Stat 437 Cancer Classification

Audrey Kimball (11659795)

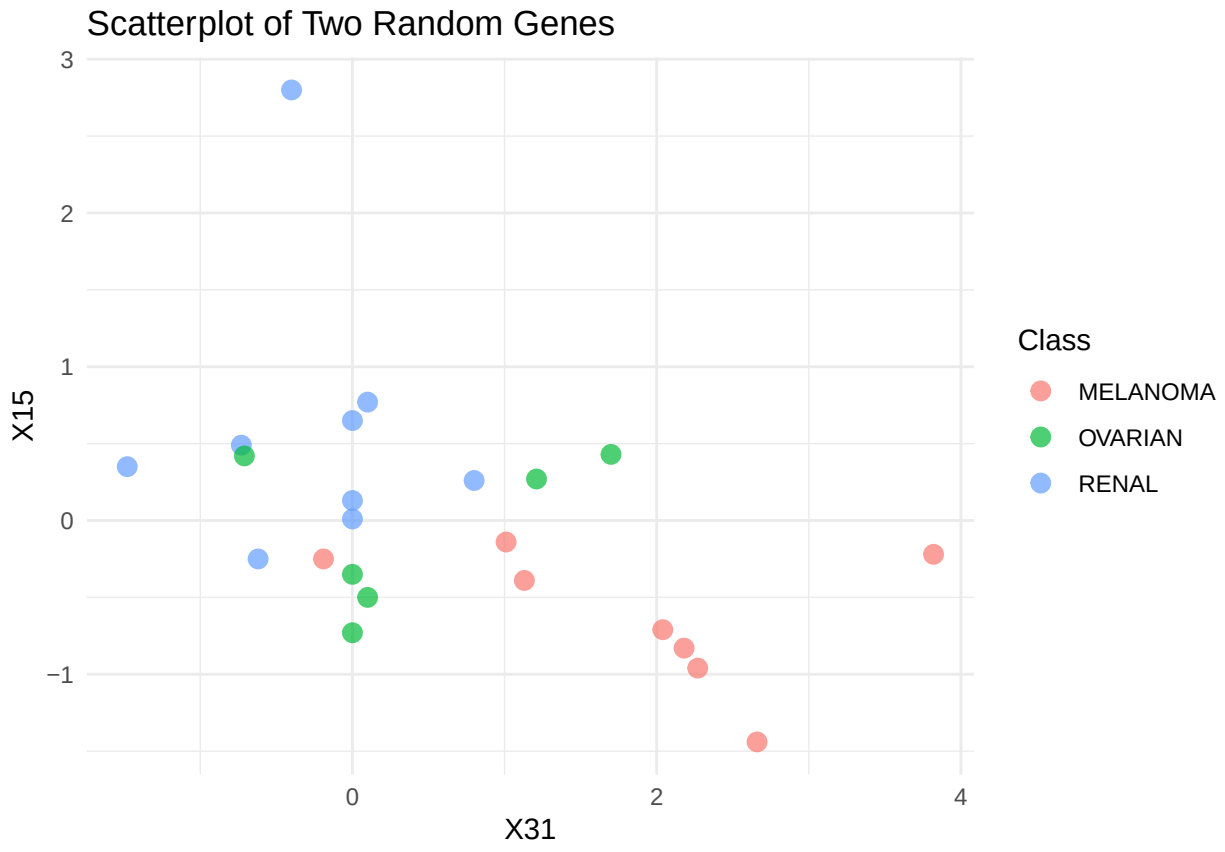
```
#load libraries
library(knitr)
library(ElemStatLearn)
library(MASS)
library(klaR)
library(reshape2)
library(ggplot2)
opts_chunk$set(fig.align="center",tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

9. We will use the human cancer microarray data that were discussed in the lectures and are provided by the R library `ElemStatLearn` (available at https://cran.r-project.org/src/contrib/Archive/ElemStatLearn/). Pick 3 cancer types "MELANOMA", "OVARIAN" and "RENAL", and randomly select the same set of 60 genes for each cancer type. Please use `set.seed(123)` for the whole of this exercise. Your analysis will be based on observations for these genes and cancer types.

```
#set seed
set.seed(123)

#get data dimensions
samples <- dim(nci)[2]
genes <- dim(nci)[1]

#randomly select 60 genes
rand_select <- sample(1:genes, size = 60, replace = FALSE)

#select cancer types
cancer_types <- colnames(nci) %in% c("MELANOMA", "OVARIAN", "RENAL")
check_select <- which(cancer_types)

#create subset
nci_sub <- nci[rand_select, check_select]
colnames(nci_sub) <- colnames(nci)[check_select]

#transpose and adding class labels
cancer_data <- data.frame(t(nci_sub))
cancer_data$Class <- colnames(nci_sub)
rownames(cancer_data) <- NULL
```

9.1) Pick 2 features and visualize the observations using the 2 features. Do you think it is hard to classify the observations based on the amount of overlap among the 3 neighborhoods of observations in each of the 3 classes? Here "a neighborhood of observations in a class" is a "open disk that contains the observations in the class".

```
#reset seed because for some reason without it the genes keep changing even with
#seed set above
set.seed(123)

#randomly pick 2 genes
gene_cols <- sample(1:60, size = 2, replace = FALSE)
gene1 <- colnames(cancer_data)[gene_cols[1]]
gene2 <- colnames(cancer_data)[gene_cols[2]]

#plot
ggplot(cancer_data, aes(x = .data[[gene1]], y = .data[[gene2]], color = Class)) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "Scatterplot of Two Random Genes", x = gene1, y = gene2) +
  theme_minimal()
```



Given the overlap in the scatter plot, it would be difficult to perfectly classify the observations using only two genes. The open disks around each class cluster are not fully separate, suggesting that these two genes are not sufficient for accurate classification.

9.2) Apply LDA and report the classwise error rate for each cancer type.

```
#apply lda
lda_fit <- lda(Class ~ ., data = cancer_data )
```

```
## Warning in lda.default(x, grouping, ...): variables are collinear
```

```
#predict using fitted model
lda_pred <- predict(lda_fit, cancer_data)

#True and predicted labels
true_labels_92 <- cancer_data$Class
pred_labels_92 <- lda_pred$class

#confusion matrix
conf_matrix_92 <- table(Predicted = pred_labels_92, Actual = true_labels_92)
print(conf_matrix_92)
```

```
##           Actual
## Predicted  MELANOMA OVARIAN RENAL
##   MELANOMA        6       0     0
##   OVARIAN         0       5     0
##   RENAL           2       1     9
```

```
#classwise error rates
class_totals_92 <- colSums(conf_matrix_92)
class_errors_92 <- class_totals_92 - diag(conf_matrix_92)
class_error_rate_92 <- round(class_errors_92 / class_totals_92, 4)

#output error rates
class_error_rate_92
```

```
## MELANOMA  OVARIAN    RENAL
##   0.2500   0.1667   0.0000
```

Melanoma: 0.2500 Ovarian: 0.1667 Renal: 0.0000

9.3) Use the library `klaR`, and apply regularized discriminant analysis (RDA) by setting the arguments `gamma` and `lambda` of `rda{klaR}` manually so that the resulting classwise error rate for each cancer type is zero.

```
#apply rda

rda_fit <- rda(Class ~ ., data = cancer_data, gamma = 0.05, lambda = 0.01)

#predict using rda model
```

```r
rda_pred <- predict(rda_fit, cancer_data)

#true predicted labels
true_labels_93 <- cancer_data$Class
pred_labels_93 <- rda_pred$class

#confusion matrix
conf_matrix_93 <- table(Predicted = pred_labels_93, Acutual = true_labels_93)
print(conf_matrix_93)
```

```
##           Acutual
## Predicted  MELANOMA OVARIAN RENAL
##   MELANOMA        8       0     0
##   OVARIAN         0       6     0
##   RENAL           0       0     9
```

```r
#classwise error rates
class_totals_93 <- colSums(conf_matrix_93)
class_errors_93 <- class_totals_93 - diag(conf_matrix_93)
class_error_rate_93 <- round(class_errors_93 / class_totals_93, 4)

class_error_rate_93
```

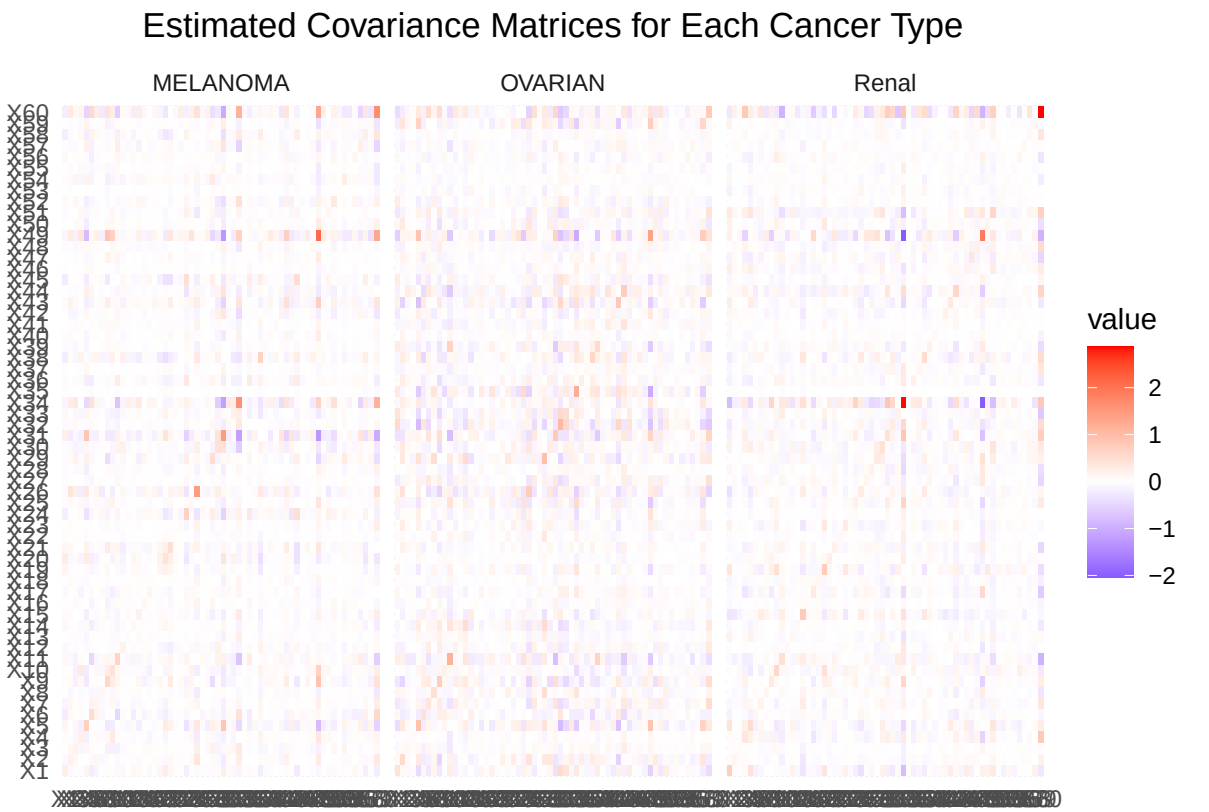
```
## MELANOMA  OVARIAN    RENAL
##        0        0        0
```

Melanoma: 0 Ovarian: 0 Renal: 0

9.4) Obtain the estimated covariance matrices from the RDA and visualize them using the same strategy in Example 3 in "LectureNotes5c_notes.pdf". What can you say about the degree of dependence among these genes for each of the three cancer types? (Hint and caution: the class labels "MELANOMA", "OVARIAN" and "RENAL" will be ordered alphabetically by R. So, you need to keep track on which estimated covariance matrix is for which class. Otherwise, you will get wrong visualization.)

```r
#extract covariance matrices from rda
sigma_melan <- rda_fit$covariances[, , 1] #melanoma
sigma_ovar <- rda_fit$covariances[, , 2] #ovarian
sigma_ren <- rda_fit$covariances[, , 3] #renal

#melt matrices
melt_melan <- melt(sigma_melan)
melt_ovar <- melt(sigma_ovar)
melt_ren <- melt(sigma_ren)

#add cancer type label
```

```r
melt_melan$Cancer <- "MELANOMA"
melt_ovar$Cancer <- "OVARIAN"
melt_ren$Cancer <- "Renal"

#combine into one data frame
sigma_all <- rbind(melt_melan, melt_ovar, melt_ren)
sigma_all$Cancer <- factor(sigma_all$Cancer)

#plot the covariance matrices
ggplot(data = sigma_all, aes(x = Var1, y = Var2, fill = value)) + geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint =  0) +
  facet_grid(~ Cancer) + xlab("") + ylab ("") +
  ggtitle("Estimated Covariance Matrices for Each Cancer Type") +
  theme_minimal() + theme(plot.title = element_text(hjust = 0.5))
```



The covariance heatmaps show that melanoma has the strongest gene-gene dependence, with clear positive and negative covariance patterns. Ovarian displays weaker, more scattered covariances, suggesting the genes act more independently. Renal shows moderate dependence with some structured patterns. This indicates that gene relationships can differ across cancer types, which may impact classification.