Microarray Assignment

Avani Karve & Siddharth Sinnarkar

2025-04-26

Brain Cancer Dataset overview:

Brain cancer, or more accurately brain tumors can be broadly classified into 2 types - benign (non-cancerous) and malignant (cancerous). Our dataset contains data pertaining to the following 4 types of brain cancers:

1. **Ependymoma:** It is a rare type of brain cancer which originates from cells lining the ventricles of the brain. It can be benign or malignant
2. **Gliobastoma:** It is a fast-growing, aggressive brain tumor and is the most common primary (originating in the brain) malignant brain cancer in adults. It develops from glial cells, which are the supporting cells of the brain and spinal cord.
3. **Medulloblastoma:** It is a malignant brain cancer that develops in the cerebellum (the part of the brain responsible for balance and co-ordination). It is the most common malignant brain cancer in children.
4. **Pilocytic Astrocytoma:** It is a slow-growing benign brain cancer, often found in the cerebellum (the part of the brain responsible for balance and co-ordination) that typically affects children and young adults. It is considered as highly treatable, with a high cure rate. It originates from astrocytes, the star-shaped cells that support and nourish neurons in the brain.

Microarray Dataset:

A single channel microarray dataset consists of gene expression values for healthy and diseased group/s of individuals. Gene expression value for a particular gene essentially gives us the amount of protein synthesized by that particular gene. If we compare the gene expression values for healthy individuals with that of diseased individuals, we can find the specific gene/s whose hypo-expression or hyper-expression results in the underlying disease.

Introduction

Suppose, we are given the gene expression values of a particular individual and we want to check if this individual is susceptible to brain cancer. How can we proceed? First, we'll need to locate or determine the genes that are responsible for causing brain cancer/tumor.

Consider the following microarray dataset that contains gene expression values of 54675 genes of 130 individuals (13 healthy individuals and 117 (= 46 + 34 + 22 + 15) individuals with the above four types of cancer (respectively).

Note: We will be treating the set of gene expression values for each individual as an independent observation, and gene expression values will be treated as a variable. Different types of cancer will be treated as different groups. So in all, we will have 5 groups/treatments.

```
data = read.csv("D:\\MSc Statistcs SPPU\\SEM
II\\Microarray\\Assignments\\Clean_Brain_5C.csv", row.names = 1)
dim(data)   # to ensure that the entire data is loaded properly

## [1] 54675   130
```

Data Preprocessing:

Usually, gene expression values are in the range 0 to 2^16. But handling values with such vast range is a difficult task. So, in order to shrink this range (to 0 to 16), we log transform the data with base 2. However, it is recommended to check if our data is already log transformed or not.

```
range(data)

## [1]  2.740766 14.928264
```

In our case, our gene expression values are in the required range. So here, we do not need to perform log transformation.

In Microarray Data Analysis, we have 2 main assumptions:

1) Number of hyper expressed genes and number of hypo expressed genes is approximately same.

2) Only a few hundred genes are differently expressed. As a result of this, the distribution of every observation is approximately same.

In order to ensure that the required assumptions are satisfied, we normalize our data (i.e. in some sense, we make the distribution of all variables the same). However, while performing analysis, we shall check if our data is already normalized. This can be done using Concordance Coefficient. Concordance Coefficient, ρ is given by

$$\rho = \frac{2 \cdot \mathrm{Cov}(X, Y)}{\mathrm{Var}(X) + \mathrm{Var}(Y) + (\mu_X - \mu_Y)^2}$$

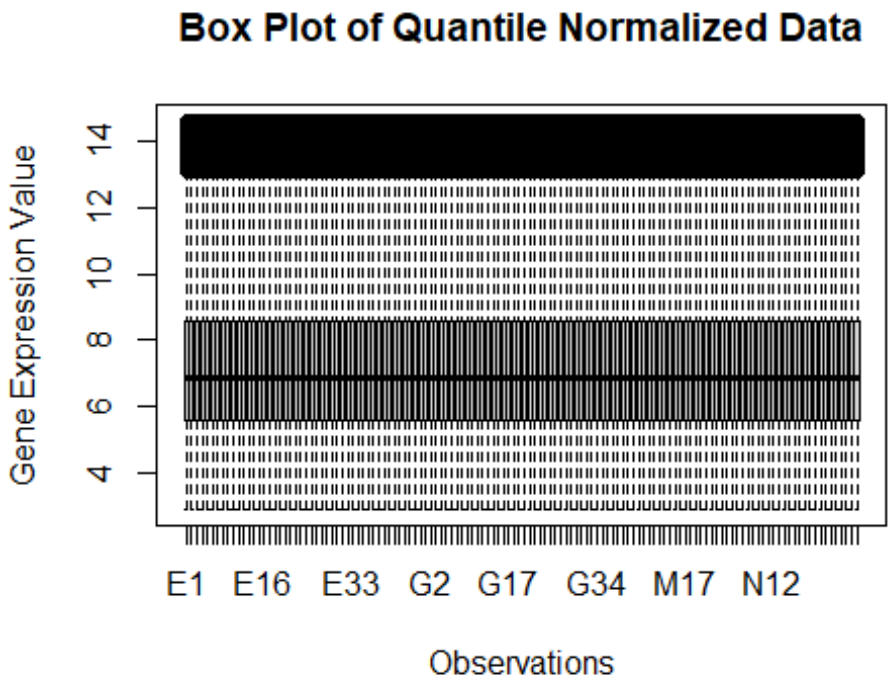
We compute concordance coefficient for all observation pairs. If all these coefficient values are greater than 0.9, then we can safely assume that our data is normalized.

```
print(min(ConMat))
```

```
## [1] 0.7424927
```

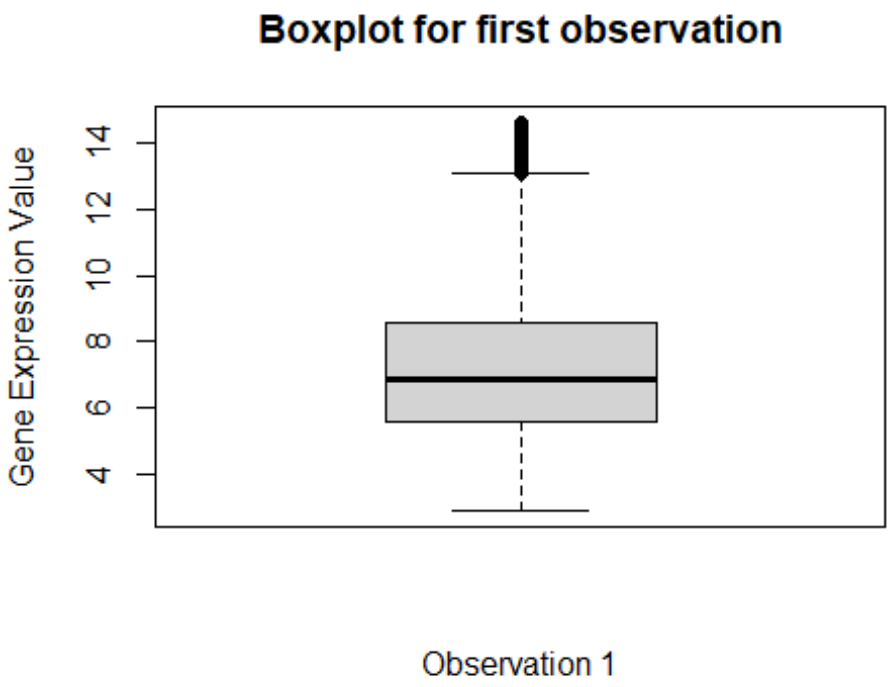
In our case, the minimum value of concordance coefficient is 0.742. Hence, we will perform normalization.

```
boxplot(QD, main = "Box Plot of Quantile Normalized Data",
        ylab = "Gene Expression Value", xlab = "Observations")
```



As we can see, all the Box plots now look the same. We have successfully carried out normalization. Let us see the Box plot for one observation to get a better understanding.

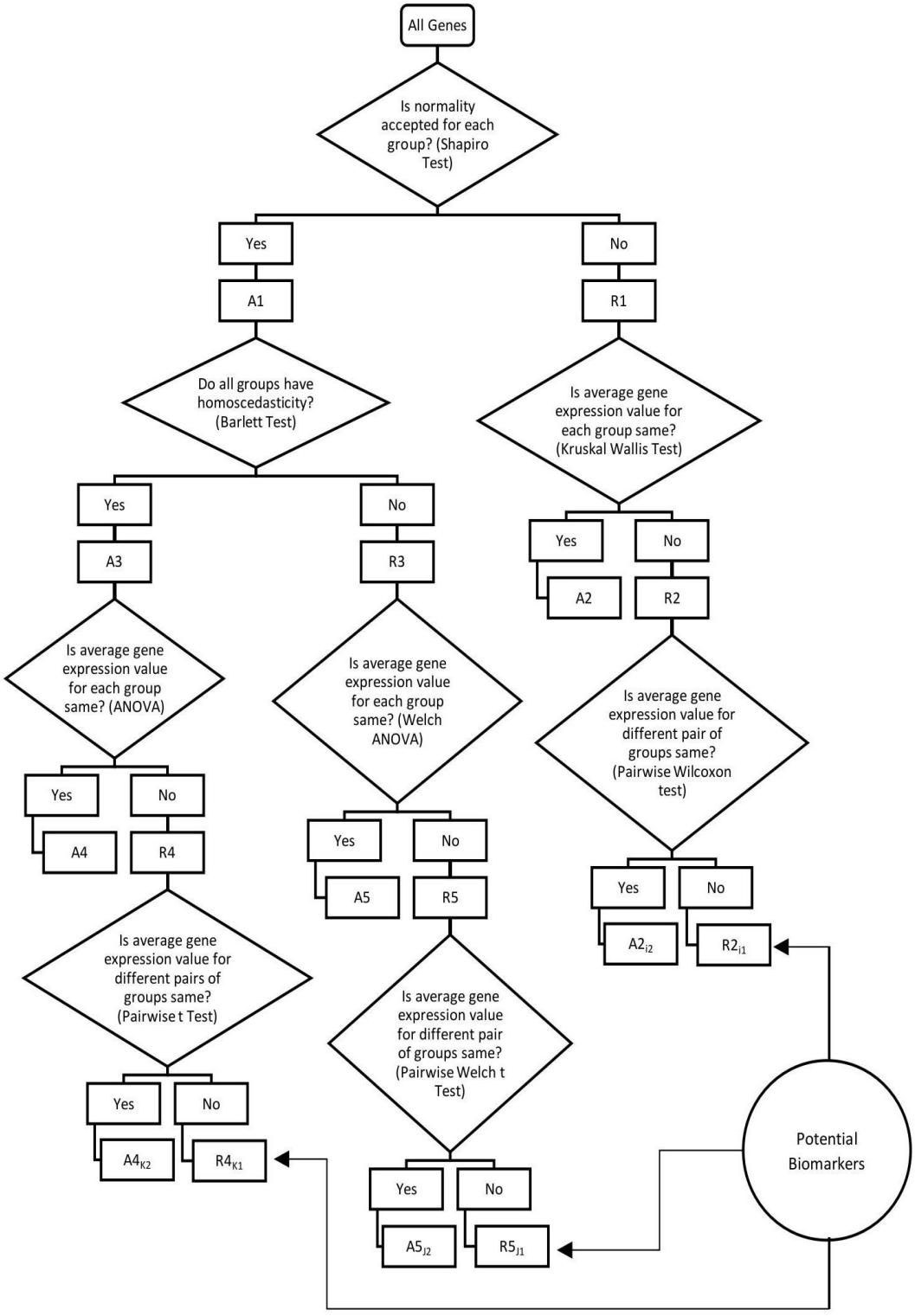
```
boxplot(QD[,1], main = "Boxplot for first observation", ylab = "Gene
Expression Value", xlab = "Observation 1")
```



Observe that the gene expression values for all the outliers lie on one side of the Boxplot. Now, that our data is in the required format, we can proceed to analysis.

Inference Procedures:

Our job is to find a set of biomarkers (i.e. differently expressed genes) and see how their gene expression is related to developing brain cancer. Inference procedures help us in identifying these biomarkers. Basically, we want to test whether the average gene expression value for each group is same. We perform a few tests as per the following tree diagram:


All Genes
Is normality accepted for each group? (Shapiro Test)
Yes
No
A1
R1
Do all groups have homoscedasticity? (Barlett Test)
Is average gene expression value for each group same? (Kruskal Wallis Test)
Yes
No
Yes
No
A3
R3
A2
R2
Is average gene expression value for each group same? (ANOVA)
Is average gene expression value for each group same? (Welch ANOVA)
Is average gene expression value for different pair of groups same? (Pairwise Wilcoxon test)
Yes
No
Yes
No
Yes
No
A4
R4
A5
R5
$A2_{i2}$
$R2_{i1}$
Is average gene expression value for different pairs of groups same? (Pairwise t Test)
Is average gene expression value for different pair of groups same? (Pairwise Welch t Test)
Potential Biomarkers
Yes
No
Yes
No
$A4_{K2}$
$R4_{K1}$
$A5_{J2}$
$R5_{J1}$

```r
# Shapiro test
cat("Number of Genes: ",length(which(PShap<alpha)))
```

```
## Number of Genes:  34558
```

34558 out of 54675 hypotheses were rejected i.e. 34558 out of 54675 genes do not come from a normal population.

```r
# Kruskal-Wallis for Normality rejected genes
cat("Number of Genes: ",length(which(KMat[,2]<=alpha)))
```

```
## Number of Genes:  23217
```

23217 genes rejected i.e. 23217 genes are differentially expressed for at least one of the 5 groups.

```r
# Pairwise Wilcoxon on rejected hypotheses.
cat("Gene: ",npbm, "\nNumber of Genes: ",length(KMat[,2]<=alpha))
```

```
## Gene:  1007_s_at
## Number of Genes:  34558
```

“1007_s_at” is the differentially expressed gene obtained after performing pairwise Wilcoxon test.

```r
# Bartlett and ANOVA on normality accepted genes
cat("Number of Genes: ",length(which(PShap>=alpha)))
```

```
## Number of Genes:  20117
```

We will perform Bartlett’s test and thereby ANOVA and Welch ANOVA on these 20117 genes.

```r
cat("Bartlett test gene count\nNumber of Accepted Genes: ", length(baccg),
"\nNumber of Rejected Genes: ", length(brejg), "\n\nWelch ANOVA test gene
count\nNumber of Accepted Genes: ", length(waaccg), "\nNumber of Rejected
Genes: ", length(warejg), "\n\nANOVA test gene count\nNumber of Accepted
Genes: ", length(aaccg), "\nNumber of Rejected Genes: ", length(arejg))
```

```
## Bartlett test gene count
## Number of Accepted Genes:  14097
## Number of Rejected Genes:  6020
##
## Welch ANOVA test gene count
## Number of Accepted Genes:  850
## Number of Rejected Genes:  5170
##
## ANOVA test gene count
## Number of Accepted Genes:  4781
## Number of Rejected Genes:  9316
```

We will perform pairwise t and Welch t test for ANOVA and Welch ANOVA rejected genes respectively.

```
## Gene:  1552257_a_at
## Number of Genes:  5170
```

“1552257_a_at” is the differential;y expressed gene obtained after performing pairwise Welch t-tests.

```
# Pairwise t for Anova Rejected Genes
cat("Gene: ", cvbm, "\nNumber of Genes: ",length(arejg))
```

```
## Gene:  1487_at
## Number of Genes:  9316
```

“1487_at” is the differentially expressed gene obtained after performing pairwise t-tests.

Overview of tests used:

1. **Shapiro Test:** Used to check normality (Failing to reject the null hypothesis suggests that all the groups come from a normal population)
2. **Bartlett Test:** Used to check homoscedasticity (Failing to reject the hull hypothesis suggests that all the groups have homoscedasticity)
3. **Kruskal Wallis Test**, **ANOVA**, **Welch ANOVA**: Used to test if average gene expression value for each group is same. (Rejection of the null hypothesis is desired).
4. **Pairwise Wilcoxon Test**, **Pairwise t test**, **Pairwise Welch t test**: Used to test if average gene expression value for two group is same. (Rejection of the null hypothesis is desired).

```
# Final Biomarkers
biomarkers=c(npbm,vvbm,cvbm)
cat("Biomarkers: ", biomarkers, sep = c("",rep(" , ",3)))
```

```
## Biomarkers: 1007_s_at , 1552257_a_at , 1487_at
```

Based on these inference procedure, we have the following 3 genes as potential biomarkers:

1. 1007_s_at: DDR1 (Discoidin Domain Receptor 1)
2. 1552257_a_at: TTLL12 (Tubulin Tyrosine Ligase-Like 12)
3. 1487_at: CTBP1 (C-terminal Binding Protein 1)

Before proceeding any further, let us try to understand the roles of these genes and assess whether their selection as biomarkers is biologically meaningful.

1. **1007_s_at: DDR1 (Discoidin Domain Receptor 1)** DDR1 is a type of protein found on the surface of cells. It becomes active when it comes into contact with collagen, a major part of our body's connective tissue. Once activated, DDR1 sends signals inside the cell that help control growth and movement of cells. DDR1 is commonly found in increased levels in several types of cancer. Its signaling can promote tumor growth and help cancer cells invade surrounding tissues.

2. **1552257_a_at: TTLL12 (Tubulin Tyrosine Ligase-Like 12)** TTLL12 is part of a family of proteins that help modify microtubules—tiny structures inside cells that act like tracks for moving things around and also help the cell divide. Since it helps with important cell processes like cell division, unusual activity of this gene might play a role in cancer development.

3. **1487_at: CTBP1 (C-terminal Binding Protein 1)** CTBP1 is a protein that helps turn off certain genes by working with other proteins. It plays an important role in deciding whether a cell grows, dies, or stays the same by controlling gene activity behind the scenes. CTBP1 is involved in many cancers as it can silence genes that would normally stop tumor growth or tell damaged cells to die. When CTBP1 is too active, it can help cancer cells survive and multiply.

Thus, it is safe to say that these genes stand as sensible choices for potential biomarkers.

Now that we have identified potential biomarkers, our next task is to study their direction of influence on cancer development. If possible, we would also like to devise a rule or a guideline that can help determine whether an individual is at a risk of brain cancer, and if so, identify which one of these 4 types of brain cancer he/she is most susceptible to.

Suppose, we are given gene expression values for these 3 genes (biomarkers) for an individual. We would first like to check if this individual is susceptible to brain cancer or not. In other words we would like to classify this observation (gene expression value for this person) in one of the 2 groups - (Brain) Cancerous & Non-(Brain)Cancerous. Since we are dealing with 2 groups, we can use logistic regression for the same.

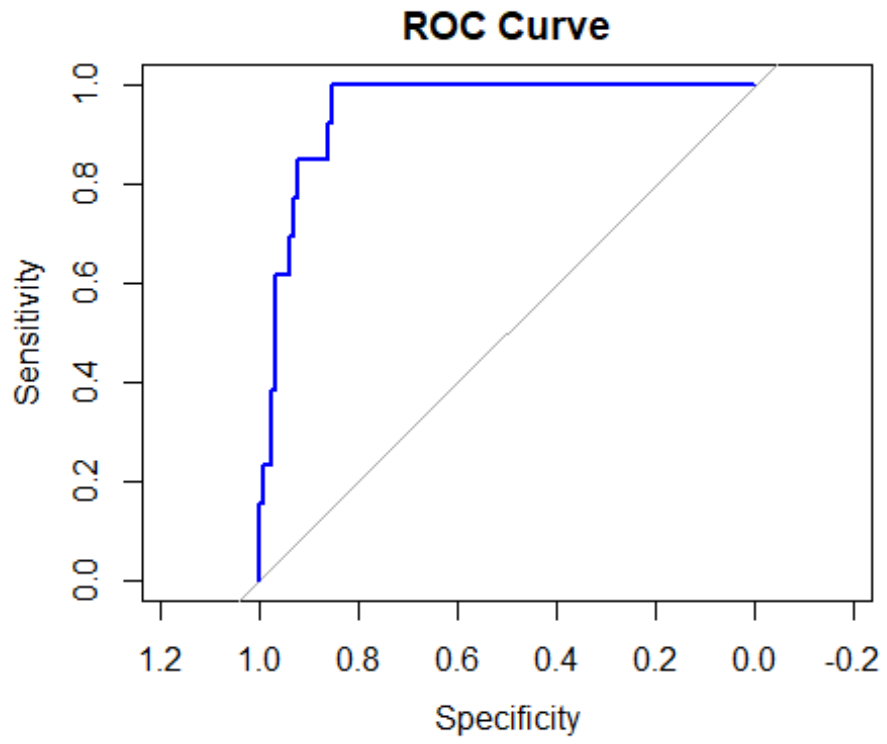
Sensitivity or True Positive Rate is the probability that the model will classify a true positive observation as a positive observation. Whereas, specificity or True Negative Rate is the probability that the model will classify a true negative observation as a negative observation. We use ROC (Receiver Operating Characteristic) curve and AUC to examine if our logistic model is a good fit or not. ROC curve is basically a plot of Sensitivity vs Specificity and AUC is the area under the ROC curve. If our value of AUC is close to 1, then we say that the underlying Logistic model is a good fit.

```
# Logistic Regression for classifying non-cancerous and cancerous groups
model1=glm(y~x1+x2+x3,data=d11,family="binomial")
prob=predict(model1, type = "response")
roc_obj=roc(d11$y, prob)

## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases

plot(roc_obj, col = "blue", main = "ROC Curve")
```




```
cutoff=coords(roc_obj, "best", ret = "threshold")
cat("Area under the curve: ", auc(roc_obj))

## Area under the curve:  0.9500329
```

Thus, we can safely say that our Logistic model is a good fit.

```
# Threshold
cat("Cutoff: ", unlist(cutoff))

## Cutoff:  0.0991748
```

For the classification rule, we select the cutoff value for which both, sensitivity and specificity are high.

Classification rule: We define the classification rule as follows:

$$y = \begin{cases} 0, & \text{if } p_{\underline{x}} < 0.0991748 \\ 1, & \text{if } p_{\underline{x}} \geq 0.0991748 \end{cases}$$

Based on gene expression values of a particular individual, Y = 0 means that individual is likely to have brain cancer, whereas Y = 1 means that the individual is unlikely to develop brain cancer and $p_{\underline{x}}$ is probability that the given individual doesn't have brain cancer.

```
s1=summary(model1)
a11=anova(model1, test = "Chisq")

cat("p values of LRT for checking adequacy of model: \n" ,a11$`Pr(>Chi)`, sep
= c("", rep(" , ", 3)))

## p values of LRT for checking adequacy of model:
## NA , 1.815434e-05 , 0.967659 , 4.621569e-07
```

Observe that, p value corresponding to "1552257_a_at" gene is 0.967659 > 0.05(los). This suggests that "1552257_a_at" gene doesn't contribute significantly to the logistic regression model.

```
model2 = glm(y~x1+x3,data=d11,family="binomial")
s2=summary(model2)
cat("AIC for model with 1552257_a_at gene: ", s1$aic, "\nAIC for model
without 1552257_a_at gene: ",s2$aic)

## AIC for model with 1552257_a_at gene:  48.73056
## AIC for model without 1552257_a_at gene:  50.64774
```

Observe that, after removing "1552257_a_at" gene from our model, AIC (Akaike Information Criterion, a relative measure used to assess model adequacies) increases. This is due to suppression effect. Suppression effect in when a variable that appears unimportant on its own (i.e. not statistically significant), but actually improves the predictive power of the model when included with other variables. Here, we observe suppression effect due to "1552257_a_at" gene.

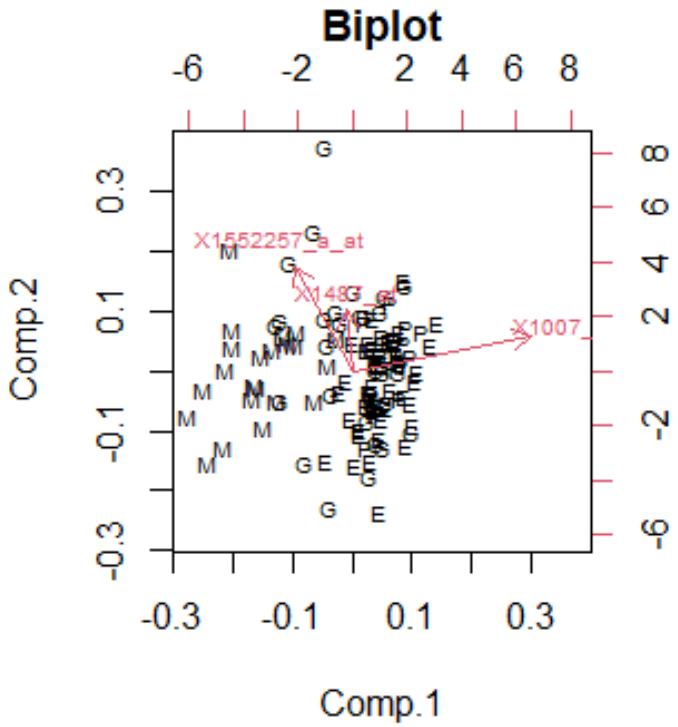
Suppose, we classify this individual in the Cancerous Group. Now that we know that this individual is susceptible to brain cancer, we would like to identify which one these 4 types of brain cancer development is more likely. Principal Component Analysis (PCA) is a technique which may help us in this situation.

```
# Principal Component Analysis
p3$loadings

##
## Loadings:
##                Comp.1 Comp.2 Comp.3
## X1007_s_at      0.951  0.282  0.127
## X1552257_a_at -0.308  0.826  0.472
## X1487_at               0.488 -0.872
##
##                 Comp.1 Comp.2 Comp.3
## SS loadings      1.000  1.000  1.000
## Proportion Var   0.333  0.333  0.333
## Cumulative Var   0.333  0.667  1.000
```

Note that the first 2 principal components are explaining about 66.7% variation in the data. Genes “1007_s_at”, “1552257_a_at”, “X1487_at” are contributing the most to the first, second and third principal components respectively.

```
# Principal Component Analysis - Biplot
biplot(p3, xlabs = k1, cex = 0.6, main = "Biplot")
```



It seems that it is possible to differentiate group M i.e. Medulloblastoma with other cancer groups using PCA. We will take the two principal components, PC1 and PC2 as regressors and use logistic regression to find the best fitted line that would separate group M from the other cancer groups.

```
# Logistic regression
m1235=glm(group ~ PC1 + PC2, data = df, family = "binomial", maxit = 100)
summary(m1235)

##
## Call:
## glm(formula = group ~ PC1 + PC2, family = "binomial", data = df,
##     maxit = 100)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.3924     0.7613  -4.456 8.35e-06 ***
## PC1          -4.3066     0.9367  -4.598 4.27e-06 ***
## PC2          -0.3701     0.7182  -0.515    0.606
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 113.106  on 116  degrees of freedom
## Residual deviance:  33.842  on 114  degrees of freedom
## AIC: 39.842
##
## Number of Fisher Scoring iterations: 7
```

```
prob1=predict(m1235, type = "response")
roc_obj1=roc(df$group, prob1)
```

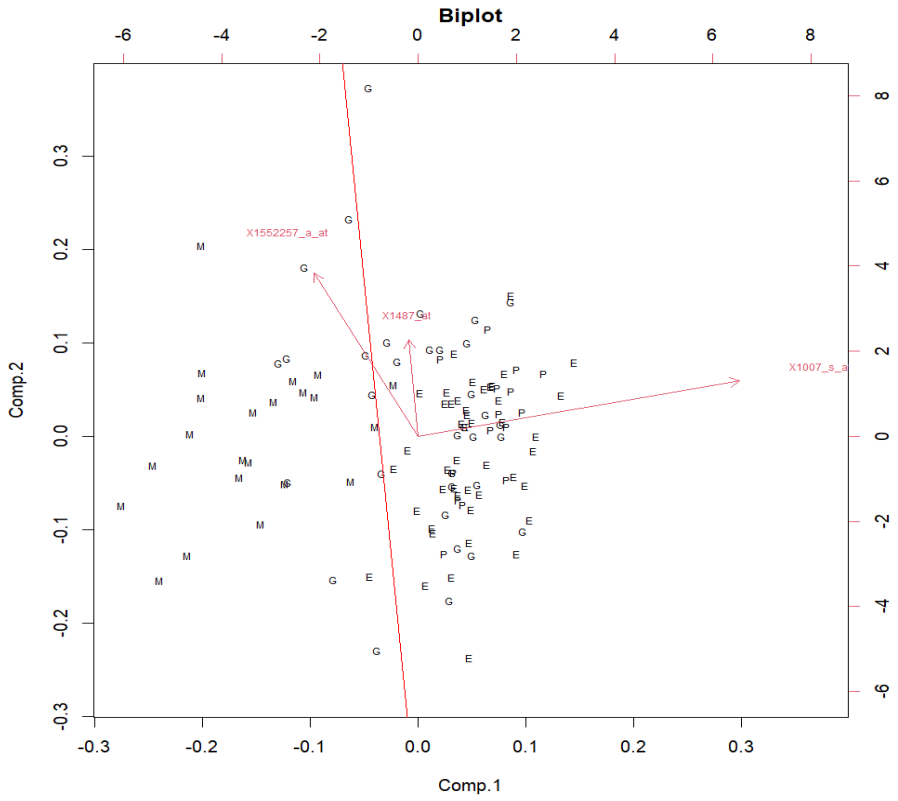
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
cat("Area under the curve: ", auc(roc_obj1))
```

```
## Area under the curve:  0.9794258
```

AUC of 0.97 suggests that our logistic model is a good fit.



Classification Rule: In order to check if a particular individual is likely to develop Medulloblastoma, we calculate the PC1 and PC2 values (using that individual's gene

expression values) and plot it on the above biplot. If the point corresponding to this individual lies on the side of Medulloblastoma, then we say that the individual is likely to develop Medulloblastoma, else we say that the individual is unlikely to develop Medulloblastoma.

However, observe that there is some misclassification possible.

Future Scope

1) More powerful pairwise tests like Tuckey's test may yield better results.
2) Integrating techniques like SVM (Support Vector Machine) with PCA may provide us with more clear classification rules.