Project
Transplant kidney rejection
High Dimensional Data Analysis

Jan Alexander* Annabel Vaessens† Steven Wallaert‡

8/4/2020

```
load('RejectionStatus.rda')
load('X_GSE21374.rda')
dim(RejectionStatus)
```

```
## [1] 282   2
```

```
dim(X_GSE21374)
```

```
## [1] 54675   282
```

```
GeneExpression <- scale(t(X_GSE21374))

GeneExpression <-
  GeneExpression[order(as.numeric(row.names(GeneExpression))), ]
```

```
## Warning in order(as.numeric(row.names(GeneExpression))): NAs introduced by
## coercion
```

```
RejectionStatus <-
  RejectionStatus[order(as.numeric(RejectionStatus$Patient_ID)), ]

all.equal(row.names(GeneExpression), as.character(RejectionStatus$Patient_ID))
```

```
## [1] TRUE
```

1 Introduction

The data is loaded as presented in the assignment.

Three research questions are defined:

*jan.alexander@ugent.be
†annabel.vaessens@vub.be
‡steven.wallaert@ugent.be

- How do the 54675 genes vary in terms of their gene expression levels? Is the variability associated with kidney rejection? (only to be answered in a data explorative manner).
- Which genes are differentially expressed between the two kidney rejection groups? You must control the False Discovery Rate at 10%.
- Can the kidney rejection be predicted from the gene expressions? What genes are most important in predicting the kidney transplant rejection? How well does the prediction model perform in terms of predicting rejection status?

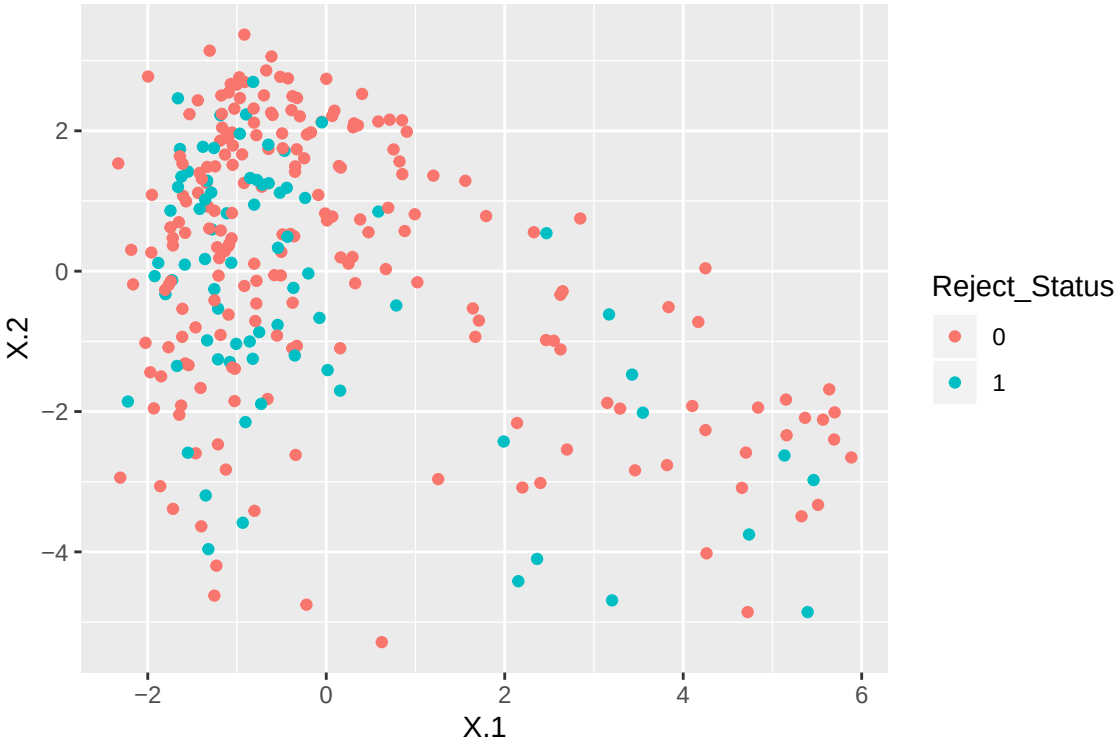
2 Data exploration

In the complete dataset, 27 % of the transplanted kidnesy were rejected.

```
Gen_spc <- PMA::SPC(GeneExpression, K = 2, sumabsv = 2)
```

```
## 1234567891011121314151617181920
## 1234567891011121314151617181920
```

```
Uk <- Gen_spc$u ; Dk <- diag(Gen_spc$d)
Zk <- data.frame(X = Uk %*% Dk, Patient_ID = row.names(GeneExpression))
Zk <- merge(Zk, RejectionStatus, by = 'Patient_ID') %>%
  mutate(Reject_Status = as.factor(Reject_Status))
ggplot(data = Zk, aes(x = X.1, y = X.2, col = Reject_Status)) +
  geom_point()
```




```
rm(Zk, Dk, Uk, X_GSE21374, Gen_spc)
```

Unfortunately, naive sparce principle component analysis does not seem to work well.

Partial least squares:

```
# GeneExpression_comb <-
#  list(genes = as.matrix(GeneExpression), Rejection = as.matrix(RejectionStatus$Reject_Status))
# pls_model <- pls::plsr(genes ~ Rejection, data = GeneExpression_comb, validation = "CV")
```

3 Differentiating genes between kidney acceptance and rejection

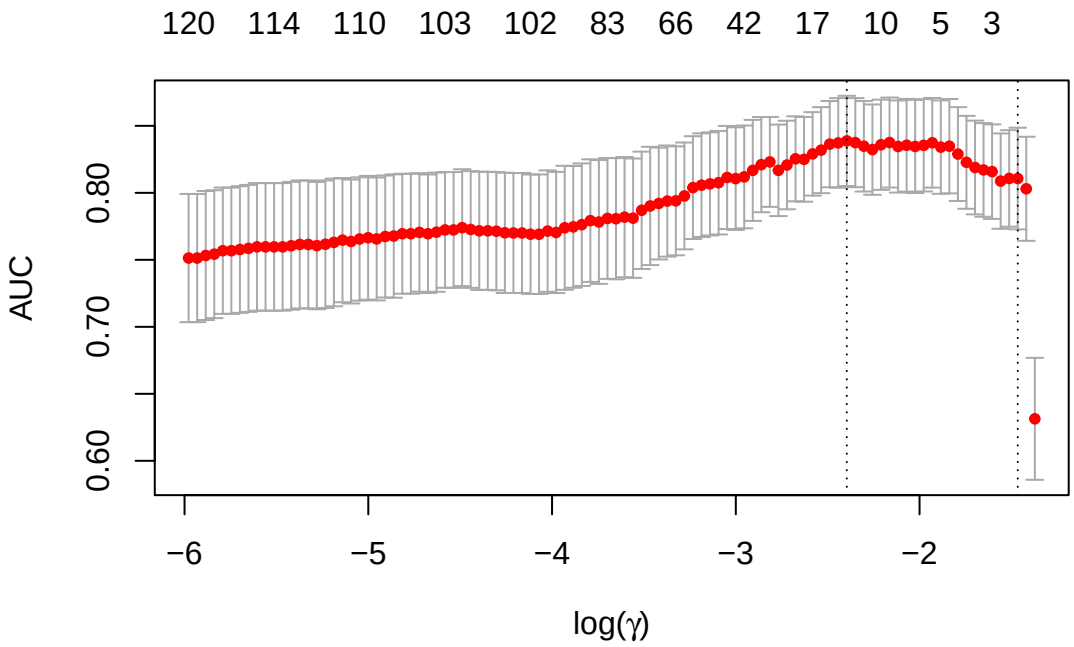
4 Prediction of kidney transplant rejection

```
ind_train <-
  sample(seq_len(nrow(RejectionStatus)), size = floor(nrow(RejectionStatus) * 0.80))

Y_train <- as.matrix(RejectionStatus[ind_train, 'Reject_Status'])
X_train <- as.matrix(GeneExpression[ind_train,])
Y_test <- as.matrix(RejectionStatus[-ind_train, 'Reject_Status'])
X_test <- as.matrix(GeneExpression[-ind_train,])
```

4.1 Lasso regression

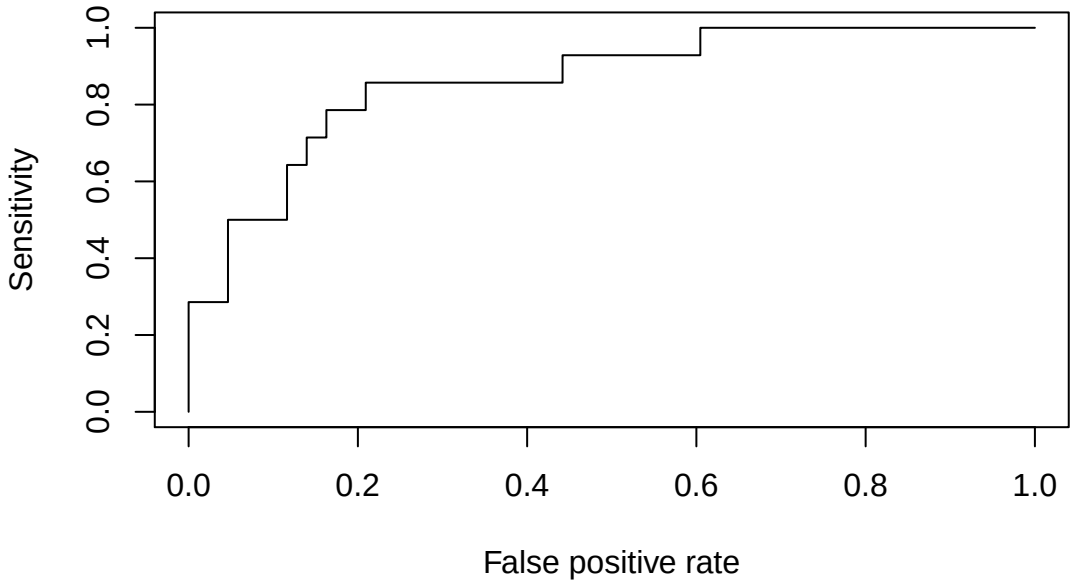
```
m.cv <-
  cv.glmnet(
    x = X_train,
    y = Y_train,
    alpha = 1,
    family = 'binomial',
    type.measure = "auc"
  )
plot(m.cv, xlab = TeX(" $ log(\\gamma ) $ "))
```



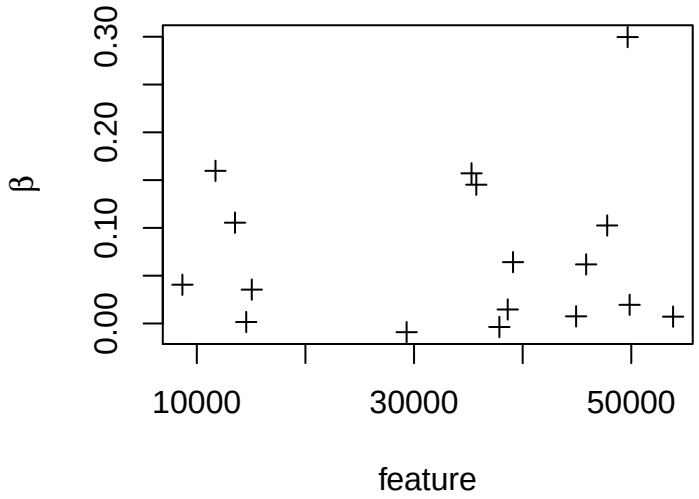
In the figure above, one can see that for γ equal to 0.0910832 , the area under the curve (AUC) is maximal for the train dataset based on a 10-fold cross-validation over the train dataset.

The ROC curve, estimated with the cross-validation dataset, is shown below:

```
m <- glmnet(
  x = X_train,
  y = Y_train,
  alpha = 1,
  family = 'binomial',
  lambda = m.cv$lambda.min
)
pred_m <-
  prediction(predict(
    m,
    newx = X_test,
    type = 'response'
  ),
  Y_test)
perf <- performance(pred_m, 'sens', 'fpr')
plot(perf)
```

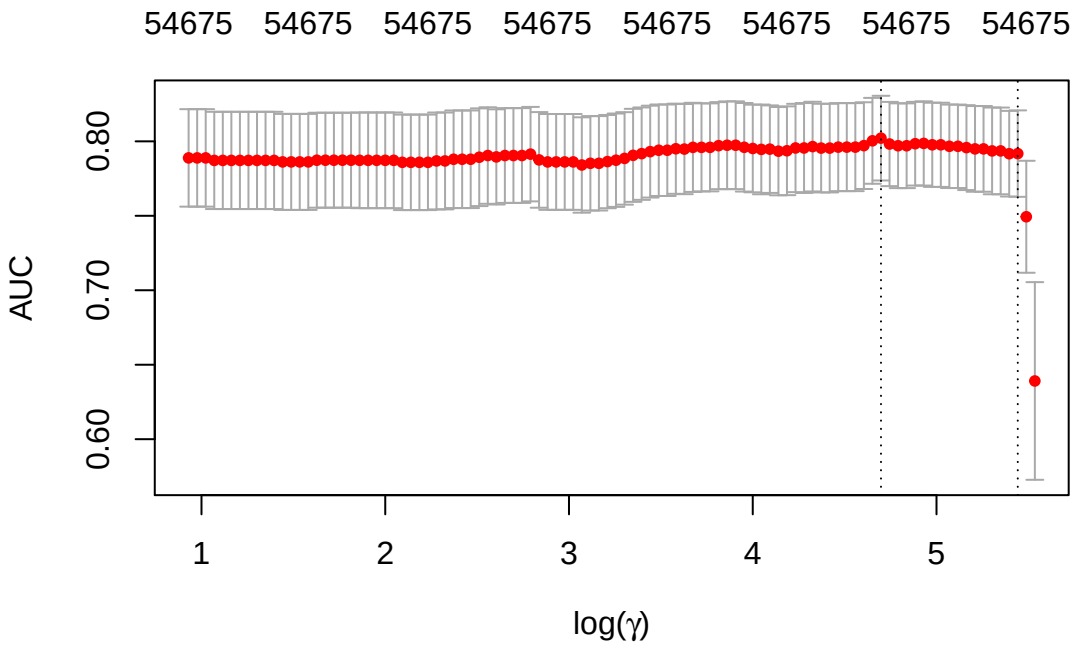


This model uses 17 of the features. This is a considerable dimensional reduction. This is illustrated below. This figure shows the loadings of the 17 selected values.



4.2 Ridge regression

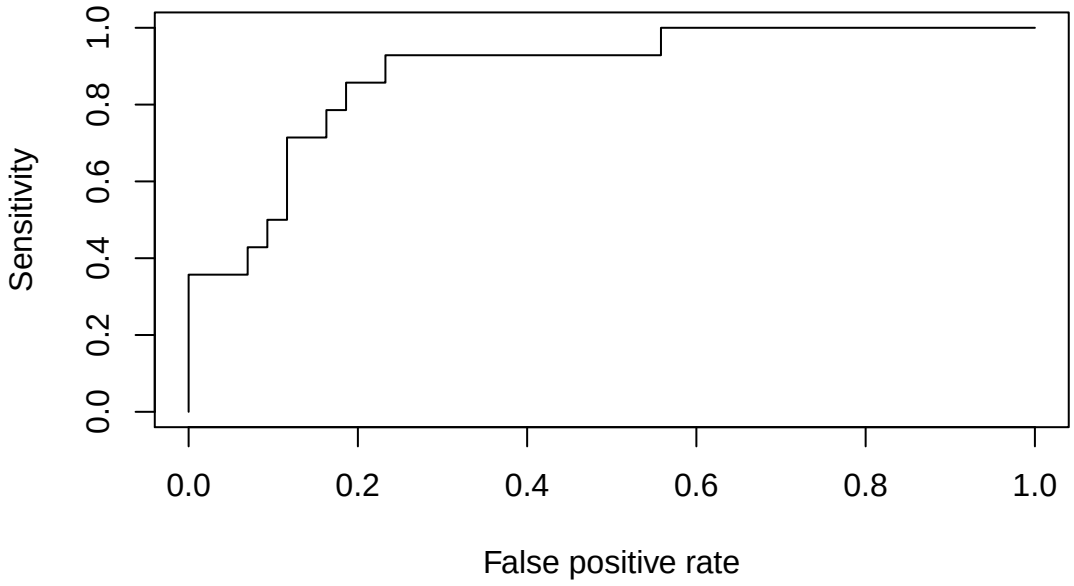
```
m.cv <-
  cv.glmnet(
    x = X_train,
    y = Y_train,
    alpha = 0,
    family = 'binomial',
    type.measure = "auc"
  )
plot(m.cv, xlab = TeX(" $ log(\\gamma ) $ "))
```



In the figure above, one can see that for γ equal to 109.7100096 , the area under the curve (AUC) is maximal for the train dataset based on a 10-fold cross-validation over the train dataset.

The ROC curve, estimated with the cross-validation dataset, is shown below:

```
m <- glmnet(
  x = X_train,
  y = Y_train,
  alpha = 0,
  family = 'binomial',
  lambda = m.cv$lambda.min
)
pred_m <-
  prediction(predict(
    m,
    newx = X_test,
    type = 'response'
  ),
  Y_test)
perf <- performance(pred_m, 'sens', 'fpr')
plot(perf)
```


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate