PCA. We can immediately see that the dimension reduce from 6830 to 64 as we know the sigma matrix to be positive denite we have n>p. So this immediate reduction in dimension is logical and we look forward to lower dimensions where we can explain most of the variance.We prove it through the inbuilt function prcomp.

Plotting the variances explained against the number of components gives us an idea about how much reduction in dimension is still possible



From the above plot we can say for about 40 components more than 90% of the variances is captured so still the dimension can be reduced from 64 to 40. Now we look forward to plot the projection of each cell on to the rst two principal components. The plot is as follows



From the plot we can say that classes like CNS & MELANOMA form a close clustering while classes like BREAST does not form a compact cluster and is scat- tered all around the plot.But we should also keep in mind that we do a tremendous dimension reduction from 6830 to 2 and if the preserved variance would have been much more we would have been condent of our deduction but in reality the 1st 2 principal components only explain 25% approximately of the total variances and thus we cannot be so certain about our conclusions.