**Utilization of Voice Data to Diagnose Parkinson’s Disease Early**

Author: Azeezat Adekanmbi

ISQS-6350 Multivariate Analysis

Dr. Alireza Sheikh-Zadeh

December 2020

# Introduction*.*

Parkinson’s disease is a degenerative neurological disorder with slowly emerging symptoms that include tremors, stiffness, and difficulty with walking, balance, and coordination. The disease primarily affects older people, with an estimated 96% of the 10 million people diagnosed with Parkinson’s being over the age of 50[[1]](#endnote-1). While the disease itself is not fatal, the neurological decline causing difficultly with motor skills can lead to things like falling, making complications from Parkinson’s the 14th leading cause of death in the United States[[2]](#endnote-2).

There is no singular medical test that diagnoses Parkinson’s disease, skilled practitioners must use various combinations of observed symptoms and diagnostic tests. Treatment for Parkinson’s disease aims to improve symptoms; there is no cure. Medications that treat the symptoms become less effective as the disease progresses,[[3]](#endnote-3) so it is important to identify the disease as early as possible. Being difficult to diagnose, combined with the importance of diagnosing it early enough for treatment to be effective, results in a need to search for innovative new ways to identify the disease as early as possible.

One potential warning sign that may be an early symptom of Parkinson’s disease is a change in voice becoming hoarse, which is referred to medically as dysphonia[[4]](#endnote-4). There are several different ways to measure voice characteristics, which could potentially be used to identify individuals experiencing a symptom of Parkinson’s earlier than other traditional methods. I found a dataset on the UCI Machine Learning Repository website that provides different voice measurements for 31 people, 23 of which had Parkinson’s disease[[5]](#footnote-1). Each record in the dataset represents one of the biomedical voice recordings for these individuals, resulting in 195 total observations. This dataset was created through a collaboration between the University of Oxford and the National Centre for Voice and Speech in Denver, Colorado, for a study looking at general voice disorders[[6]](#endnote-5). A description of the variables included in this dataset are detailed in Table 1.

Table : Parkinson's Data Set Variable Descriptions

| **Variable** | **Description** |
| --- | --- |
| name | Subject name and recording number |
| MDVP:Fo(Hz) | Average vocal fundamental frequency |
| MDVP:Fhi(Hz) | Maximum vocal fundamental frequency |
| MDVP:Flo(Hz) | Minimum vocal fundamental frequency |
| MDVP:Jitter(%) | MDVP[[7]](#footnote-2) jitter as a percentage |
| MDVP:Jitter(Abs) | MDVP absolute jitter in microseconds |
| MDVP:RAP | MDVP Relative Amplitude Peturbation |
| MDVP:PPQ | MDVP five-point Period Petrubation Quotient |
| Jitter:DDP | Average absolute difference of differences between cycles, divided by the average period |
| MDVP:Shimmer | MDVP local shimmer |
| MDVP:Shimmer(dB) | MDVP local shimmer in decibels |
| Shimmer:APQ3 | Three point Amplitude Petrubation Quotient |
| Shimmer:APQ5 | Five point Amplitude Peturbation Quotient |
| MDVP:APQ | MDVP 11-point Amplitude Petrubation Quotient |
| Shimmer:DDA | Average absolute difference between consecutive differences between the amplitudes of consecutive periods |
| NHR | Noise-to-Harmonics Ratio |
| HNR | Harmonics-to-Noise Ratio |
| status | Health status of the subject (1) – Parkinson’s, (0) – healthy |
| RPDE | Recurrence Period Density Entropy |
| D2 | Correlation dimension |
| DFA | Detrended Fluctuation Analysis |
| spread1 | First nonlinear measure of fundamental frequency variation |
| spread2 | Second nonlinear measure of fundamental frequency variation |
| PPE | Pitch Period Entropy |

My motivation behind choosing this topic and this data set is my interest in exploring how data can be used in medicine to improve health outcomes. Using data to identify a new early diagnostic tool for Parkinson’s is potentially a great example of this. This report will focus on two questions:

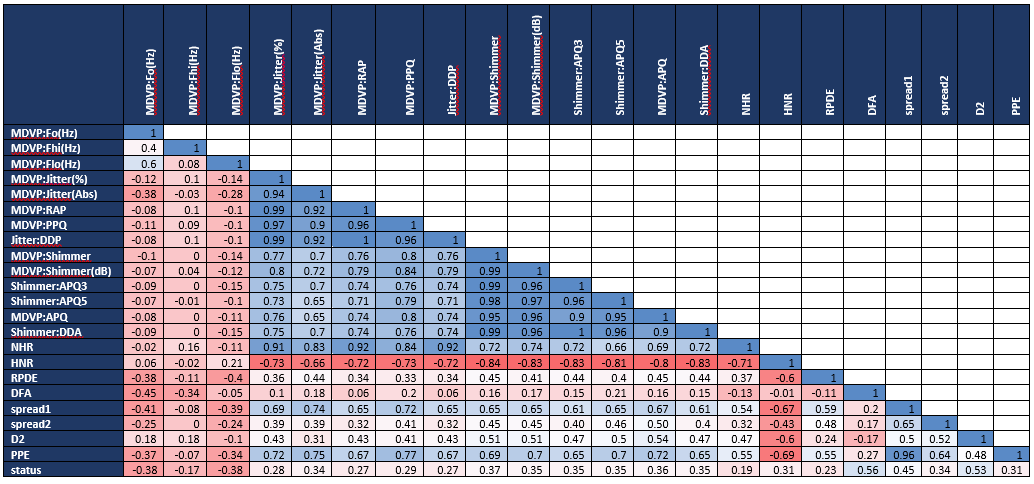
1. Which voice measurement variables are most associated with having Parkinson’s disease?
2. Are the expected groups in the data, of those that do and do not have Parkinson’s disease, distinct and observable?

I performed an initial review of the data to assess whether the majority of the correlations in the data fell into the appropriate range for this project, between 0.15 and 0.9 (absolute value). There are 23 numeric variables in the dataset, resulting in 253 distinct variable pairs. 182 of those pairs had a correlation in the appropriate range, which is certainly a majority.

A full correlation matrix for the variables is shown in Table 2. A gradient formatting is applied to highlight the greatest positive and negative correlations. The highest correlation of 1.00 can be seen between Jitter:DDP and MDVP:Fo(Hz) as well as Shimmer:DDA and MDVP:Fo(Hz). The greatest negative correlation of -0.84 exists between HNR and MDVP:Fo(HZ). Given that all of these are similar measurements of voice, strong correlations like those noted are expected.

The status variable, which identifies the patients diagnosed with Parkinson’s disease, has the highest positive correlation with DFA and the greatest negative correlation with MDVP:Fo and MDFP:Flo.

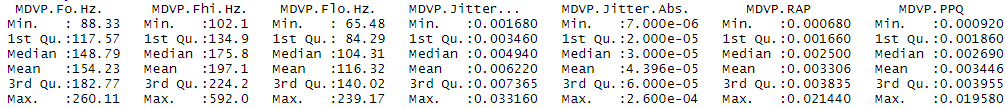
Table : Correlation Matrix for Parkinson's Variables



# Data Cleaning and Summary Statistics

The available dataset was cleaned in preparation for data analysis. The data cleaning process involves fixing missing value issues and removing irrelevant/unimportant variables from the dataset. The original dataset has no missing values at all, hence, missing value computations were not necessary. The variable name ‘name’ consists of pseudo-names of patients in the Parkinson’s dataset and was removed from the dataset since it has no meaningful contribution to the cluster analysis.

Summary statistics of Parkinson’s data variables were computed, and the results for 7 of the variables are displayed below:

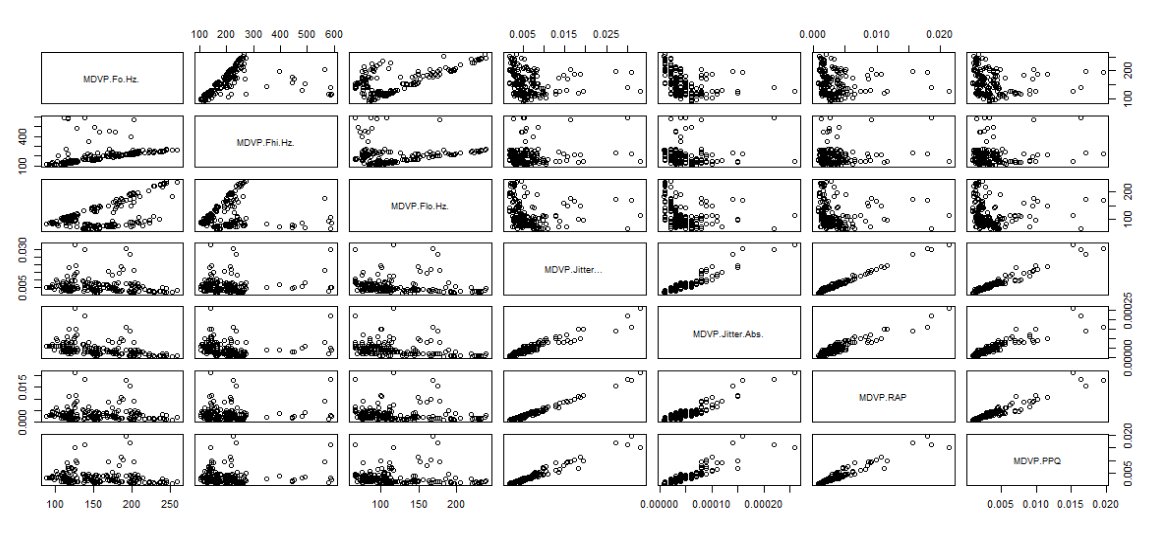


*Check appendix for full results*

# Data Visualization

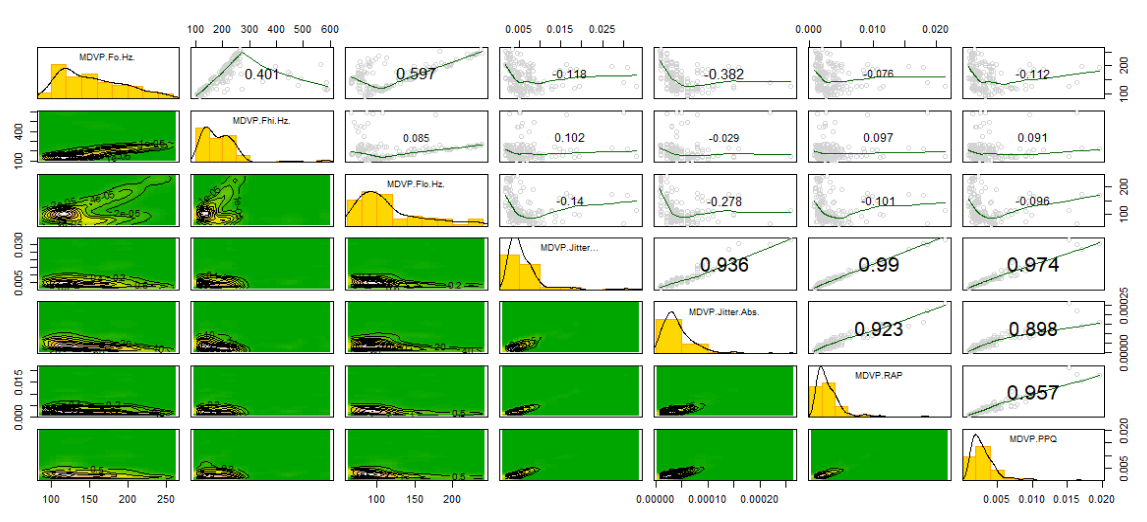
After performing some data cleaning activities on the dataset, I continued to explore or visualize what’s going on in the dataset; I had to use scatterplots which is one of the technics available out there. This will help us to understand how a few pairs of the variables are related and their linear association; in this case I took the first 7 pairs of variables to explore the relationships and correlations. This can be seen in figure (1) below. From the results in the graph, I can see the existence of some outliers in the dataset. This might in the course of the data exploration and analysis have a negative impact on the results.

Figure



Another technique is also applied to understand the relationships, variations and or correlation between the first few pairs of the variables. This method or technic adds estimated bivariate densities of the data to the scatterplot. The results can be seen in figure (2) below. From the results I can see that for some pairs of the variables, there is a positive correlation between them and for other pairs, there is a negative correlation while others also do not seem to be correlated or are less correlated.

Figure



# Dimension Reduction Analysis

There are 22 variables in the dataset providing different voice measurements. I applied principal components analysis to reduce this dimensionality while accounting for as much of the original variation as possible. The units differ between the voice measurements, some record frequency in hertz (Hz), some measure sound intensity in decibels (dB), while others are ratios or proportions. Because of the inconsistency in measurement and scale between the variables, I performed the principal components analysis using the correlation matrix to such that the data is appropriately scaled.

The table below shows the variance accounted for by each of the principal components. There are many different methodologies for selecting a cutoff for the number of principal components to use, which is a subjective process. Given that the first three components account for 77% of the cumulative variance, exceeding the simple rule of thumb of trying to account for 75% of total variance, I limited the analysis to using those three. Principal component 1 explains 58.9% of the total variance, with principal components 2 and 3 having much smaller explanatory power with 11.3% and 7.0%, respectively.

Table : PCA Variance

| **Component** | **Variance** | **Cu. Variance** | **Component** | **Variance** | **Cu. Variance** |
| --- | --- | --- | --- | --- | --- |
| 1 | 0.589 | 0.589 | 12 | 0.005 | 0.991 |
| 2 | 0.113 | 0.702 | 13 | 0.003 | 0.994 |
| 3 | 0.070 | 0.772 | 14 | 0.002 | 0.996 |
| 4 | 0.067 | 0.839 | 15 | 0.001 | 0.997 |
| 5 | 0.044 | 0.883 | 16 | 0.001 | 0.998 |
| 6 | 0.033 | 0.916 | 17 | 0.001 | 0.999 |
| 7 | 0.025 | 0.941 | 18 | 0.001 | 1.000 |
| 8 | 0.016 | 0.957 | 19 | < 0.001 | 1.000 |
| 9 | 0.013 | 0.970 | 20 | < 0.001 | 1.000 |
| 10 | 0.010 | 0.980 | 21 | < 0.001 | 1.000 |
| 11 | 0.006 | 0.986 | 22 | < 0.001 | 1.000 |

Loadings for the three principal components selected are shown in the table below. The correlations between almost all variables and PC1 are of the same approximate moderate value and in the same direction, with 14 of the 22 variables having a loading between -0.2 and -0.3. The notable exceptions to this include the three variables measuring vocal fundamental frequency (MDVP:Fo, Fhi, Flo) and Detrended Fluctuation Analysis measure (DFA) which did not have meaningful loading coefficients, and the Harmonics-to-Noise Ratio (HNR) having a moderate correlation in the opposite direction from the other variables (0.242). Given the overall similarity of most loadings in this component, I name PC1 “overall voice quality score.” Given the negative values for most loadings, I expected those scoring high in PC1 to have low values for most measurements of voice quality.

The measurements of vocal fundamental frequency (MDVP:Fo, Fhi, Flo) have the greatest correlations with PC2, with the average fundamental frequency (MDVP:Fo) being the highest at 0.533. Given these loadings, I named PC2 “fundamental frequency score.” Those scoring higher in component 2 would have a higher average vocal fundamental frequency than others.

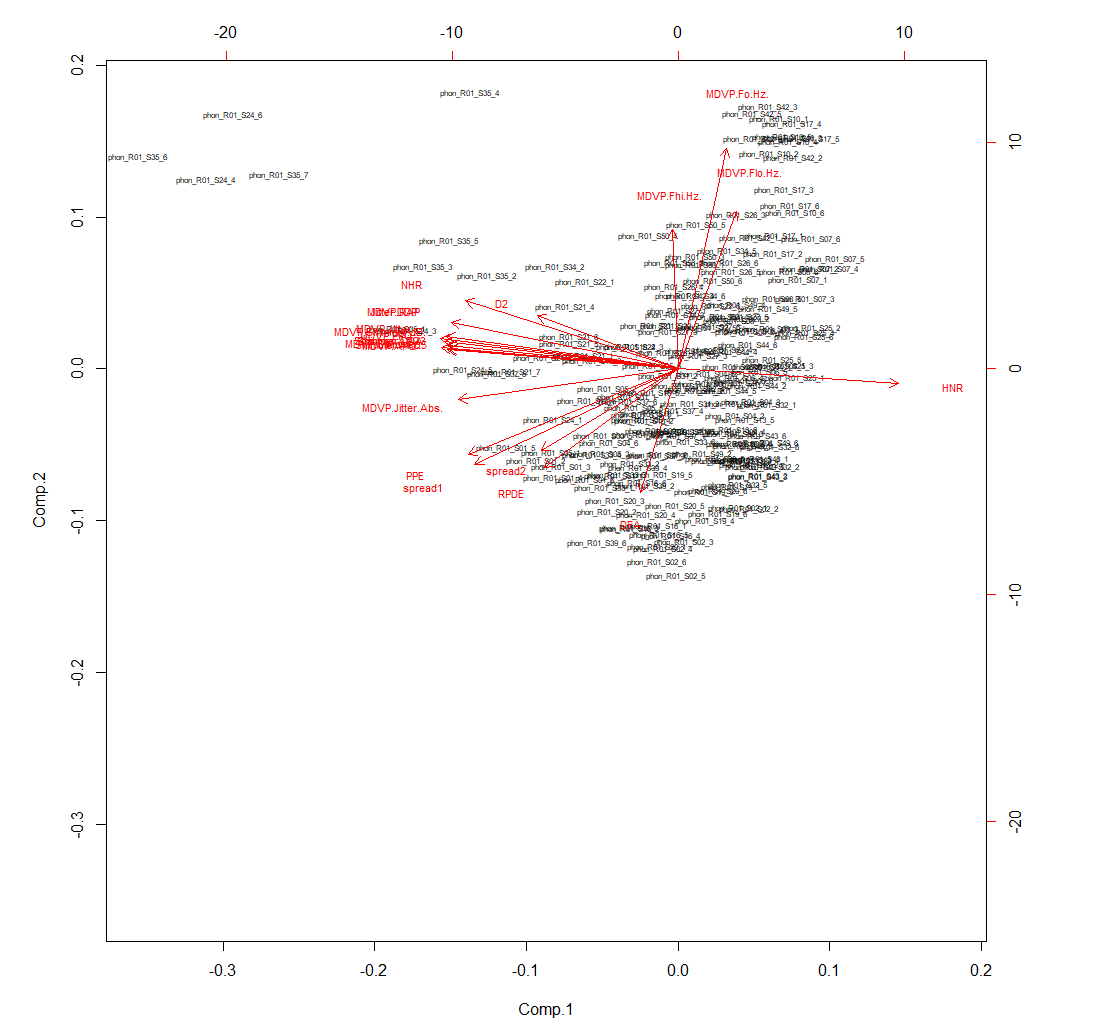
The variables with the greatest correlation to PC3 are signal fractal scaling exponent (DFA), with a value of -0.466, and dynamical complexity measurement (D2) with a value of 0.428. It’s notable that these correlations are opposite directions from each other. Those scoring higher in PC3 would have a greater dynamical complexity measurement and a lower fractal scaling exponent. Because of this differential dynamic between the two variable being important to PC3, I name it the “DFA D2 difference score.”

Table : PCA Loadings

| **Variable** | **PC1** | **PC2** | **PC3** |
| --- | --- | --- | --- |
| MDVP:Fo(Hz) |  | 0.553 | 0.128 |
| MDVP:Fhi(Hz) |  | 0.349 | 0.268 |
| MDVP:Flo(Hz) |  | 0.395 | -0.233 |
| MDVP:Jitter(%) | -0.255 |  | -0.151 |
| MDVP:Jitter(Abs) | -0.242 |  | -0.185 |
| MDVP:RAP | -0.250 | 0.116 | -0.169 |
| MDVP:PPQ | -0.257 |  | -0.179 |
| Jitter:DDP | -0.250 | 0.116 | -0.169 |
| MDVP:Shimmer | -0.260 |  |  |
| MDVP:Shimmer(dB) | -0.262 |  |  |
| Shimmer:APQ3 | -0.253 |  |  |
| Shimmer:APQ5 | -0.252 |  |  |
| MDVP:APQ | -0.254 |  |  |
| Shimmer:DDA | -0.253 |  |  |
| NHR | -0.234 | 0.171 |  |
| HNR | 0.242 |  | -0.162 |
| RPDE | -0.147 | -0.249 | 0.353 |
| DFA |  | -0.311 | -0.466 |
| spread1 | -0.224 | -0.239 | 0.152 |
| spread2 | -0.151 | -0.204 | 0.349 |
| D2 | -0.155 | 0.133 | 0.428 |
| PPE | -0.231 | -0.214 |  |

The biplot blow provides a graphical representation of the first two principal components and the data. The name variable, which identifies the person the measurements were taken from, is used to mark points on the plot. From the graph I can see that the angle between the arrows for the variables measuring vocal fundamental frequency (MDVP:Fo, Fhi, Flo) are quite small, indicating high correlation with each other, which makes sense intuitively as they’re all measuring the same vocal component. The direction of those arrows for the vocal fundamental frequency variables show their measurements are greater for those scoring higher in PC2, as previously described. A large number of the variable arrows with a negative relationship to PC1 are almost overlapping, identifying high correlation with each other and confirming the previous analysis of PC1 – those scoring higher have lower values for most measurements other than HNR.

Figure : PCA Biplot



Given that the dataset includes a status variable to identify patients that have Parkinson’s disease, I can evaluate principal component’s ability to differentiate between the groups by running a correlation between the principal component scores and the status variable. The correlation between the principal component scores and the Parkinson’s disease status indicator is shown in the table below. These show that there is only a moderate correlation between how a patient scores in these components and whether they have Parkinson’s disease or not, suggesting that this data does not do a great job of indicating whether someone is sick and should not be the sole factor for diagnosing a patient with Parkinson’s disease.

Table : PCA Parkinson's Status Correlation

|  |  |
| --- | --- |
| **Principal Component** | **Correlation with Parkinson’s Disease Status** |
| PC1 – overall voice quality score | -0.42 |
| PC2 – fundamental frequency score | -0.39 |
| PC3 – DFA D2 difference score | 0.16 |

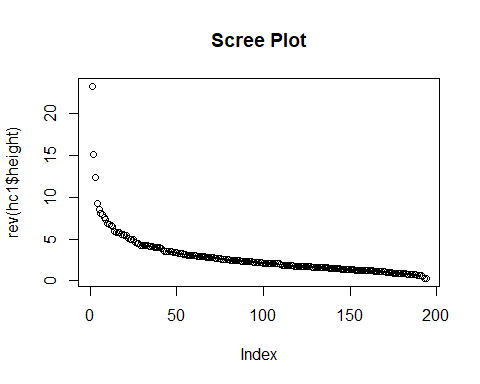
Expanded analysis tables and the R code used to generate these results can be found in the appendix for the dimension reduction analysis section.

# Cluster Analysis

In this instance, I wanted to see I could easily identify the two main clusters in the data – Patients with and without Parkinson’s using Hierarchical and Model-Based clustering. I chose Hierarchical clustering because I wanted to see if the data would show us what I assumed (that there are two main clusters within the data) without supplying this number. Model-Based clustering was selected for comparison, and so that I could outright try to find 2 clusters by ourselves.

The figure below shows the scree plot of the result of hierarchical clustering

Figure : Hierarchical Clustering Scree Plot



The elbow test suggests that there are 4 clusters to be observed within the dataset, which goes against the assumptions of 2 clusters. When viewing the contingency tables of the clustering results using the three Hierarchical clustering methods, I can see that none of the methods do a great job of separating patients with or without Parkinson’s.

Table : Contingency Table - HC Complete Linkage

|  |  |  |
| --- | --- | --- |
| Cluster | Has Parkinson’s | Healthy |
| 1 | 7 | 77 |
| 2 | 41 | 56 |
| 3 | 0 | 10 |
| 4 | 0 | 4 |

Table : Contingency Table - HC Average Linkage

|  |  |  |
| --- | --- | --- |
| Cluster | Has Parkinson’s | Healthy |
| 1 | 48 | 124 |
| 2 | 0 | 19 |
| 3 | 0 | 2 |
| 4 | 0 | 2 |

Table : Contingency Table - HC Single Linkage

|  |  |  |
| --- | --- | --- |
| Cluster | Has Parkinson’s | Healthy |
| 1 | 48 | 143 |
| 2 | 0 | 1 |
| 3 | 0 | 1 |
| 4 | 0 | 2 |

While not being a great separator, Complete Linkage gives us the most variance between patients with and without Parkinson’s. Clusters 1, 3, and 4 are clearly clusters for patients with Parkinson’s, while Cluster two seems to go either way. Based on these results I can say that the data does not provide strong indicators of whether or not a patient has Parkinson’s disease.

Using Model-Based clustering, the results were not much better. Examining the results of Model-Based clustering using two clusters, I have the following results:

Table : Contingency Table - Model-Based Clustering

|  |  |  |
| --- | --- | --- |
| Cluster | Has Parkinson’s | Healthy |
| 1 | 22 | 107 |
| 2 | 26 | 40 |

When I supplied the recommended number of clusters to the model, I had the exact same results as when I did not. The clusters are not clearly separated by whether or not they have Parkinson’s disease or not, although it could be said that falling into Cluster 2 could mean that a patient has a higher chance of being ill.

Finally, I created clusters using only the variables that make up Principal Component 1 from the earlier analysis using Model-Based clustering. The results were as follows

Table : Contingency Table - Model Based Clustering with Variables making up PC1

|  |  |  |
| --- | --- | --- |
| Cluster | Has Parkinson’s | Healthy |
| 1 | 5 | 79 |
| 2 | 43 | 68 |

Clustering based on these variables gives us a bit more separation than clustering on the entire dataset. I can see that patients in Cluster 1 are very likely to be healthy, while patients in Cluster 2 could almost go either way.

The results are not conclusive enough, and the clusters are not stark enough to be a source of final diagnoses on Parkinson’s disease. As a lot of tests to diagnose Parkinson’s involve ruling out other causes vi, this could also be used in a similar way: Based on these results, If a patient falls into Cluster 2, they should be recommended for further testing for Parkinson’s disease as there is clearly a higher chance of illness.

The R Code used to generate the tables and images used for this section can be seen in Appendix B.

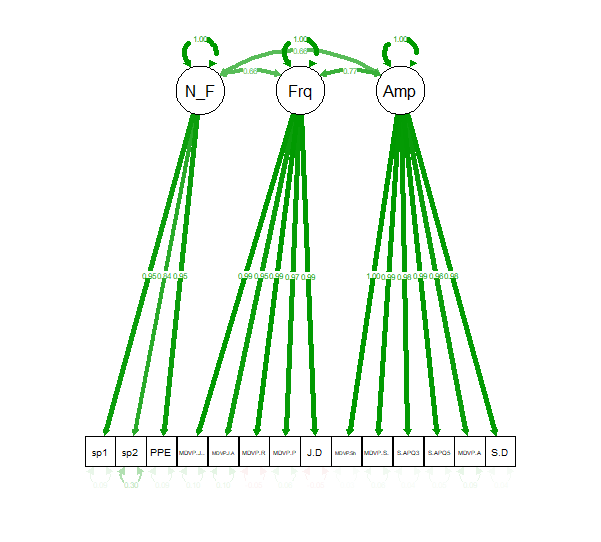
# Confirmatory Factor Analysis

The description of the dataset 1 outlines a few groups of data within the dataset. I will examine the largest 3 (our latent variables) to confirm that the factors within them fit as defined. The variables in question are:

1. **Measures of Variation in Fundamental Frequency**: The UCI Machine Learning Repository describes **MDVP:Jitter(%)**, **MDVP:Jitter(Abs)**, **MDVP:RAP**, **MDVP:PPQ**, and **Jitter:DDP** as Measures of variation in fundamental frequency
2. **Measures of Variation in Amplitude**: The Repository describes **MDVP:Shimmer**, **MDVP:Shimmer(dB)**, **Shimmer:APQ3**, **Shimmer:APQ5**, **MDVP:APQ**, and **Shimmer:DDA** as measures of variation in amplitude
3. **Non-linear Measures of Variation in Fundamental Frequency**: The Repository describes **spread1**, **spread2**, and **PPE** as non-linear measures of variation in fundamental frequency

I built the model using the variables above as guidelines (See Appendix C). the Path Diagram confirms that the Model is as I designed it.

Figure : CFA Path Diagram



Right off the bat, I can see that the correlation coefficients relating the factors to the latent variables are quite strong, and the relationships I defined are meaningful. While this is not conclusive of the model, it suggests I are on the right path.

Examining the model deeper gives us the following results:

|  |
| --- |
| Model Chisquare = 3563.587 Df = 74 Pr(>Chisq) = 0  Goodness-of-fit index = 0.5071418  Adjusted goodness-of-fit index = 0.3006741  SRMR = 0.04372491  Normalized Residuals  Min. 1st Qu. Median Mean 3rd Qu. Max.  -1.46648 -0.04966 0.01014 0.11832 0.23458 1.59425  R-square for Endogenous Variables  MDVP.Jitter... MDVP.Jitter.Abs. MDVP.RAP  0.9806 0.8518 1.0000  MDVP.PPQ Jitter.DDP MDVP.Shimmer  0.9165 1.0000 1.0050  MDVP.Shimmer.dB. MDVP.APQ Shimmer.APQ3  0.9675 0.9169 0.9688  Shimmer.APQ5 Shimmer.DDA spread1  0.9596 0.9688 0.9399  spread2 PPE  0.4262 0.9856  Parameter Estimates  Iterations = 464 |

Using the various fit indices, I can conclude that the model is not an adequate fit to the actual data. Examining the SRMR alone, I would assume that the model actually performed well (Our SRMR of 0.04 is < 0.05, which is good). The other fit indices however tell a different story. the Chi-Squared p-value of 0 is > 0.05, which rejects the null hypothesis that the restricted and unrestricted covariance matrices are equal. The GFI and AGFI are both under 0.95 which is also poor.

As the model was based on the variable groups detailed on the source (UCI Machine Learning Repository), the results would suggest that the defined groups within the dataset are not as strongly related as one would expect.

# Conclusion

I posed two questions at the beginning of this report:

1. Which voice measurement variables are most associated with having Parkinson’s disease?
2. Are the expected groups in the data, of those that do and do not have Parkinson’s disease, distinct and observable?

The first question is easily answered by looking at the correlation matrix shown in Table 2. I see that two variables have a strong correlation with the status variable that identifies patients that have been diagnosed with Parkinson’s disease. These two variables are Detrended Fluctuation Analysis (DFA, correlation 0.56) and the correlation dimension (D2, correlation 0.53).

Given the large number of variables in this dataset that are difficult to understand for those without expertise in vocal measurements, I attempted to reduce the dimensions of the dataset through principal components analysis to create a few meaningful components that would explain the majority of the variance in the data. Performing PCA resulted in three components – “overall voice quality score”, “fundamental frequency score”, and “DFA D2 difference score.” While this analysis accomplished simplifying the data, it’s notable that these three components were not meaningfully correlated with the Parkinson’s disease status variable, having correlations of -0.42, -0.39, and 0.16, respectively. None of these are greater than the correlations seen between status and the individual variables of DFA (0.56) and D2 (0.53), suggesting that the individual variables are better if you’re solely interested in helping to identify Parkinson’s disease in patients.

To answer the second question, I performed cluster analysis on the data. An advantage of the dataset is that the status variable should, in theory, identify the patients that belong in two distinct clusters – healthy and Parkinson’s disease. Hierarchical and model-based clustering on their own failed to meaningfully separate the groups. When performing model-based clustering on a subset of variables identified by PCA to have meaningful loadings, I got better results where one cluster contained almost all healthy patients. The other cluster was evenly split between healthy and Parkinson’s disease patients. So presence in one cluster could potentially identify healthy patients, but the presence in the other is essentially a coin flip on whether or not they were healthy.

The biggest pro of this study was having access to the status variable that allowed us to evaluate the ability for the factors and clusters to identify those with Parkinson’s disease. Cons of this study include the presence of multiple measurements taken from the same patient that may have had an uncontrollable effect on this results and the complexity of the variables being difficult to interpret.

Future work for post-analysis should first focus on simplifying the data by limiting the number of variables such that there are less measurements of the same thing. There are more than 20 variables measuring essentially three different categories of vocal quality – variation in fundamental frequency, variation in amplitude, and non-linear measurements of variation in fundamental frequency. Limiting the variables to a more meaningful and distinct set that are best associated with Parkinson’s disease may allow for more meaningful and easy to understand results. Furthermore, the unknown effect of having multiple measurements from the same person should be studied and controlled for as appropriate.

For now, the primary conclusion I reached is that using multivariate analysis methods on the captured data does not easily or distinctly identify those with Parkinson’s disease. This is not particularly surprising, given the complexity of healthcare data in humans that may have multiple other underlying conditions that could affect vocal measurements and health.

# References

1. Parkinson’s Foundation. “Parkinson’s Foundation Statistics.” <https://www.parkinson.org/Understanding-Parkinsons/Statistics> [↑](#endnote-ref-1)
2. Parkinson’s Foundation. “What is Parkinson’s?” <https://www.parkinson.org/understanding-parkinsons/what-is-parkinsons> [↑](#endnote-ref-2)
3. Sveinbjornsdottir S (October 2016). “The clinical symptoms of Parkinson’s disease. Journal of Neurochemistry. <https://onlinelibrary.wiley.com/doi/full/10.1111/jnc.13691> [↑](#endnote-ref-3)
4. Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering (to appear). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3051371/> [↑](#endnote-ref-4)
5. UCI Machine Learning Repository, Parkinson’s Data Set. <https://archive.ics.uci.edu/ml/datasets/parkinsons> [↑](#footnote-ref-1)
6. Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. “Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection”, BioMedical Engineering OnLine 2007, 6:23 (26 June 2007) <https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-6-23>

   vi Clinic, M. (n.d.). Parkinson's Disease. Retrieved from Mayo Clinic: <https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/diagnosis-treatment/drc-20376062>

   # Appendix A – Data Cleaning and Summary Statistics R Code and Full Output

   #read in the original data  
   dat=read.csv('Park (1).csv')

   # report summary statistics of the variables in the data  
   summary(dat)

   ## name MDVP.Fo.Hz. MDVP.Fhi.Hz. MDVP.Flo.Hz.   
   ## phon\_R01\_S01\_1: 1 Min. : 88.33 Min. :102.1 Min. : 65.48   
   ## phon\_R01\_S01\_2: 1 1st Qu.:117.57 1st Qu.:134.9 1st Qu.: 84.29   
   ## phon\_R01\_S01\_3: 1 Median :148.79 Median :175.8 Median :104.31   
   ## phon\_R01\_S01\_4: 1 Mean :154.23 Mean :197.1 Mean :116.32   
   ## phon\_R01\_S01\_5: 1 3rd Qu.:182.77 3rd Qu.:224.2 3rd Qu.:140.02   
   ## phon\_R01\_S01\_6: 1 Max. :260.11 Max. :592.0 Max. :239.17   
   ## (Other) :189   
   ## MDVP.Jitter... MDVP.Jitter.Abs. MDVP.RAP MDVP.PPQ   
   ## Min. :0.001680 Min. :7.000e-06 Min. :0.000680 Min. :0.000920   
   ## 1st Qu.:0.003460 1st Qu.:2.000e-05 1st Qu.:0.001660 1st Qu.:0.001860   
   ## Median :0.004940 Median :3.000e-05 Median :0.002500 Median :0.002690   
   ## Mean :0.006220 Mean :4.396e-05 Mean :0.003306 Mean :0.003446   
   ## 3rd Qu.:0.007365 3rd Qu.:6.000e-05 3rd Qu.:0.003835 3rd Qu.:0.003955   
   ## Max. :0.033160 Max. :2.600e-04 Max. :0.021440 Max. :0.019580   
   ##   
   ## Jitter.DDP MDVP.Shimmer MDVP.Shimmer.dB. Shimmer.APQ3   
   ## Min. :0.002040 Min. :0.00954 Min. :0.0850 Min. :0.004550   
   ## 1st Qu.:0.004985 1st Qu.:0.01650 1st Qu.:0.1485 1st Qu.:0.008245   
   ## Median :0.007490 Median :0.02297 Median :0.2210 Median :0.012790   
   ## Mean :0.009920 Mean :0.02971 Mean :0.2823 Mean :0.015664   
   ## 3rd Qu.:0.011505 3rd Qu.:0.03789 3rd Qu.:0.3500 3rd Qu.:0.020265   
   ## Max. :0.064330 Max. :0.11908 Max. :1.3020 Max. :0.056470   
   ##   
   ## Shimmer.APQ5 MDVP.APQ Shimmer.DDA NHR   
   ## Min. :0.00570 Min. :0.00719 Min. :0.01364 Min. :0.000650   
   ## 1st Qu.:0.00958 1st Qu.:0.01308 1st Qu.:0.02474 1st Qu.:0.005925   
   ## Median :0.01347 Median :0.01826 Median :0.03836 Median :0.011660   
   ## Mean :0.01788 Mean :0.02408 Mean :0.04699 Mean :0.024847   
   ## 3rd Qu.:0.02238 3rd Qu.:0.02940 3rd Qu.:0.06080 3rd Qu.:0.025640   
   ## Max. :0.07940 Max. :0.13778 Max. :0.16942 Max. :0.314820   
   ##   
   ## HNR status RPDE DFA   
   ## Min. : 8.441 Min. :0.0000 Min. :0.2566 Min. :0.5743   
   ## 1st Qu.:19.198 1st Qu.:1.0000 1st Qu.:0.4213 1st Qu.:0.6748   
   ## Median :22.085 Median :1.0000 Median :0.4960 Median :0.7223   
   ## Mean :21.886 Mean :0.7538 Mean :0.4985 Mean :0.7181   
   ## 3rd Qu.:25.076 3rd Qu.:1.0000 3rd Qu.:0.5876 3rd Qu.:0.7619   
   ## Max. :33.047 Max. :1.0000 Max. :0.6852 Max. :0.8253   
   ##   
   ## spread1 spread2 D2 PPE   
   ## Min. :-7.965 Min. :0.006274 Min. :1.423 Min. :0.04454   
   ## 1st Qu.:-6.450 1st Qu.:0.174350 1st Qu.:2.099 1st Qu.:0.13745   
   ## Median :-5.721 Median :0.218885 Median :2.362 Median :0.19405   
   ## Mean :-5.684 Mean :0.226510 Mean :2.382 Mean :0.20655   
   ## 3rd Qu.:-5.046 3rd Qu.:0.279234 3rd Qu.:2.636 3rd Qu.:0.25298   
   ## Max. :-2.434 Max. :0.450493 Max. :3.671 Max. :0.52737   
   ##

   # remove the first column from the data  
   dat=dat[,-1]

   #check for missing values in the data  
   table(is.na(dat))

   ##   
   ## FALSE   
   ## 4485

   # Appendix B – Dimension Reduction Analysis R Code and Full Output

   R Code

   #read in cleaned parkinsons data and assign name variable to row name  
   parkinsons <- read.csv("C:/ parkinsons.csv", row.names="name")  
     
   #remove status variable prior to PCA  
   parkinsons.c <- parkinsons[,-17]  
     
   #run PCA using the correlation matrix  
   parkinsons\_pca <- princomp(parkinsons.c, cor=T)  
     
   #print summary of PCA results to generate tables of cumulative variance and loading coefficients  
   summary(parkinsons\_pca, loading=T)

   #run biplot on first two principal components  
   biplot(parkinsons\_pca, cex=c(0.5,0.7))

   #assess the correlation between the PCA scores and Parkinson's Disease indicator  
   cor(parkinsons$status, parkinsons\_pca$scores[,1:3])

   R Output

   PCA

   ## Importance of components:  
   ## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
   ## Standard deviation 3.599738 1.5766657 1.24178490 1.21036614 0.98687187  
   ## Proportion of Variance 0.589005 0.1129943 0.07009226 0.06659028 0.04426891  
   ## Cumulative Proportion 0.589005 0.7019993 0.77209160 0.83868189 0.88295080  
   ## Comp.6 Comp.7 Comp.8 Comp.9 Comp.10  
   ## Standard deviation 0.85387845 0.74313181 0.60199945 0.53836618 0.47341977  
   ## Proportion of Variance 0.03314129 0.02510204 0.01647288 0.01317446 0.01018756  
   ## Cumulative Proportion 0.91609209 0.94119413 0.95766701 0.97084147 0.98102903  
   ## Comp.11 Comp.12 Comp.13 Comp.14  
   ## Standard deviation 0.374920185 0.323792112 0.26407751 0.195361922  
   ## Proportion of Variance 0.006389325 0.004765515 0.00316986 0.001734831  
   ## Cumulative Proportion 0.987418353 0.992183869 0.99535373 0.997088560  
   ## Comp.15 Comp.16 Comp.17 Comp.18  
   ## Standard deviation 0.148363381 0.1333699311 0.1116082339 0.0849360690  
   ## Proportion of Variance 0.001000531 0.0008085245 0.0005661999 0.0003279153  
   ## Cumulative Proportion 0.998089091 0.9988976158 0.9994638157 0.9997917310  
   ## Comp.19 Comp.20 Comp.21 Comp.22  
   ## Standard deviation 0.0591317799 3.293866e-02 6.015269e-04 1.819996e-04  
   ## Proportion of Variance 0.0001589349 4.931616e-05 1.644703e-08 1.505629e-09  
   ## Cumulative Proportion 0.9999506659 1.000000e+00 1.000000e+00 1.000000e+00  
   ##   
   ## Loadings:  
   ## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9  
   ## MDVP.Fo.Hz. 0.553 0.128 0.131 0.115 0.148 0.217   
   ## MDVP.Fhi.Hz. 0.349 0.268 -0.241 0.188 -0.719 -0.370 -0.153  
   ## MDVP.Flo.Hz. 0.395 -0.233 0.220 0.286 0.468 -0.455   
   ## MDVP.Jitter... -0.255 -0.151 -0.252   
   ## MDVP.Jitter.Abs. -0.242 -0.185 -0.308   
   ## MDVP.RAP -0.250 0.116 -0.169 -0.258   
   ## MDVP.PPQ -0.257 -0.179 -0.157 0.142   
   ## Jitter.DDP -0.250 0.116 -0.169 -0.258   
   ## MDVP.Shimmer -0.260 0.242 -0.133   
   ## MDVP.Shimmer.dB. -0.262 0.207 -0.112   
   ## Shimmer.APQ3 -0.253 0.240 -0.198 -0.117   
   ## Shimmer.APQ5 -0.252 0.304 -0.103 0.132  
   ## MDVP.APQ -0.254 0.234 0.153  
   ## Shimmer.DDA -0.253 0.240 -0.198 -0.117   
   ## NHR -0.234 0.171 -0.274 -0.102 -0.296   
   ## HNR 0.242 -0.162 0.159 -0.271 0.477  
   ## RPDE -0.147 -0.249 0.353 -0.345 0.255 -0.511 -0.358  
   ## DFA -0.311 -0.466 0.287 0.391 -0.231 -0.520  
   ## spread1 -0.224 -0.239 0.152 0.226 0.403 0.275  
   ## spread2 -0.151 -0.204 0.349 0.460 -0.184 -0.629 0.123  
   ## D2 -0.155 0.133 0.428 0.308 0.123 0.573 -0.328  
   ## PPE -0.231 -0.214 0.267 0.434 0.247  
   ## Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16  
   ## MDVP.Fo.Hz. 0.645 0.107 0.256 0.137 0.221   
   ## MDVP.Fhi.Hz. -0.174   
   ## MDVP.Flo.Hz. -0.459   
   ## MDVP.Jitter... 0.110 -0.157   
   ## MDVP.Jitter.Abs. -0.139 0.257 0.153 -0.221 0.318 0.275 0.633   
   ## MDVP.RAP 0.120 -0.125 -0.233 -0.245   
   ## MDVP.PPQ 0.152 -0.191 -0.111 -0.222 0.278 -0.481 -0.206   
   ## Jitter.DDP 0.119 -0.125 -0.233 -0.246   
   ## MDVP.Shimmer   
   ## MDVP.Shimmer.dB. -0.159 -0.101 0.264   
   ## Shimmer.APQ3 0.338 0.179 -0.222   
   ## Shimmer.APQ5 0.356 -0.461 0.113   
   ## MDVP.APQ -0.663 -0.142 -0.158 0.417 0.112   
   ## Shimmer.DDA 0.338 0.178 -0.222   
   ## NHR -0.191 0.803 -0.117   
   ## HNR -0.132 -0.248 0.668 0.155   
   ## RPDE -0.197 0.388   
   ## DFA 0.190 0.172 0.215   
   ## spread1 -0.123 0.154 0.182 -0.549 -0.307 0.266   
   ## spread2 0.268 0.167 -0.165   
   ## D2 -0.330 -0.126 0.274   
   ## PPE 0.212 0.444 0.380 -0.322   
   ## Comp.17 Comp.18 Comp.19 Comp.20 Comp.21 Comp.22  
   ## MDVP.Fo.Hz.   
   ## MDVP.Fhi.Hz.   
   ## MDVP.Flo.Hz.   
   ## MDVP.Jitter... -0.164 0.446 -0.745   
   ## MDVP.Jitter.Abs. 0.104 0.219   
   ## MDVP.RAP 0.193 -0.346 -0.707   
   ## MDVP.PPQ -0.210 0.324 0.449 0.133   
   ## Jitter.DDP 0.194 -0.345 0.707   
   ## MDVP.Shimmer 0.895   
   ## MDVP.Shimmer.dB. -0.696 -0.460 -0.164   
   ## Shimmer.APQ3 0.213 0.114 -0.204 -0.707   
   ## Shimmer.APQ5 0.517 -0.216 -0.307 -0.200   
   ## MDVP.APQ 0.227 0.242 0.206 -0.160   
   ## Shimmer.DDA 0.213 0.114 -0.205 0.707   
   ## NHR   
   ## HNR   
   ## RPDE   
   ## DFA   
   ## spread1 0.107   
   ## spread2   
   ## D2   
   ## PPE -0.132 -0.186

   # Appendix C: Clustering Analysis R Code:

   Input Data and Create Distance Matrix

   mydata <- read.csv("cleaned\_data.csv", row.names='name')  
   mydata.s <- scale(mydata[2:23])  
   dist <- dist(mydata.s)

   Create Scree Plot  
   hc1 <- hclust(dist, 'complete')

   plot(rev(hc1$height), main="Scree Plot")

   Create Contingency Table for Complete Linkage

   hc1 <- hclust(dist, 'complete')  
   ct <- cutree(hc1, 4)   
   table(ct, mydata$status)

   Create Contingency Table for Complete Linkage

   hc2 <- hclust(dist, 'average')  
   ct <- cutree(hc2, 4)   
   table(ct, mydata$status)

   Create Contingency Table for Complete Linkage

   hc3 <- hclust(dist, 'single')  
   ct <- cutree(hc3, 4)   
   table(ct, mydata$status)

   # Appendix D: Confirmatory Factor Analysis Model and R Code:

   Model:

   Frequency -> MDVP.Jitter..., lambda1, NA

   Frequency -> MDVP.Jitter.Abs., lambda2, NA

   Frequency -> MDVP.RAP, lambda3, NA

   Frequency -> MDVP.PPQ, lambda4, NA

   Frequency -> Jitter.DDP, lambda5, NA

   Amplitude -> MDVP.Shimmer, lambda6, NA

   Amplitude -> MDVP.Shimmer.dB., lambda7, NA

   Amplitude -> MDVP.APQ, lambda8, NA

   Amplitude -> Shimmer.APQ3, lambda9, NA

   Amplitude -> Shimmer.APQ5, lambda10, NA

   Amplitude -> Shimmer.DDA, lambda11, NA

   Nonlin\_Freq -> spread1, lambda12, NA

   Nonlin\_Freq -> spread2, lambda13, NA

   Nonlin\_Freq -> PPE, lambda14, NA

   MDVP.Jitter... <-> MDVP.Jitter..., theta1, NA

   MDVP.Jitter.Abs. <-> MDVP.Jitter.Abs., theta2, NA

   MDVP.RAP <-> MDVP.RAP, theta3, NA

   MDVP.PPQ <-> MDVP.PPQ, theta4, NA

   Jitter.DDP <-> Jitter.DDP, theta5, NA

   MDVP.Shimmer <-> MDVP.Shimmer, theta6, NA

   MDVP.Shimmer.dB. <-> MDVP.Shimmer.dB., theta7, NA

   MDVP.APQ <-> MDVP.APQ, theta8, NA

   Shimmer.APQ3 <-> Shimmer.APQ3, theta9, NA

   Shimmer.APQ5 <-> Shimmer.APQ5, theta10, NA

   Shimmer.DDA <-> Shimmer.DDA, theta11, NA

   spread1 <-> spread1, theta12, NA

   spread2 <-> spread2, theta13, NA

   PPE <-> PPE, theta14, NA

   Nonlin\_Freq <-> Nonlin\_Freq, NA, 1

   Frequency <-> Frequency, NA, 1

   Amplitude <-> Amplitude, NA, 1

   Amplitude <-> Frequency, rho1, NA

   Amplitude <-> Nonlin\_Freq, rho2, NA

   Nonlin\_Freq <-> Frequency, rho3, NA

   Read Dataset

   voice <- read.csv('cfa\_final.csv')

   Create CFA Model using our defined model

   library(sem)  
   voice\_model <- specifyModel(file = 'voice\_model\_final.txt')

   voice\_sem <- sem(voice\_model, cor(voice), nrow(voice))

   Create SEM Plot

   library(semPlot);

   semPaths(voice\_sem, 'est')

   View Fit Indices and measure against Thresholds

   options(fit.indices = c("GFI", "AGFI", "SRMR")) # Some fit indices   
   criteria = summary(voice\_sem)   
     
   criteria$SRMR

   criteria$GFI

   criteria$AGFI

   criteria$SRMR < 0.05

   criteria$GFI > 0.95

   criteria$AGFI > 0.95 [↑](#endnote-ref-5)
7. Multi-Dimensional Voice Program [↑](#footnote-ref-2)