Jacob Harris

Sex Determination from Skeletal Remains

Abstract

Measurements of thoracic vertebrae from 28 skeletons were taken and used to determine the sex of 7 unknown individuals. Using data imputation, principal components analysis, and models such as logistic regression, linear discriminant analysis, and quadratic discriminant analysis, the sex of the individuals is predicted to be either male or female. After using the stated methodology, 4 of the skeletons were considered to be female and 3 were considered to be male.

Introduction

This project comes from a fellow graduate student at Texas A&M - Corpus Christi, Leah Fuentes. Near Waco, numerous skeletons were dug up using machinery, which led to the mistreatment of many individual remains. These remains were sent to the forensic science department for study. As a result of the mishandling, the sex of several of these individual remains were unidentified and needed further analysis. In this project, data from 28 individuals are used, with only 21 identified as male or female and 7 remaining as unknown. From the known sample, 14 were found to be female and 7 were found to be male.

Using 6 measurements taken from the thoracic vertebrae, predictions for sex were made. Thoracic vertebrae are found along the base of the neck to the bottom of the ribs [1]. These vertebrae have the least flexion/extension ability of the spine [2]. Some of the functions of thoracic vertebrae include protecting the spinal cord, providing attachment for ribs, and support of the chest and abdomen [3].

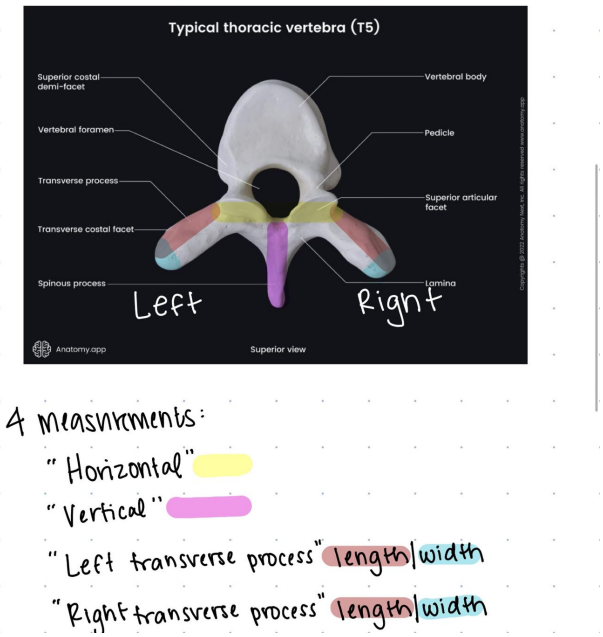


Figure 1: Thoracic Vertebrae Measurements

The picture above shows the measurements taken for each vertebra. Each measurement is in mm. The skeletons used in this project had all 12 thoracic vertebrae intact, with the exception of one skeleton missing one single vertebra.

Methodology

The first task in this project involved using MANOVA to determine if there are differences of the 6 measurements among the two sexes and twelve vertebrae. Once done, data imputation was used to replace missing values followed by principal component analysis. The first two components resulting from PCA were used as predictors in a multiple logistic regression, linear discriminant analysis, and quadratic discriminant analysis.

Data

The data was provided by Leah Fuentes, another graduate student whose project this data comes from. The 6 different measurements for 12 vertebrae for 28 individuals gives a total of 2016 different measurements taken. In this project, the data is structured multiple ways. In order to perform a MANOVA, the following data was used;

```
head(Known)
```

```
# A tibble: 6 x 9
  Individual Vertebrae Horizontal Vertical lTPlength lTPwidth rTPlength rTPwidth
  <fct>      <fct>          <dbl>    <dbl>     <dbl>    <dbl>     <dbl>    <dbl>
1 MB34       T1              50.0     27.8      NA       7.56      13.6     15
2 MB34       T2              33.8     27.9      18.5     12.8      19.6     13.4
3 MB34       T3              29.3     27.3      17.6     11.3      18.6     13.2
4 MB34       T4              26.5     27.9      21.4     12.2      17.9     11.8
5 MB34       T5              29.9     27.4      18.4     12.8      18.9     12.6
6 MB34       T6              26.7     27.7      20.3     11.0      18.2     12.0
# i 1 more variable: Sex <fct>
```

This data has each of the 12 vertebrae as individual records. This allows a two-way MANOVA to use Sex and Vertebrae as factors. In order to make predictions, a single entry for each individual was needed. The following data is what is used in making predictions;

```
head(indiv)
```

```
# A tibble: 6 x 74
  Sex   Individual Horizontal.T1 Horizontal.T10 Horizontal.T11 Horizontal.T12
  <fct> <fct>              <dbl>          <dbl>          <dbl>          <dbl>
1 <NA>  MB125               40.1           30.8           36.5           35.5
2 <NA>  MB130               41             32.0           33.1           NA
3 <NA>  MB151               46.8           33.6           39.3           44.3
4 <NA>  MB47                48.5           32.4           35.6           39.4
5 <NA>  MB49                44.8           32.2           36.3           38.7
6 <NA>  MB65                51.7           36.2           43.8           46.9
# i 68 more variables: Horizontal.T2 <dbl>, Horizontal.T3 <dbl>,
#   Horizontal.T4 <dbl>, Horizontal.T5 <dbl>, Horizontal.T6 <dbl>,
#   Horizontal.T7 <dbl>, Horizontal.T8 <dbl>, Horizontal.T9 <dbl>,
#   Vertical.T1 <dbl>, Vertical.T10 <dbl>, Vertical.T11 <dbl>,
#   Vertical.T12 <dbl>, Vertical.T2 <dbl>, Vertical.T3 <dbl>,
#   Vertical.T4 <dbl>, Vertical.T5 <dbl>, Vertical.T6 <dbl>, Vertical.T7 <dbl>,
#   Vertical.T8 <dbl>, Vertical.T9 <dbl>, Left.TP.Length.T1 <dbl>, ...
```

The first 7 rows contain individuals with unknown Sex. This data frame shows the 72 different types of measurement taken, 6 measurements for 12 vertebrae. As a result, the number of predictors (72) is much greater than the sample size (21 known, 28 total).

Analysis

The following libraries are used in the analysis of the given data.

```
library(readxl)
library(tidyverse)
library(MVN)
library(GGally)
library(biotools)
library(softImpute)
```

MANOVA

The first MANOVA is set up by the following equation;

$$Y_{ijk} = \beta + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \epsilon_{ijk}$$

where α_i is the term associated with either male or female, γ_j corresponds to one of the 12 vertebrae, and ϵ_{ijk} is the error for the k-th measurement. The term $\alpha\gamma_{ij}$ represents an interaction between Sex and Vertebrae. The response Y_{ijk} is a vector of the different measurements taken for each vertebrae for an individual. The following table shows the results of the MANOVA.

```
man.fit <- manova(cbind(Horizontal,Vertical,lTPlength,
                        lTPwidth,rTPlength,rTPwidth)~Sex*Vertebrae,data=Known)
summary(man.fit, test = 'Wilks')
```

```
               Df   Wilks approx F num Df den Df Pr(>F)
Sex             1 0.54789  29.7066      6  216.0 <2e-16 ***
Vertebrae      11 0.04551  13.7496     66 1161.2 <2e-16 ***
Sex:Vertebrae  11 0.79321   0.7785     66 1161.2 0.9021
Residuals     221
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table above, we can see the interaction term is not significant. This would lead to a second MANOVA with the following equation

$$Y_{ijk} = \beta + \alpha_i + \gamma_j + \epsilon_{ijk}$$

The MANOVA from this equation omits the interaction from before. The table below shows the output from the subsequent MANOVA.

```
man.fit2 <- manova(cbind(Horizontal,Vertical,lTPlength,
                        lTPwidth,rTPlength,rTPwidth)~Sex+Vertebrae,data=Known)
summary(man.fit2, test = 'Wilks')
```

```
           Df   Wilks approx F num Df den Df    Pr(>F)
Sex         1 0.55352   30.518      6  227.0 < 2.2e-16 ***
Vertebrae  11 0.04812   14.105     66 1220.1 < 2.2e-16 ***
Residuals 232
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From both MANOVA tables, Sex and Vertebrae are found to be significant in relation to the measurements taken from each vertebrae. Post-hoc tests were not run to determine which vertebrae were found to be different.

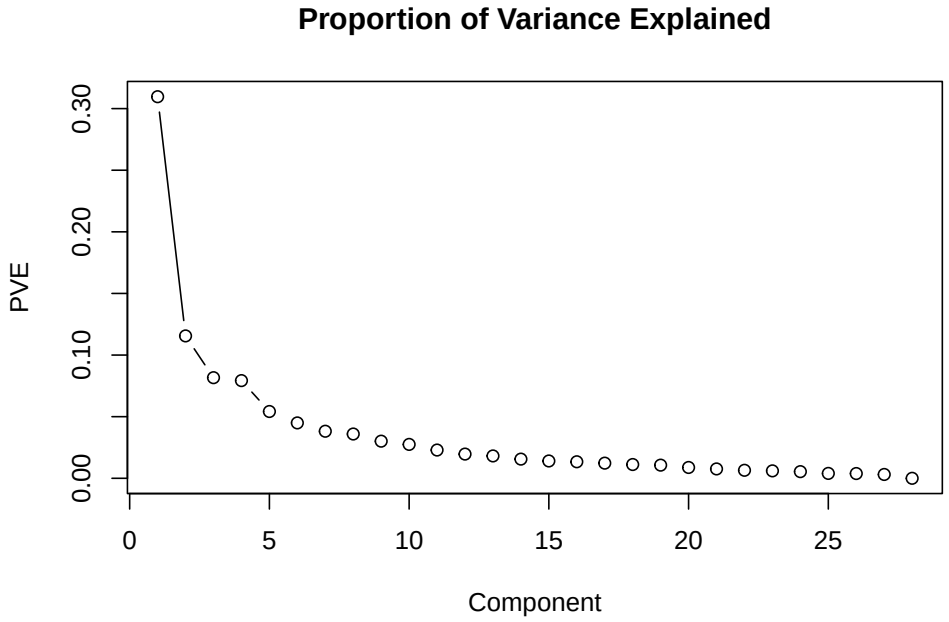
PCA

In order to run PCA, data imputation was used to fill in missing values in the 28 row, 74 column data set. The imputation was done using the 'softImpute' package as shown below. The code below also performs the principal component analysis

```
set.seed(1335)
imp <- softImpute(as.matrix(indiv[-c(1,2)]))
indiv.comp <- as.data.frame(complete(as.matrix(indiv[-c(1,2)]),imp))
indiv.comp$Sex <- indiv$Sex
indiv.comp$Individual <- indiv$Individual
pc.out <- prcomp(indiv.comp[-c(73,74)],scale. = TRUE)
pca.comp <- data.frame(pc.out$x)
pca.comp$Sex <- indiv$Sex
pca.comp$Individual <- indiv$Individual
pve <- pc.out$sdev^2 / sum(pc.out$sdev^2)
```

The plot shows the variances explained by each principal component.

```
plot(pve,type = 'b', main = 'Proportion of Variance Explained',
     xlab = 'Component', ylab = 'PVE')
```



From the scree plot above, the first 3 to 5 components could reasonably be used in models to predict Sex. In the following models, only the first 2 components are used. As a result of PCA, 72 predictors is reduced to only 2. The first component explains 31% of the variance in the data used in PCA. Combining the first two components gives 43% of the variance.

We can also look at the coefficients of the first two principal components to see what variables influence them. To do so, we can look at the greatest absolute values in each loading.

```
sort(abs(pc.out$rotation[,1]),decreasing = T)[1:10]
```

```
 Left.TP.Width.T5  Left.TP.Width.T3  Left.TP.Width.T4     Horizontal.T3
        0.1815766         0.1730041         0.1716748         0.1630511
Right.TP.Width.T5 Right.TP.Width.T4  Left.TP.Width.T6       Vertical.T7
        0.1629101         0.1607222         0.1584009         0.1579753
      Vertical.T1 Right.TP.Width.T3
```

```
0.1516597         0.1484722
```

The code output above shows the magnitude of each coefficient for the 10 largest absolute values in the first principal component. These variables are the most important in calculating the first principal component. As seen above, most of these variables correspond to TP Width measurements for different vertebrae.

```
sort(abs(pc.out$rotation[,2]),decreasing = T)[1:10]
```

```
 Right.TP.Width.T2        Vertical.T4        Vertical.T3   Left.TP.Width.T7
         0.2444582          0.2232226          0.2192700          0.2169360
       Vertical.T6        Vertical.T5        Vertical.T8 Right.TP.Length.T7
         0.1961512          0.1926996          0.1881701          0.1853647
 Right.TP.Width.T6   Left.TP.Width.T8
         0.1852916          0.1794489
```

The code output above looks at the 10 largest absolute values for calculating the second principal component. The second principal component uses mainly Vertical and some TP Width measurements in creating the second component.

Logistic Regression

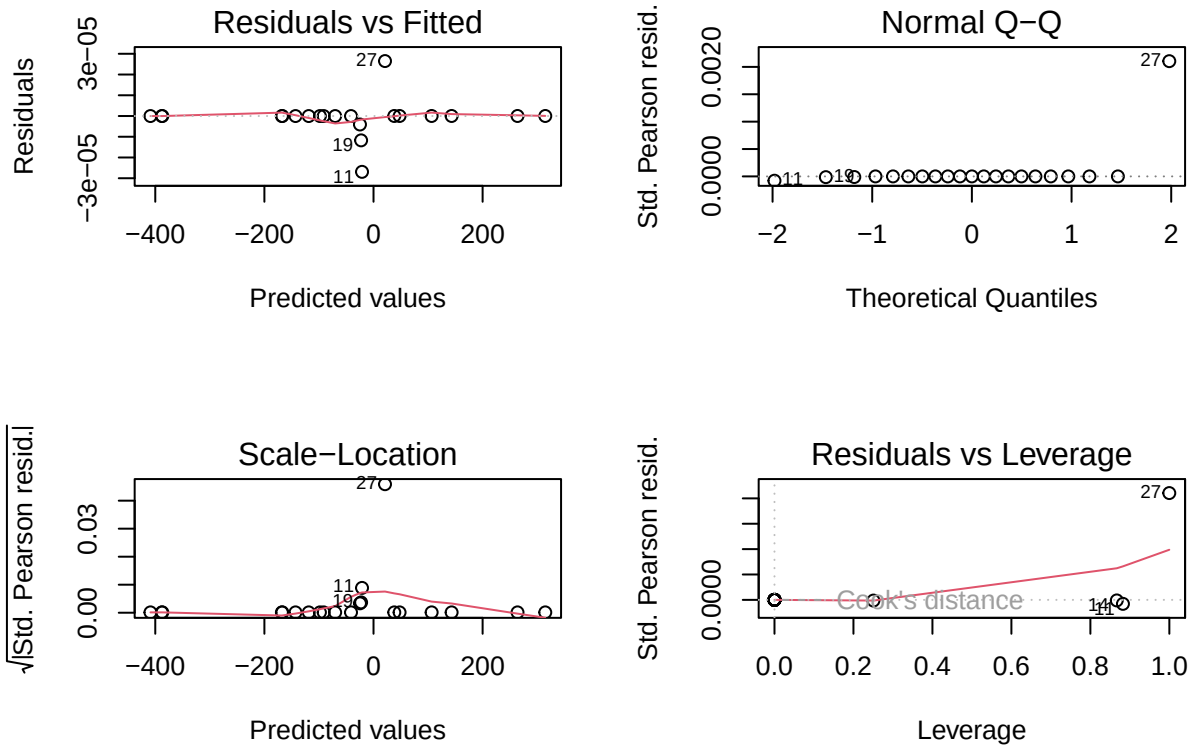
The first model used to predict Sex is a multiple logistic regression with the first 2 components from PCA. Originally the first 3 were used, but the model was found to be unstable. The following code shows the model creation and summary.

```
log.fit.pca <- glm(Sex~PC1+PC2, data = pca.comp,family = 'binomial')
log.pred <- rep('F',28)
log.pred[predict(log.fit.pca,newdata = pca.comp[c(1,2)], type = 'response')>0.5] <- 'M'
log.pred <- data.frame(Individual <- pca.comp$Individual,
                       Prediction <- log.pred)
summary(log.fit.pca)
```

```
Call:
glm(formula = Sex ~ PC1 + PC2, family = "binomial", data = pca.comp)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-3.807e-05  -2.100e-08  -2.100e-08   2.100e-08   3.746e-05

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -58.47  127556.01   0.000    1.000
PC1             39.20   29627.77   0.001    0.999
PC2            -34.28   37165.40  -0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.6734e+01  on 20  degrees of freedom
Residual deviance: 3.1625e-09  on 18  degrees of freedom
  (7 observations deleted due to missingness)
AIC: 6

Number of Fisher Scoring iterations: 25
```

The following plot shows residuals and leverage points for the logistic regression above.

```
par(mfrow=c(2,2))
plot(log.fit.pca)
```



We can see point 27 is a large residual and considered to be a high leverage point. This model could also be considered unstable, as points 8-21 are Female, point 11 and all points to the left in the top-left plot, and 22-28 are Male, point 27 and all points to the right in the top-left plot. The top-right plot shows a Q-Q residual plot comparing the residuals from the model to a Normal sample. The only point in this plot that causes concern would be point 27. The bottom-right graph shows leverage points for the logistic regression model. Point 27 is seen to have the highest leverage among all points.

```
table(log.pred[8:28,]$Prediction....log.pred,pca.comp[8:28,]$Sex,
      dnn = c('Predicted','Observed'))
```

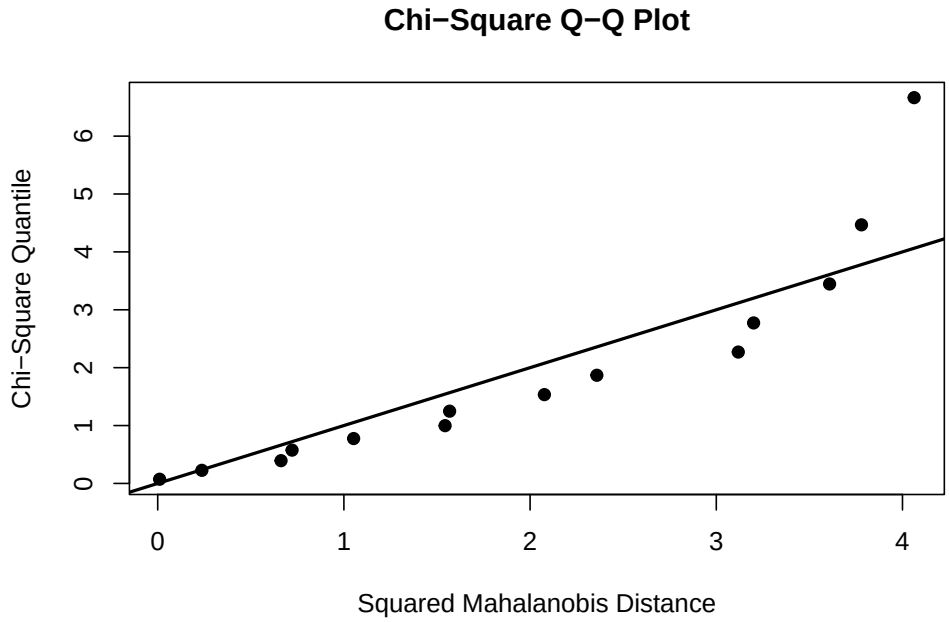
```
         Observed
Predicted  F  M
        F 14  0
        M  0  7
```

As we can see from the confusion matrix, there is no error in predicting the known individuals.

LDA

The next model used is a linear discriminant analysis. We can check the assumptions that the two principal components come from a multivariate normal distribution with a shared covariance matrix. The following results show the tests for multivariate Normality.

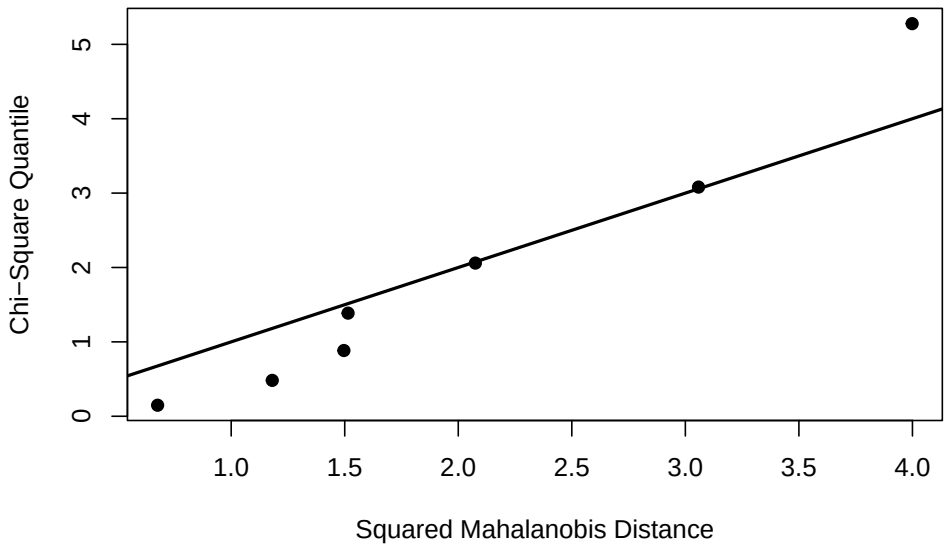
```
mvn(pca.comp[8:21,1:2],
    multivariatePlot = 'qq',mvnTest = 'mardia', univariateTest = 'SW')
```




```
$multivariateNormality
             Test         Statistic           p value Result
1 Mardia Skewness  6.34370329562054 0.174908609951883    YES
2 Mardia Kurtosis -1.04835542465742 0.294474882619959    YES
3             MVN              <NA>              <NA>    YES

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk    PC1       0.9647    0.7991    YES
2 Shapiro-Wilk    PC2       0.9642    0.7904    YES

$Descriptives
     n       Mean  Std.Dev     Median       Min      Max      25th       75th
PC1 14 -3.0488623 2.650193 -2.8341032 -7.427043 1.163232 -4.435268 -1.5952221
PC2 14 -0.7119317 2.102378 -0.4854651 -3.815780 2.951247 -2.425462  0.6021574
            Skew  Kurtosis
PC1 -0.006919509 -1.168906
PC2  0.094726976 -1.303774
```

The output above shows the test performed for the known Female individuals. From a Mardia test for multinormality, we see the sample is considered to be from a multivariate normal distribution and the Q-Q plot seems to agree with only one obvious outlier. The Shapiro-Wilk tests for Normality also show each principal component seems to come from a Normal distribution as well.

```
mvn(pca.comp[22:28,1:2],
    multivariatePlot = 'qq',mvnTest = 'mardia',univariateTest = 'SW')
```




```
$multivariateNormality
             Test         Statistic           p value Result
1 Mardia Skewness  3.47197563806491 0.482152284738597    YES
2 Mardia Kurtosis -0.94335588911062 0.345498897674591    YES
3             MVN               <NA>              <NA>    YES

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk    PC1       0.9183    0.4560    YES
2 Shapiro-Wilk    PC2       0.8679    0.1779    YES

$Descriptives
    n       Mean  Std.Dev    Median       Min      Max      25th     75th
PC1 7  4.5680899 3.240629  5.169858 -1.372524 8.051451  3.272868 6.791054
PC2 7 -0.3801742 3.012107 -1.716277 -3.889962 3.325175 -2.627149 2.437071
          Skew  Kurtosis
PC1 -0.6597062 -1.096646
PC2  0.1425864 -2.055031
```

The output above shows the same test performed for known Male individuals. Due to the small sample size, the Q-Q plot is beneficial in showing the sample seems to be a multivariate Normal sample. The next assumption to check is the shared covariance, which is tested with a Box M test.

```
boxM(pca.comp[8:28,1:2],grouping = pca.comp[8:28,]$Sex)
```

```
    Box's M-test for Homogeneity of Covariance Matrices

data:  pca.comp[8:28, 1:2]
Chi-Sq (approx.) = 3.3644, df = 3, p-value = 0.3388
```

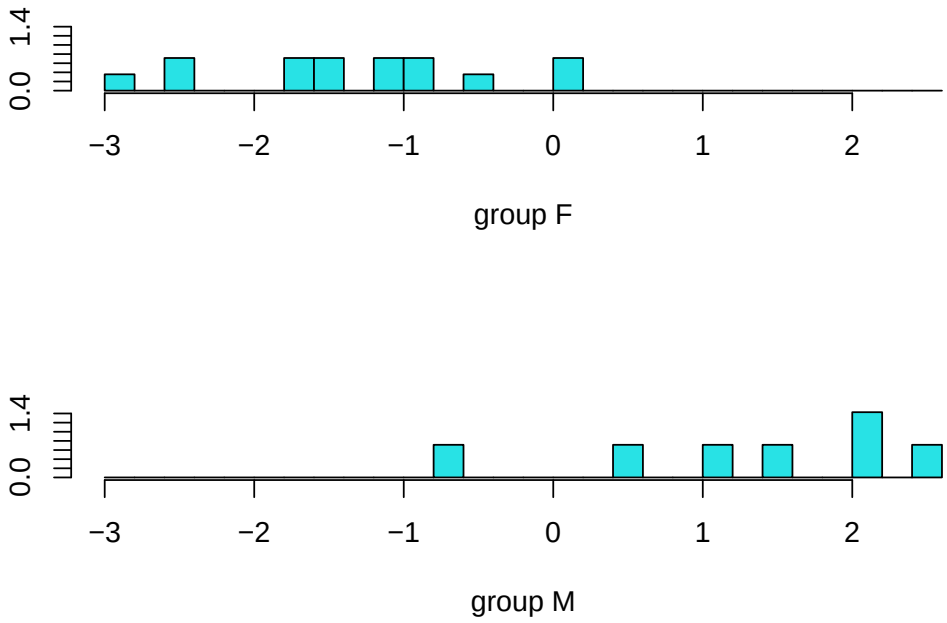
In the Box M test, the null hypothesis is that the two groups have the same covariance matrices. Due to the p-value above, we fail to reject the null at the $\alpha = 0.05$ significance level and consider the covariances as shared. Both this test and the multivariate tests above show that the assumptions for Linear Discriminant Analysis are met. The next output shows a LDA model.

```
lda.fit <- lda(Sex~PC1+PC2,data = pca.comp)
lda.pred <- predict(lda.fit,newdata = pca.comp[c(1,2)])
lda.pred <- data.frame(Individual <- pca.comp$Individual,
                       Prediction <- lda.pred)
lda.fit
```

```
Call:
lda(Sex ~ PC1 + PC2, data = pca.comp)

Prior probabilities of groups:
        F         M
0.6666667 0.3333333

Group means:
        PC1        PC2
F -3.048862 -0.7119317
M  4.568090 -0.3801742

Coefficients of linear discriminants:
            LD1
PC1  0.35213814
PC2 -0.01960364
```

The first LDA uses the sample proportion of Male and Female as the prior. The following plot shows the discriminant calculated for each group.

```
plot(lda.fit)
```



This plot shows the calculated discriminants for the LDA model for observations in each group. From this graph, we can see some overlap between the two groups. This is seen in the confusion matrix as well, as there is some error in predicting the known individuals.

```
table(lda.pred[8:28,]$class,pca.comp[8:28,]$Sex,
      dnn = c('Predicted','Observed'))
```

```
          Observed
```

```
Predicted  F  M
        F 14  1
        M  0  6
```

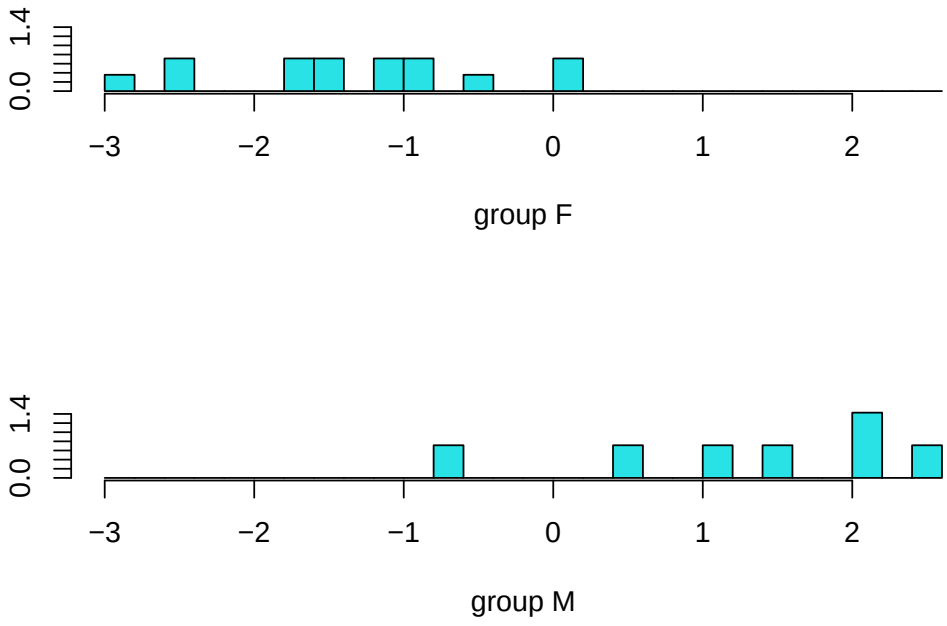
This model incorrectly predicts only 1 male and correctly identifies all females. Another LDA could be fit using priors with equal probability. The following model below shows the effects of this change.

```
lda.fit2 <- lda(Sex~PC1+PC2,data = pca.comp, prior=c(.5,.5))
lda.pred2 <- predict(lda.fit2,newdata = pca.comp[c(1,2)])
lda.pred2 <- data.frame(Individual <- pca.comp$Individual,
                        Prediction <- lda.pred2)
lda.fit2
```

```
Call:
lda(Sex ~ PC1 + PC2, data = pca.comp, prior = c(0.5, 0.5))

Prior probabilities of groups:
  F   M
0.5 0.5

Group means:
        PC1        PC2
F -3.048862 -0.7119317
M  4.568090 -0.3801742

Coefficients of linear discriminants:
            LD1
PC1  0.35213814
PC2 -0.01960364
```

The following plot shows little difference in the discriminant value between the two models.

```
plot(lda.fit2)
```



The biggest difference is in the confusion matrix. As seen below, the equal prior model performs worse than the first LDA.

```
table(lda.pred2[8:28,]$class,pca.comp[8:28,]$Sex,
      dnn = c('Predicted','Observed'))
```

```
         Observed
Predicted  F  M
        F 12  1
        M  2  6
```

Both LDA models predict the same sex for the unknown individuals, but perform differently for the known individuals. There is no known reason for the difference in number of known male and female individuals, so an equal prior probability is not unreasonable.

QDA

Even though the covariances between the two sexes are considered to be similar, a quadratic discriminant analysis can still be run. The following shows such a model.

```
qda.fit <- qda(Sex~PC1+PC2,data = pca.comp)
qda.pred <- predict(qda.fit,newdata = pca.comp[c(1,2)])
qda.pred <- data.frame(Individual <- pca.comp$Individual,
                       Predicted <- qda.pred)
qda.fit
```

```
Call:
qda(Sex ~ PC1 + PC2, data = pca.comp)

Prior probabilities of groups:
        F         M
0.6666667 0.3333333

Group means:
        PC1        PC2
F -3.048862 -0.7119317
M  4.568090 -0.3801742
```

Similar to the LDA models, the first model uses the sample proportion as the prior probabilities of male and female. The confusion matrix is shown below.

```
table(qda.pred[8:28,]$class,pca.comp[8:28,]$Sex,
      dnn = c('Predicted','Observed'))
```

```
         Observed
Predicted  F  M
        F 14  1
        M  0  6
```

This model performs the same as the first LDA, only one male is misclassified as female. Another QDA using equal prior probabilities is shown below.

```
qda.fit2 <- qda(Sex~PC1+PC2,data = pca.comp, prior = c(0.5,0.5))
qda.pred2 <- predict(qda.fit2,newdata = pca.comp[c(1,2)])
qda.pred2 <- data.frame(Individual <- pca.comp$Individual,
                       Predicted <- qda.pred2)
qda.fit2
```

```
Call:
qda(Sex ~ PC1 + PC2, data = pca.comp, prior = c(0.5, 0.5))
```

```
Prior probabilities of groups:
  F   M
0.5 0.5

Group means:
        PC1        PC2
F -3.048862 -0.7119317
M  4.568090 -0.3801742
```

Similar to before, the confusion matrix is shown below.

```
table(qda.pred2[8:28,]$class,pca.comp[8:28,]$Sex,
      dnn = c('Predicted','Observed'))
```

```
         Observed
Predicted  F  M
        F 12  1
        M  2  6
```

Just as before, changing the prior probabilities affects the performance of predicting the known individuals. However, the prediction of the unknown individuals remains the same.

Models with PCA Results

The following table shows the predictions of the multiple logistic regression, LDAs, and QDAs.

```
  Individual Logistic.Prediction LDA.Prediction QDA.Prediction
1      MB125                  F              F              F
2      MB130                  F              F              F
3      MB151                  F              F              F
4       MB47                  M              M              M
5       MB49                  M              M              M
6       MB65                  M              M              M
7       MB81                  F              F              F
```

All models are in agreement that out of the unknown, 4 are Female (MB 81, MB 125, MB 130, MB 151) and 3 are Male (MB 47, MB 49, MB 65). Changing the prior probabilities for the LDA and QDA models did not affect the final prediction of the individuals, although the predicted probabilities did shift slightly. Further analysis may look into the arguments for or against the different priors of both models.

Further Analysis

One negative aspect of using PCA is the lack of interpretability in using models. To account for that, backwards step-wise model selection is used to create a logistic regression. The variables used in that model are then used in LDA and QDA models as well. In order to avoid PCA, only variables with complete data are available for selection.

Backward Model Selection

The backwards step-wise model selection uses Bayesian Information Criterion (BIC) to compare models. The end model will have the lowest BIC out of the models proposed.

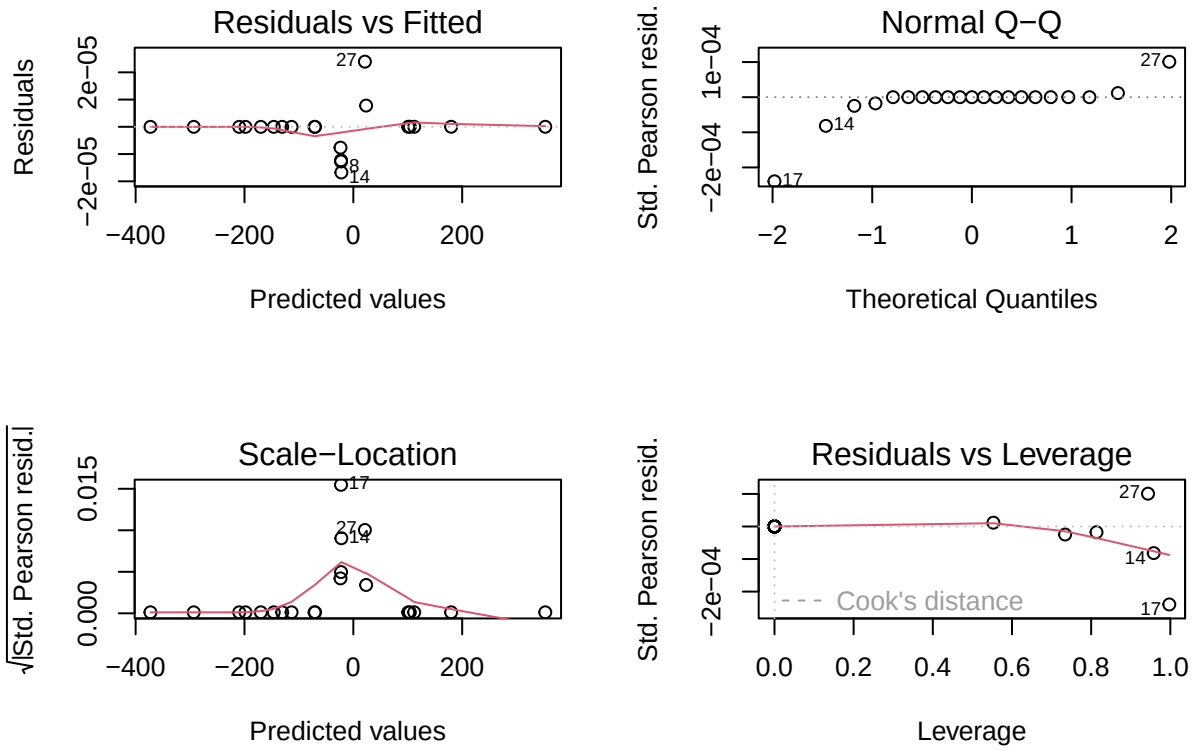
```
t.comp <- complete.cases(t(indiv))
comp.indiv <- cbind(indiv[,1],indiv[,t.comp])
# step wise logistic regression
full.log <- glm(Sex~.,data = comp.indiv[,-2], family = 'binomial')
# k = log(nrow()) - BIC
step.log <- step(full.log,scope = formula(Sex~0),
```

```
                direction = 'backward',k = log(nrow(comp.indiv)),trace = 0)
summary(step.log)
```

```
Call:
glm(formula = Sex ~ Horizontal.T2 + Left.TP.Length.T11 + Left.TP.Length.T5 +
    Left.TP.Width.T10, family = "binomial", data = comp.indiv[,
    -2])

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.369e-05  -2.100e-08  -2.100e-08   2.100e-08   3.382e-05

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2314.60 1805463.71  -0.001    0.999
Horizontal.T2           23.19   19236.44   0.001    0.999
Left.TP.Length.T11      34.57   28739.33   0.001    0.999
Left.TP.Length.T5       31.66   32408.51   0.001    0.999
Left.TP.Width.T10       33.01   26860.40   0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.6734e+01  on 20  degrees of freedom
Residual deviance: 2.5692e-09  on 16  degrees of freedom
  (7 observations deleted due to missingness)
AIC: 10

Number of Fisher Scoring iterations: 25
```

From the output above, we can see the model uses T2 Horizontal, T11 Left TP Length, T5 Left TP Length, and T10 Left TP Width measurements as predictors. We can get a residual plot as before.

```
par(mfrow=c(2,2))
plot(step.log)
```



Same as the logistic regression model from before, point 27 is a high leverage point and considered to be a larger residual. The top-right plot also shows the residuals deviate from Normality more than before. We also see two more high leverage points in the bottom-right plot, points 14 and 17.

```
step.pred <- rep('F',28)
step.pred[predict(step.log, newdata = indiv,type = 'response')>0.5] <- 'M'
table(step.pred[8:28],indiv$Sex[8:28],
      dnn = c('Predicted','Observed'))
```

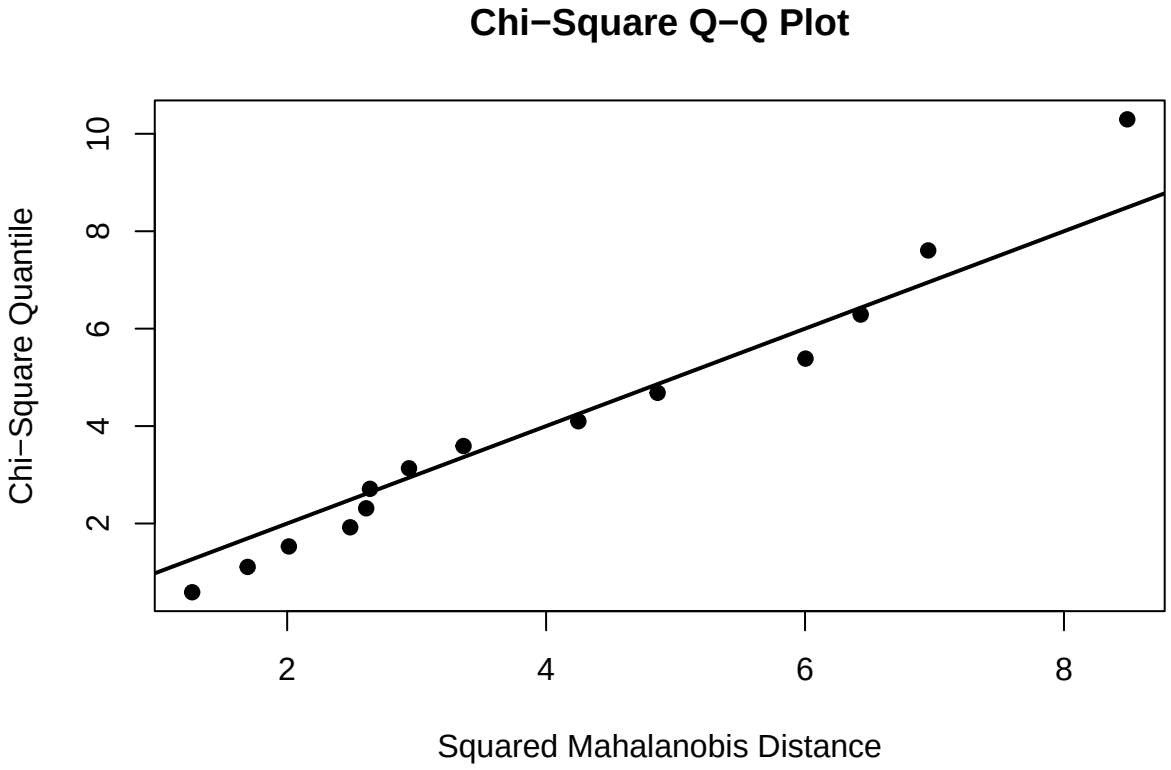
```
         Observed
Predicted  F  M
        F 14  0
        M  0  7
```

Despite the new suspect points from the residual and leverage point plots, the model still is able to retroactively predict the sex for each individual.

LDA

We can check the assumptions for LDA and QDA again. LDA assumes a shared covariance matrix for predictors regardless of sex, while QDA assumes each sex has a different covariance matrix for predictors. Both assume the sample is taken from a multivariate normal distribution.

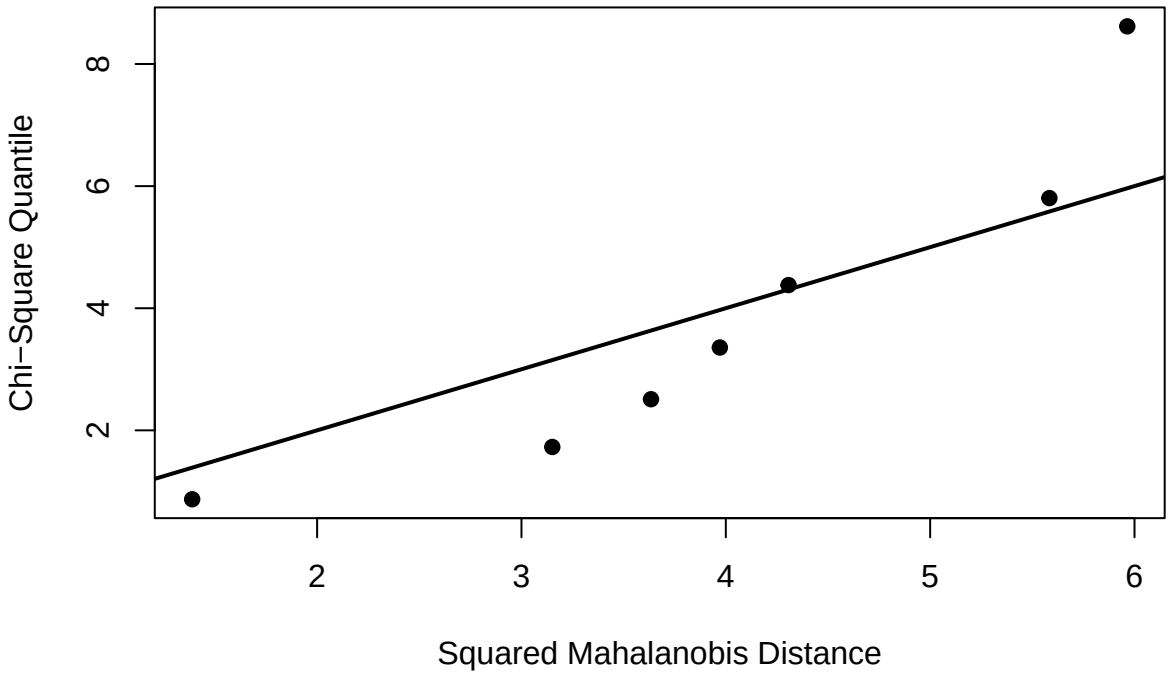
```
stepVars <- labels(terms(step.log))
mvn(comp.indiv[8:21,stepVars], mvnTest = 'mardia',
    univariateTest = 'SW', multivariatePlot = 'qq')
```




```
$multivariateNormality
             Test        Statistic           p value Result
1 Mardia Skewness 21.3522157728383 0.376668377017385    YES
2 Mardia Kurtosis -0.9273635684699 0.353737799406758    YES
3             MVN             <NA>              <NA>    YES

$univariateNormality
          Test           Variable Statistic   p value Normality
1 Shapiro-Wilk   Horizontal.T2       0.9704    0.8820    YES
2 Shapiro-Wilk Left.TP.Length.T11    0.8978    0.1047    YES
3 Shapiro-Wilk Left.TP.Length.T5     0.9504    0.5664    YES
4 Shapiro-Wilk Left.TP.Width.T10     0.9502    0.5633    YES

$Descriptives
                     n     Mean  Std.Dev Median   Min   Max    25th    75th
Horizontal.T2       14 35.24571 2.309957 34.955 31.31 39.74 33.7875 36.4950
Left.TP.Length.T11  14 11.31286 2.518383 11.185  8.01 14.96  8.9700 13.5825
Left.TP.Length.T5   14 18.22286 1.810127 18.025 15.74 22.29 17.2400 19.1100
Left.TP.Width.T10   14 11.99214 2.160332 11.600  8.38 15.35 10.2900 13.3200
                         Skew   Kurtosis
Horizontal.T2       0.30513945 -0.8495922
Left.TP.Length.T11  0.01071956 -1.7649075
Left.TP.Length.T5   0.59695170 -0.4362628
Left.TP.Width.T10   0.17624250 -1.2818829
```

From the Mardia test above, we can consider the female sample to come from a multivariate normal distribution. The Q-Q plot also shows that the sample seems to follow multivariate normality.

```
mvn(comp.indiv[22:28,stepVars], mvnTest = 'mardia',
    univariateTest = 'SW', multivariatePlot = 'qq')
```




```
$multivariateNormality
             Test        Statistic           p value Result
1 Mardia Skewness 19.7196863526697 0.475582941479236    YES
2 Mardia Kurtosis -1.1419064617455 0.253492902425781    YES
3             MVN             <NA>              <NA>    YES

$univariateNormality
          Test           Variable Statistic   p value Normality
1 Shapiro-Wilk   Horizontal.T2       0.9081    0.3828    YES
2 Shapiro-Wilk Left.TP.Length.T11    0.9830    0.9727    YES
3 Shapiro-Wilk Left.TP.Length.T5     0.9378    0.6190    YES
4 Shapiro-Wilk Left.TP.Width.T10     0.8223    0.0675    YES

$Descriptives
                    n     Mean   Std.Dev Median   Min   Max   25th   75th
Horizontal.T2       7 38.57571 3.2577490  38.76 32.72 42.28 37.500 40.635
Left.TP.Length.T11  7 13.47429 3.1301750  13.65  8.59 17.67 11.590 15.615
Left.TP.Length.T5   7 19.87143 0.5780262  19.73 19.00 20.59 19.535 20.355
Left.TP.Width.T10   7 13.71429 0.8125035  13.71 12.86 14.63 12.930 14.470
                          Skew  Kurtosis
Horizontal.T2       -0.43427705 -1.050021
Left.TP.Length.T11  -0.18486586 -1.556774
Left.TP.Length.T5   -0.11648936 -1.734090
Left.TP.Width.T10    0.02333501 -2.110286
```

The male sample is also considered to be from a multivariate normal distribution. Despite a smaller sample size, we can look at the Q-Q plot and see small deviations from a multivariate normal distribution.

The next test is a Box M test for homogeneity of covariances. The following code shows such a test being performed.

```
boxM(comp.indiv[8:28,stepVars],grouping = comp.indiv[8:28,]$Sex)
```

```
    Box's M-test for Homogeneity of Covariance Matrices

data:  comp.indiv[8:28, stepVars]
Chi-Sq (approx.) = 15.611, df = 10, p-value = 0.1113
```

From the test above, we fail to reject the null hypothesis that the covariance matrices are the same at the $\alpha = 0.05$ level. This leads to this data satisfying the assumptions for LDA and breaking the assumption of different covariances for sexes required for QDA

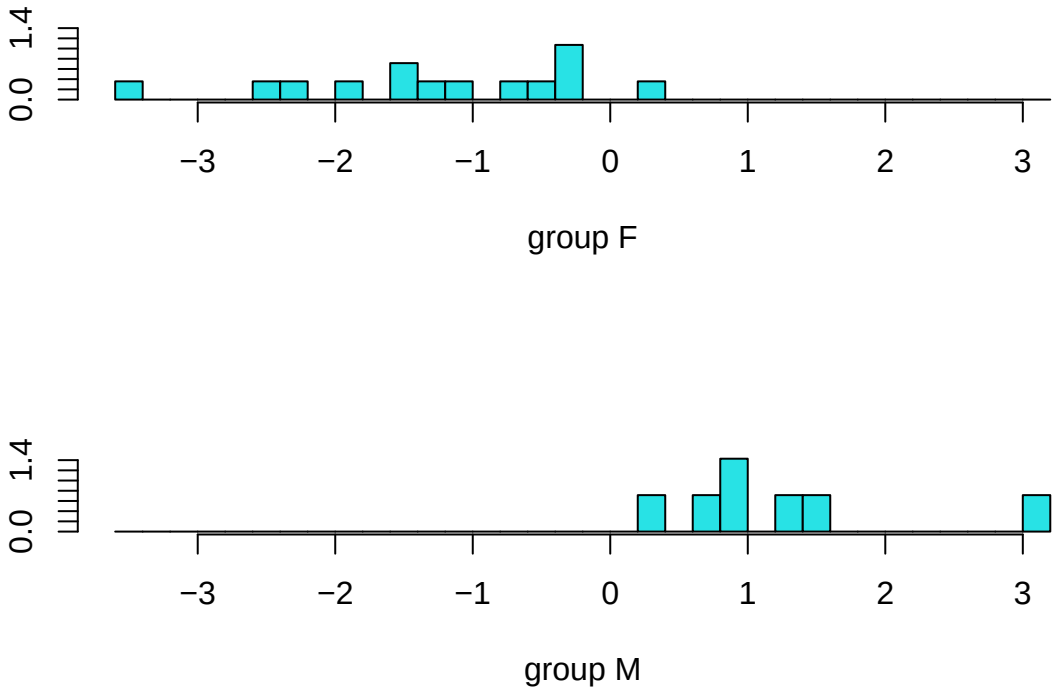
The following LDA model uses sample proportions as priors.

```
lda.step <- lda(formula(step.log),data = comp.indiv)
lda.step
```

```
Call:
lda(formula(step.log), data = comp.indiv)

Prior probabilities of groups:
        F         M
0.6666667 0.3333333

Group means:
  Horizontal.T2 Left.TP.Length.T11 Left.TP.Length.T5 Left.TP.Width.T10
F      35.24571           11.31286          18.22286          11.99214
M      38.57571           13.47429          19.87143          13.71429

Coefficients of linear discriminants:
                       LD1
Horizontal.T2      0.2331343
Left.TP.Length.T11 0.2323548
Left.TP.Length.T5  0.3638031
Left.TP.Width.T10  0.3604889
```

We can also plot the discriminant values for each observation in each group.

```
plot(lda.step)
```



The plot above shows a small overlap in discriminant values for each group. We can look at the confusion matrix below to see how well the model predicts known observations.

```
lda.step.pred <- predict(lda.step,newdata = comp.indiv)
table(lda.step.pred$class[8:28],comp.indiv[8:28,]$Sex,
      dnn = c('Predicted','Observed'))
```

```
         Observed
Predicted  F  M
        F 13  0
        M  1  7
```

This LDA model misclassifies one female as male, a slight reversal from before where a male would be classified as female.

Results

Using the previous two models, the following predictions for the unknown individuals are made.

```
unknown.pred2 <- data.frame(Individual = pca.comp[1:7,]$Individual,
                            Step.Log.Pred = step.pred[1:7],
                            Step.LDA.Pred = lda.step.pred$class[1:7])
unknown.pred2
```

```
  Individual Step.Log.Pred Step.LDA.Pred
1      MB125             F             F
2      MB130             F             F
3      MB151             M             F
4       MB47             M             M
5       MB49             M             M
6       MB65             M             M
7       MB81             F             F
```

From the table above, the LDA model using the variables selected in the backwards step wise logistic regression has the same predictions as the models using principal components analysis. This model is getting

the same results only using 4 measurements rather than all 72 used for the PCA models.

Limitations

The study looks only at adult skeletons, with no adjustment for age of the individual. Data for specific vertebrae were also missing, leading to some measurements not being included as possible selections for backwards step-wise variable selection. The test accuracy of the models used may be affected due to the small sample sizes taken. A look into LOOCV may provide more insight into the accuracy of predictions.

Conclusion

The models used in this project predict that of the 7 skeletons for which sex was unknown, 4 are female (MB 81, MB 125, MB 130, MB 151) and 3 are male (MB 47, MB 49, MB 65). In order to use all measurements taken, Principal Component Analysis was performed with data imputation in order to replace missing values. The first 2 principal components were used as predictors and after checking the performance of a multiple logistic along with the model assumptions for Linear Discriminant Analysis and Quadratic Discriminant analysis, the LDA model using the first two principal components was determined to be most useful for predicting sex of unknown skeletons. After the first models were created, step-wise variable selection was used to create a multiple logistic regression model using BIC as the comparison metric. This model resulted in using T2 Horizontal, T11 Left TP Length, T5 Left TP Length, and T10 Left TP Width measurements as predictors. These measurements were then used to create a LDA model, and after checking assumptions, that had similar results to the models using PCA and data imputation. For future predictions, if data is scarce, the model using T2 Horizontal, T11 Left TP Length, T5 Left TP Length, and T10 Left TP Width measurements would be preferable in order to limit the need for data imputation. If data is not scarce, the LDA model using PCA components as predictors wouold be preferable.

Acknowledgments

This project would not be possible without the help of Leah Fuentes. As part of her Masters' project, this data was made available for analysis and the initial problem of predicting sex of the skeletons was posed. She has also been very helpful in answering questions about the data and problem originally posed.

The guidance of Dr. Guardiola was also instrumental in performing this analysis. The advice given made the analysis in this project easier and much more clear. His instruction throughout the semester has also been informative and provided valuable tools used in this project.

References

[1], [3] - Professional, C. C. medical. (n.d.). Thoracic spine: What it is, Function & Anatomy. Cleveland Clinic. https://my.clevelandclinic.org/health/body/22460-thoracic-spine

[2] - Waxenbaum JA, Reddy V, Futterman B. Anatomy, Back, Thoracic Vertebrae. [Updated 2023 Aug 1]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK459153/

Figure 1 - Provided by Leah Fuentes

Appendix

R Code

```
# Libraries
library(readxl)
library(tidyverse)
library(MVN)
library(GGally)
library(biotools)
```

```
library(softImpute)
# Load Data
Known <- read_excel("...",sheet = "Known", col_names = TRUE)
Known$Sex <- as.factor(Known$Sex)
Known$Vertebrae <- as.factor(Known$Vertebrae)
Known$Individual <- as.factor(Known$Individual)
Unknown <- read_excel("...",sheet = "Unknown", col_names = TRUE)
Unknown$Vertebrae <- as.factor(Unknown$Vertebrae)
Unknown$Individual <- as.factor(Unknown$Individual)
indiv <- read_excel("...",col_names = TRUE)
indiv$Sex <- as.factor(indiv$Sex)
indiv$Individual <- as.factor(indiv$Individual)
colnames(indiv) <- make.names(colnames(indiv),unique = TRUE)
NAs <- complete.cases(Known)
na.known <- Known[NAs,]
# MANOVA with Interaction
man.fit <- manova(cbind(Horizontal,Vertical,lTPlength,
                        lTPwidth,rTPlength,rTPwidth)~Sex*Vertebrae,data=Known)
summary(man.fit, test = 'Wilks')
# MANOVA without Interaction
man.fit2 <- manova(cbind(Horizontal,Vertical,lTPlength,
                        lTPwidth,rTPlength,rTPwidth)~Sex+Vertebrae,data=Known)
summary(man.fit2, test = 'Wilks')
# Data Imputation
set.seed(1335)
imp <- softImpute(as.matrix(indiv[-c(1,2)]))
indiv.comp <- as.data.frame(complete(as.matrix(indiv[-c(1,2)]),imp))
indiv.comp$Sex <- indiv$Sex
indiv.comp$Individual <- indiv$Individual
# Principal Component Analysis
pc.out <- prcomp(indiv.comp[-c(73,74)],scale. = TRUE)
pca.comp <- data.frame(pc.out$x)
pca.comp$Sex <- indiv$Sex
pca.comp$Individual <- indiv$Individual
## Proportion of Variance Explained
pve <- pc.out$sdev^2 / sum(pc.out$sdev^2)
# Multivariate Normality Tests
## Female
mvn(pca.comp[8:21,1:2],
    multivariatePlot = 'qq',mvnTest = 'mardia', univariateTest = 'SW')
## Male
mvn(pca.comp[22:28,1:2],
    multivariatePlot = 'qq',mvnTest = 'mardia',univariateTest = 'SW')
# Box M-test for Homogeneity of Covariance Matrices
boxM(pca.comp[8:28,1:2],grouping = pca.comp[8:28,]$Sex)
# Multiple Logistic Regression
log.fit.pca <- glm(Sex~PC1+PC2, data = pca.comp,family = 'binomial')
log.pred <- rep('F',28)
log.pred[predict(log.fit.pca,newdata = pca.comp[c(1,2)], type = 'response')>0.5] <- 'M'
log.pred <- data.frame(Individual <- pca.comp$Individual,
                       Prediction <- log.pred)
summary(log.fit.pca)
table(log.pred[8:28,]$Prediction....log.pred,pca.comp[8:28,]$Sex,
```

```
      dnn = c('Predicted','Observed'))
# LDA
lda.fit <- lda(Sex~PC1+PC2,data = pca.comp)
lda.pred <- predict(lda.fit,newdata = pca.comp[c(1,2)])
lda.pred <- data.frame(Individual <- pca.comp$Individual,
                       Prediction <- lda.pred)
lda.fit
plot(lda.fit)
table(lda.pred[8:28,]$class,pca.comp[8:28,]$Sex,
      dnn = c('Predicted','Observed'))
# QDA
qda.fit <- qda(Sex~PC1+PC2,data = pca.comp)
qda.pred <- predict(qda.fit,newdata = pca.comp[c(1,2)])
qda.pred <- data.frame(Individual <- pca.comp$Individual,
                       Predicted <- qda.pred)
qda.fit
table(qda.pred[8:28,]$class,pca.comp[8:28,]$Sex,
      dnn = c('Predicted','Observed'))
# Prediction Table
unknown.pred <- data.frame(Individual = pca.comp[1:7,]$Individual,
                           Logistic.Prediction = log.pred[1:7,]$Prediction....log.pred,
                           LDA.Prediction = lda.pred[1:7,]$class,
                           QDA.Prediction = qda.pred[1:7,]$class)
unknown.pred
```