Using PCA for Gender Recognition Based on Audio

Phillip Efthimion

Most people would think that it is easy to guess a person's gender based on the sound of one's voice. There are a lot of factors that make up one's voice. We would like to see if we can use specific pieces of the one's voice, such as the frequencies, to determine gender.

The data consists of 3,168 recordings of voices, both male and female. Each voice was analyzed through R using the warbleR, seewave, and tuneR packages which measured 21 properties about each voice (Becker). They were analyzed only at frequencies between 0 and 280 hertz. This is because this is the human vocal range (Becker). It is unknown if subjects read from a set passage or if there was any other type of control on what was said while recordings were taken. This dataset of voices is a superset of 4 datasets: "The Harvard-Haskins Database of Regularly-Timed Speech", "Telecommunications & Signal Processing Laboratory (TSP) Speech Database at McGill University", "VoxForge Speech Corpus", and "Festvox CMU_ARTIC Speech Database at Carnegie Mellon University.

The objective of this analysis is to first reduce the dimension of the data through principal component analysis, but to still explain at least 80% of our data. Then, we will be using linear regression analysis to see if we can identify one's gender by hearing their voice. Principal component analysis is explaining the maximum amount of variance with as few principal components possible. It will be helpful to avoid multicollinearity by reducing the number of variables.

The response variable, gender, identifies one as either male or female for the purposes of this study. Males have been coded as 1 and females have been coded as 0. This was recorded when the subject had their voice recorded, it is not a educated guess based on other variables.

The first predictor variable is the mean frequency of each voice (meanfreq). It is measured in kilohertz (kHz). As one speaks, one does not talk at the same pitch. This variables measures the average frequency of each subject's voice.

The second predictor is the standard deviation of frequency (sd). It measures the standard deviation of each subject's voice. Standard deviation is a measure of amount of variation in a set of data.

The third predictor, median, refers to the median frequency of each voice. The median is the "middle" frequency of each voice. That is, if one was to lay out all of the frequencies in numerical order that was spoken, the median would be the one right in the middle. It is measured in kilohertz.

The fourth, fifth, and sixth predictors are the first quartile (Q25), the third quartile (Q75), and the interquartile range (IQR). The first quartile refers to the midpoint between the minimum and median numbers of the data when the data would be ranked into 4 equal groups. The third quartile refers to the midpoint between the median and the maximum. The IQR is Q75 - Q25, and is the middle 50% of the data. They are all measured in kilohertz. These are all basic statistical terms.

The seventh predictor term is skew. Skewness is a measure of the asymmetry of the probability distribution. Skewness affects a term's normality. According to R's seewave package, skewness is calculated by S = sum((x-mean(x))^3) / (N-1) / sd^3.

The eighth predictor term is kurtosis (kurt). Kurtosis is a statistical measurement related to skewness. Kurtosis measures the tails of distributions and is a measure of peakness. This could also affect normal distributions. According to R's seewave package, kurtosis is calculated by K = sum((x-mean(x))^4) / (N-1) / sd^4.

The ninth predictor term is spectral entropy (sp.ent). It is a measure for the complexity of the noise using the spectrum and amplitude. (Sueur and Lelouch)

The tenth predictor term refers to spectral flatness (sfm). Spectral flatness, also known as Wiener entropy, is a measurement that quantifies how noise-like as opposed to tone-like a sound is. It is measured from 0 to 1 with 0 being a pure tone and one being white noise. (Sueur and Simonis)

The eleventh predictor term is the mode frequency. The mode of the dataset refers to the value that occurs more times than any other value. This is a basic statistical measurement.

The twelfth predictor is centroid. Centroid is computed by C = sum(x*y) with y being the dependent variable gender. This is another measure for finding the "center" of the data. (Sueur and Simonis)

The thirteenth predictor refers to the peak frequency (peakf). The peak frequency is the maximum frequency that the subject vocalizes. It is measured in kilohertz (kHz).

The fourteenth, fifteenth, and sixteenth predictors are the average (meanfun), minimum (minfun), and maximum (maxfun) fundamental frequency measured across acoustic signal. Fundamental frequency is the lowest frequency of a periodic waveform.

The seventeenth, eighteenth, and nineteenth predictors refers to the average (meandom), minimum (mindom), and maximum (maxdom) dominant frequency measured across acoustic signal. The dominant frequency is the frequency of the wave with the highest amplitudes (Becker). People speak with multiple inflections and each inflection has different frequency peaks, the dominant sound wave, that is the one with the highest peaks, is the one we use to measure the dominant frequency. These are measured in kilohertz.

The twentieth predictor refers to the range of the dominant frequency measured across acoustic signal (dfrange). It is the range, therefore it is calculated by dfrange = maxdom - mindom. (Sueur and Simonis)

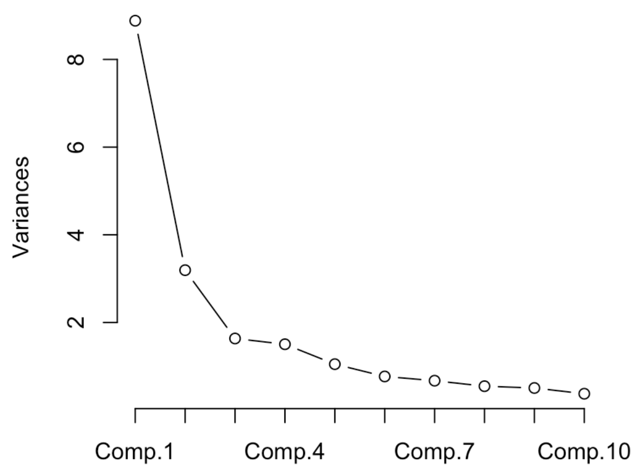
The twenty-first predictor refers to the modulation index (modindx). It is calculated as the sum of the absolute differences between adjacent measurements of fundamental frequencies divided by the frequency range (Sueur and Simonis).

Since we will be eventually performing multiple linear regression, we will create histograms for all of the predictor variables in order make each variable normally distributed. We will use log, square root, inverse, and any other transformations in order to make each as normal as possible. Based upon the histogram of each of the predictor variables, we are going to take the log transformation of skewness, the minimum dominant frequency measurement, and the modulation index. We are going to take the square root transformation of spectral entropy, average dominant frequency measurement, and the range of dominant frequency. We will perform an inverse transformation on the interquartile range and kurtosis. Additionally, we will be performing the inverse of the log transformation on the maximum fundamental frequency. Also, because some of the values for the modulation index, which are being log transformed, equal 0, we will being adding 1 to each of the values before the transformation. This needs to be done because log(0) does not equal a finite number. Principal component analysis is unable to be performed unless all of the values are finite.

Now we will compute the principal components.

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     2.9803140 1.7869164 1.2787119 1.22636748 1.02453386 0.87708982 0.82013179 0.73964731 0.71127693
Proportion of Variance 0.4441136 0.1596535 0.0817552 0.07519886 0.05248348 0.03846433 0.03363081 0.02735391 0.02529574
Cumulative Proportion  0.4441136 0.6037671 0.6855223 0.76072115 0.81320463 0.85166896 0.88529977 0.91265367 0.93794942
                           Comp.10    Comp.11     Comp.12     Comp.13    Comp.14     Comp.15    Comp.16     Comp.17
Standard deviation     0.61275163 0.55839710 0.392126548 0.369051935 0.32863202 0.270470269 0.20083478 0.185583660
Proportion of Variance 0.01877323 0.01559037 0.007688161 0.006809967 0.00539995 0.003657708 0.00201673 0.001722065
Cumulative Proportion  0.95672264 0.97231301 0.980001172 0.986811138 0.99221109 0.995868797 0.99788553 0.999607592
                            Comp.18      Comp.19      Comp.20
Standard deviation     0.0874052395 1.443883e-02 1.668215e-08
Proportion of Variance 0.0003819838 1.042398e-05 1.391471e-17
Cumulative Proportion  0.9999895760 1.000000e+00 1.000000e+00
```

The principal components analysis give us 20 components. We want to narrow the amount of dimensions of our analysis so we will use a scree plot in order to help us decide how many dimensions are required.



The scree plot lists the variances of each component. We are to find the “elbow” of the plot to help us determine where to stop adding components to our analysis. The elbow will be where there is the point of diminishing returns in regards to variance. Here, it looks like component 3 is the elbow of the graph. Therefore, we should take components 1, 2, and 3.

To reduce noise, we will cutoff the value of each variable in each component to anything above 0.3. This gives us for component 1: the mean frequency and centroid. Component 2 is the interquartile range, skew, kurtosis, and spectral entropy. Component 3 is the maximum fundamental frequency, the minimum dominant frequency, maximum dominant frequency, the range of the dominant frequency, and the modulation index.

```
Loadings:
              Comp.1 Comp.2 Comp.3
meanfreq       0.311        -0.194
sd            -0.277 -0.147
median         0.279        -0.230
Q25            0.299  0.114 -0.111
Q75            0.186 -0.213 -0.287
inv_IQR        0.206  0.340  0.103
log_skew              0.403
inv_kurt             -0.418
sqrt_sp.ent   -0.227 -0.308
mode           0.243 -0.143 -0.147
sfm           -0.274 -0.153
centroid       0.311        -0.194
meanfun        0.198  0.188  0.227
minfun         0.160
inv_log_maxfun -0.114 0.132 -0.380
sqrt_meandom   0.235 -0.225  0.171
log_mindom     0.132  0.118 -0.305
sqrt_maxdom    0.229 -0.237  0.320
sqrt_dfrange   0.227 -0.242  0.326
log_modindx                 -0.357
gender        -0.157 -0.284 -0.256
```

```
> summary.lm(pclm, correlation = T)

Call:
lm(formula = voice2$gender ~ comp1 + comp2 + comp3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.38081 -0.24245  0.04039  0.32054  1.10059

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.500000   0.006976   71.68   <2e-16 ***
comp1       -0.071027   0.002341  -30.35   <2e-16 ***
comp2       -0.119324   0.003904  -30.57   <2e-16 ***
comp3        0.059414   0.005455   10.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3926 on 3164 degrees of freedom
Multiple R-squared:  0.3842,	Adjusted R-squared:  0.3836
F-statistic:   658 on 3 and 3164 DF,  p-value: < 2.2e-16

Correlation of Coefficients:
      (Intercept) comp1 comp2
comp1 0.00
comp2 0.00        0.00
comp3 0.00        0.00  0.00
```

The first component deals with the “center” of the graph. The mean is the average for the data. The centroid also deals with the center of the frequencies. We will refer to this as the Center component.

The second component deals with normality. Skewness and kurtosis measure normality and tail lengths. Interquartile range deals with the middle 50% of data. Spectral entropy deals with noise which effects the normality of a distribution. Therefore, the second component is the Normal component.

The third component deals with range. The maximum fundamental frequency and the minimum and maximum dominant frequency deal with absolute high and low points. The modulation index is calculated with absolute differences and frequency ranges. Therefore, the third component is the Range component.

As shown above on the right, we now calculate a multilinear regression with the Center, Normality, and Range components against the dependent variable gender. The output shows that all of the components are statistically significant. Therefore, we reject the null hypothesis with our p value of 2.2e-16 so we can identify gender based on these components. Our equation will be Gender = 0.5 - 0.7 Center - 0.11 Normality + 0.6 Range. Also, we can see that there is no correlation of coefficients which means that the PCA was performed correctly.

Now we have to go back and cross validate. With cross validation, we can see that one of our assumptions was not satisfied. The PCA only attributes for ~68% of our variance and we would like it to account for 80%. Therefore we will add components 4 and 5 to our model.

```
Loadings:
               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
meanfreq        0.311        -0.194 -0.157
sd             -0.277 -0.147        -0.171 -0.160
median          0.279        -0.230 -0.212 -0.103
Q25             0.299  0.114 -0.111
Q75             0.186 -0.213 -0.287 -0.394 -0.164
inv_IQR         0.206  0.340  0.103  0.199
log_skew               0.403        -0.335  0.209
inv_kurt              -0.418         0.197 -0.183
sqrt_sp.ent    -0.227 -0.308         0.116 -0.159
mode            0.243 -0.143 -0.147
sfm            -0.274 -0.153         0.104 -0.179
centroid        0.311        -0.194 -0.157
meanfun         0.198  0.188  0.227  0.130 -0.491
minfun          0.160                0.133 -0.142
inv_log_maxfun -0.114  0.132 -0.380  0.289  0.296
sqrt_meandom    0.235 -0.225  0.171  0.209  0.161
log_mindom      0.132  0.118 -0.305  0.316  0.282
sqrt_maxdom     0.229 -0.237  0.320  0.102  0.267
sqrt_dfrange    0.227 -0.242  0.326         0.260
log_modindx                  -0.357  0.433 -0.230
gender         -0.157 -0.284 -0.256 -0.177  0.386
```

```
Call:
lm(formula = voice2$gender ~ comp1 + comp2 + comp3 + comp4 +
    comp5)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9179 -0.2350  0.0287  0.2648  1.3886

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.500000   0.005902   84.72   <2e-16 ***
comp1       -0.071027   0.001980  -35.87   <2e-16 ***
comp2       -0.119324   0.003303  -36.13   <2e-16 ***
comp3        0.059414   0.004616   12.87   <2e-16 ***
comp4       -0.073574   0.004813  -15.29   <2e-16 ***
comp5       -0.184339   0.005761  -32.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3322 on 3162 degrees of freedom
Multiple R-squared:  0.5594,	Adjusted R-squared:  0.5587
F-statistic:   803 on 5 and 3162 DF,  p-value: < 2.2e-16

Correlation of Coefficients:
      (Intercept) comp1 comp2 comp3 comp4
comp1 0.00
comp2 0.00        0.00
comp3 0.00        0.00  0.00
comp4 0.00        0.00  0.00  0.00
comp5 0.00        0.00  0.00  0.00  0.00
```

We can see that this model has a much higher adjusted R^2 than our last with 3 components, 0.38 to 0.56. 0.56 is a much more respectable adjusted R^2 score. Looking back at our scree plot, the fifth component is not as significant an elbow as the third component, but it is still valid. Component 4 consists of the third quartile (Q75), skew, minimum dominant frequency, and modulation index. I would say that this component is the tail component because it deals more with the area where outliers may occur. The fifth component is average fundamental frequency.

We still reject the null hypothesis and all of our components are still statistically significant. Our equation has now been expanded to Gender = 0.5 - 0.7 Center - 0.11 Normality + 0.6 Range - 0.07 Tail - 0.18 Meanfun. The final model tells us that males have a larger vocal range, since the range component is positive and we have male coded as 1. However, Center, Normality, Tail, and Meanful are all negatively correlated.

A look at the residual plots of each component against the residuals do not show any trends. This helps validate our model and its assumptions.

Now we will test this model against the model without performing principal component regression.

```
> rawlm <- lm(voice2$gender ~ ., data = voice_x)
> summary.lm(rawlm, correlation = T)

Call:
lm(formula = voice2$gender ~ ., data = voice_x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.92217 -0.11347  0.01642  0.13097  1.16741 

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -0.135107   0.576924  -0.234  0.81486    
meanfreq       -0.239963   1.636918  -0.147  0.88346    
sd              4.367008   1.049543   4.161 3.26e-05 ***
median         -1.128977   0.462909  -2.439  0.01479 *  
Q25            -3.774023   0.440251  -8.572  < 2e-16 ***
Q75             6.667022   0.753214   8.851  < 2e-16 ***
inv_IQR         0.007750   0.001081   7.168 9.46e-13 ***
log_skew       -0.362020   0.041512  -8.721  < 2e-16 ***
inv_kurt       -0.529932   0.114657  -4.622 3.96e-06 ***
sqrt_sp.ent     1.522512   0.586888   2.594  0.00952 ** 
mode            0.486293   0.081284   5.983 2.44e-09 ***
sfm            -0.649537   0.079706  -8.149 5.23e-16 ***
centroid              NA         NA      NA       NA    
meanfun       -12.681818   0.184728 -68.651  < 2e-16 ***
minfun          2.978265   0.243951  12.208  < 2e-16 ***
inv_log_maxfun -0.144085   0.035622  -4.045 5.36e-05 ***
sqrt_meandom   -0.136057   0.032510  -4.185 2.93e-05 ***
log_mindom      0.010916   0.012898   0.846  0.39743    
sqrt_maxdom    -0.184538   0.216608  -0.852  0.39431    
sqrt_dfrange    0.213524   0.212507   1.005  0.31508    
log_modindx     0.308266   0.146948   2.098  0.03600 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2204 on 3148 degrees of freedom
Multiple R-squared:  0.807,	Adjusted R-squared:  0.8058 
F-statistic: 692.6 on 19 and 3148 DF,  p-value: < 2.2e-16
```

The full model has a large change from our previous model that we ran PCA through. First, it agrees with our conclusion to reject the null hypothesis. It has a much larger adjusted R squared value of 0.8058. However, not all of the model's variables are statistically significant. Mean frequency, minimum and maximum dominant frequency, and dominant frequency range are all not deemed statistically significant at alpha = 0.05. This model is suggesting that one's dominant frequency is not statistically signifiant so it is not the dominant tones of one's voice, but the small inflections that are a give away to one's gender. This makes some sense. For example, choirs, there are those in either genders able to sing with altos, tenors, or baritones. All different groups that sing in different octaves.

Though our model we performed PCA on accounts for 80% of the variance, the full model above has all of the variance. This alone does not account for such a large difference in adjusted R squared. More analysis must be done. For example, why centroid is fine in all of the analysis until performing multiple linear regression on the full model.

I believe that our conclusion makes sense. When I answer the phone, one seems to usually be able to inherently know what gender is talking on the other end of the line. Our analysis backs up this claim. Further analysis that I would like to do would be training the model using machine learning techniques and the “caret” package in R. This will allow us to train our model we performed principal component analysis on and test how accurate the model is.

Bibliography

Becker, Kory. “GenderRecognition by Voice | Kaggle.” *Gender Recognition by Voice | Kaggle*. N.p., n.d. Web.

Becker, Kory. "Identifying the Gender of a Voice using Machine Learning." Primary Objects. N.p., 14 Jan. 2017. Web.

Sueur, Jerome, and Caroline Simonis. “R: Spectral Properties.” *R: Spectral Properties*. N.p., n.d. Web.

Sueur, Jerome, and Laurent Lelouch. “R: Shannon and Renyi Spectral Entropy.” *R: Shannon and Renyi Spectral Entropy*. N.p., n.d. Web.

“Wiener Entropy.” *Wiener Entropy - Sound Analysis Pro*. N.p., n.d. Web.

R Code

```
# PCA on Voice Recognition
# Phillip Efthimion

# Input Data

voice <- read.csv("/Users/Phillip/Downloads/voice.csv", sep=",", header = TRUE)
names(voice)
head(voice$label)
str(voice)



#################
# Transformation#
#################

# meanfreq. No transformation improves normality

hist(voice$meanfreq)
voice$log_meanfreq <- log10(voice$meanfreq)
voice$sqrt_meanfreq <- (voice$meanfreq)^0.5
voice$inv_meanfreq <- 1 / (voice$meanfreq)
hist(voice$log_meanfreq)
hist(voice$sqrt_meanfreq)
hist(voice$inv_meanfreq)

# sd. No transformation improves normality

hist(voice$sd)
voice$log_sd <- log10(voice$sd)
voice$sqrt_sd <- (voice$sd)^0.5
voice$inv_sd <- 1 / (voice$sd)
hist(voice$log_sd)
hist(voice$sqrt_sd)
hist(voice$inv_sd)

# median. No transformation improves normality

hist(voice$median)
voice$log_median <- log10(voice$median)
voice$sqrt_median <- (voice$median)^0.5
voice$inv_median <- 1 / (voice$median)
hist(voice$log_median)
hist(voice$sqrt_median)
hist(voice$inv_median)
```

```
# Q25. No transformation improves normality

hist(voice$Q25)
voice$log_Q25 <- log10(voice$Q25)
voice$sqrt_Q25 <- (voice$Q25)^0.5
voice$inv_Q25 <- 1 / (voice$Q25)
hist(voice$log_Q25)
hist(voice$sqrt_Q25)
hist(voice$inv_Q25)

# Q75. No transformation improves normality. None look normal.

hist(voice$Q75)
voice$log_Q75 <- log10(voice$Q75)
voice$sqrt_Q75 <- (voice$Q75)^0.5
voice$inv_Q75 <- 1 / (voice$Q75)
hist(voice$log_Q75)
hist(voice$sqrt_Q75)
hist(voice$inv_Q75)

# IQR. Inverse transformation helps improve normality

hist(voice$IQR)
voice$log_IQR <- log10(voice$IQR)
voice$sqrt_IQR <- (voice$IQR)^0.5
voice$inv_IQR <- 1 / (voice$IQR)
hist(voice$log_IQR)
hist(voice$sqrt_IQR)
hist(voice$inv_IQR)

# skew. Log transformation helps improve normality.

hist(voice$skew)
voice$log_skew <- log10(voice$skew)
voice$sqrt_skew <- (voice$skew)^0.5
voice$inv_skew <- 1 / (voice$skew)
hist(voice$log_skew)
hist(voice$sqrt_skew)
hist(voice$inv_skew)

# kurt. Inverse transformation helps improve normality.

hist(voice$kurt)
voice$log_kurt <- log10(voice$kurt)
voice$sqrt_kurt <- (voice$kurt)^0.5
voice$inv_kurt <- 1 / (voice$kurt)
hist(voice$log_kurt)
hist(voice$sqrt_kurt)
hist(voice$inv_kurt)
```

```
# sp.ent. SQRT transformation helps improve normality

hist(voice$sp.ent)
voice$log_sp.ent <- log10(voice$sp.ent)
voice$sqrt_sp.ent <- (voice$sp.ent)^0.5
voice$inv_sp.ent <- 1 / (voice$sp.ent)
hist(voice$log_sp.ent)
hist(voice$sqrt_sp.ent)
hist(voice$inv_sp.ent)

# Mode. No transformation improves normality.

hist(voice$mode)
voice$log_mode <- log10(voice$mode)
voice$sqrt_mode <- (voice$mode)^0.5
voice$inv_mode <- 1 / (voice$mode)
hist(voice$log_mode)
hist(voice$sqrt_mode)
hist(voice$inv_mode)

# Centroid. No transformation improves normality.

hist(voice$centroid)
voice$log_centroid <- log10(voice$centroid)
voice$sqrt_centroid <- (voice$centroid)^0.5
voice$inv_centroid <- 1 / (voice$centroid)
hist(voice$log_centroid)
hist(voice$sqrt_centroid)
hist(voice$inv_centroid)

# Meanful. o transformation improves normality.

hist(voice$meanfun)
voice$log_meanfun <- log10(voice$meanfun)
voice$sqrt_meanfun <- (voice$meanfun)^0.5
voice$inv_meanfun <- 1 / (voice$meanfun)
hist(voice$log_meanfun)
hist(voice$sqrt_meanfun)
hist(voice$inv_meanfun)

# minion. No transformation improves normality

hist(voice$minfun)
voice$log_minfun <- log10(voice$minfun)
voice$sqrt_minfun <- (voice$minfun)^0.5
voice$inv_minfun <- 1 / (voice$minfun)
hist(voice$log_minfun)
hist(voice$sqrt_minfun)
```

```
hist(voice$inv_minfun)

# max fun. invlog best transformation.

hist(voice$maxfun)
voice$log_maxfun <- log10(voice$maxfun)
voice$sqrt_maxfun <- (voice$maxfun)^0.5
voice$inv_maxfun <- 1 / (voice$maxfun)
hist(voice$log_maxfun)
hist(voice$sqrt_maxfun)
hist(voice$inv_maxfun)
voice$inv_log_maxfun <- 1 / (log10(voice$maxfun))
hist(voice$inv_log_maxfun)

# meandom. SQRT transformation best

hist(voice$meandom)
voice$log_meandom <- log10(voice$meandom)
voice$sqrt_meandom <- (voice$meandom)^0.5
voice$inv_meandom <- 1 / (voice$meandom)
hist(voice$log_meandom)
hist(voice$sqrt_meandom)
hist(voice$inv_meandom)

# mindom. log transform.

hist(voice$mindom)
voice$log_mindom <- log10(voice$mindom)
voice$sqrt_mindom <- (voice$mindom)^0.5
voice$inv_mindom <- 1 / (voice$mindom)
hist(voice$log_mindom)
hist(voice$sqrt_mindom)
hist(voice$inv_mindom)

# maxdom. SQRT transformation.

hist(voice$maxdom)
voice$log_maxdom <- log10(voice$maxdom)
voice$sqrt_maxdom <- (voice$maxdom)^0.5
voice$inv_maxdom <- 1 / (voice$maxdom)
hist(voice$log_maxdom)
hist(voice$sqrt_maxdom)
hist(voice$inv_maxdom)

# dfrange. SQRT transformation

hist(voice$dfrange)
voice$log_dfrange <- log10(voice$dfrange)
voice$sqrt_dfrange <- (voice$dfrange)^0.5
```

```
voice$inv_dfrange <- 1 / (voice$dfrange)
hist(voice$log_dfrange)
hist(voice$sqrt_dfrange)
hist(voice$inv_dfrange)

# modindx. LOG transformation.

hist(voice$modindx)

voice$modindx <- voice$modindx + 1

voice$log_modindx <- log10(voice$modindx)
voice$sqrt_modindx <- (voice$modindx)^0.5
voice$inv_modindx <- 1 / (voice$modindx)
hist(voice$log_modindx)
hist(voice$sqrt_modindx)
hist(voice$inv_modindx)

#################
#################

# Convert labels male & female to 0 & 1

voice$gender[1:1584] <- 1
voice$gender[1585:3168] <- 0


# add 1 to $modindx so we can take the log transformation
voice$modindx <- voice$modindx + 1

voice2 <- voice[, c("meanfreq", "sd", "median", "Q25", "Q75", "inv_IQR", "log_skew", "inv_kurt",
"sqrt_sp.ent", "mode", "sfm", "centroid", "meanfun", "minfun", "inv_log_maxfun",
"sqrt_meandom", "log_mindom", "sqrt_maxdom", "sqrt_dfrange", "log_modindx", "gender")]

voice_x <- voice_x[, -21]
names(voice_x)

pca_x <- princomp(voice_x, cor = T)
plot(pca_x, type = "l")
summary(pca, loadings = T)

comp1 <- pca_x$scores[,1]
comp2 <- pca_x$scores[,2]
comp3 <- pca_x$scores[,3]

pclm <- lm(voice2$gender ~ comp1 + comp2 + comp3)
summary.lm(pclm, correlation = T)

rawlm <- lm(voice2$gender ~ ., data = voice_x)
```

```
summary.lm(rawlm, correlation = T)

pclm.res <- resid(pclm)
plot(voice2$gender, pclm.res)
plot(voice2$meanfreq, pclm.res)
plot(voice2$sd, pclm.res)
plot(comp1, pclm.res)
plot(comp2, pclm.res)
plot(comp3, pclm.res)
plot(comp4, pclm.res)
plot(comp5, pclm.res)

# Unused

# pca <- princomp(voice, cor = T)
# summary(pca, loadings = T)

# plot(pca, type = "l")
```