Parkinson's Disease: EFA ja Classification Analysis

Practical Work for DATA.STAT.450 Monimuuttujamenetelmät

Aada Kerkkonen

22.2.2025

Table of Contents

Introduction

Parkinson's disease (PD) is a neurodegenerative disease that causes tremor, stiffness and other motor issues. As the disorder progresses, the patient may also experience cognitive decline: mental health issues, hallucinations and issues with attentiveness as well as speech and other auditory issues. For example, people with Parkinson's disease tend to speak monotonously, sometimes stuttering or forgetting words (Michael J. Fox Foundation 2024).

This project will analyse voice-related features in PD using Exploratory Factory Analysis (EFA) and Classification Analysis. The former will be used in hopes of discovering latent variables among the data and reducing redundancy within the original dataset. These results will then be used as input for the Classification Analysis. The aim is to classify the observations into two groups: diagnosed with Parkinson's and not.

Data

The data for this project consists of 195 observations of different biomedical voice measurements of 31 individuals (Little 2007). Out of those 31, 23 are diagnosed with Parkinson's disease with time since diagnosis ranging between 0 and 28 years. The data was originally used for a study classifying people (46-85yo) with Parkinson's from those without the disease by detecting dysphonia (Little et al. 2009). Different phonations were recorded from the subjects, ranging between a second and 36 seconds in length. To assure that the measures would be comparable to those of other studies in the area, the researchers utilised a computer program called MVDP (Multidimensional Voice Program Analysis).

The data consists of 24 variables: one for identification of the subject (name), a boolean variable for the diagnosis (status) while the rest of the variables are measurements of the phonations. These can be divided into six categories:

1. Frequency-Related measures in Hertz (relating to the pitch of the voice)
    - MDVP:Fo - the average vocal fundamental frequency
    - MDVP:Fhi - maximum vocal fundamental frequency
    - MDVP:Flo - minimum vocal fundamental frequency
2. Variation in Frequency (jitter, i.e. variance in timing and stability of vocal vibrations)
    - MDVP:Jitter(%)
    - MDVP:Jitter(abs)
    - Jitter:DDP - difference between cycles
    - MDVP:RAP - Relative Amplitude Perturbation
    - MDVP:PPQ - Period Perturbation Quetient
      $\rightarrow$ High jitter value indicates unstable voice quality: can be a symptom of PD
3. Amplitude Variation (shimmer, i.e. variations in the loudness of vocal vibrations)

- MDVP:Shimmer
- MDVP:Shimmer(dB)
- Shimmer:DDA - difference between cycles
- Shimmer:APQ3 - 3 point Amplitude Perturbation Quetient
- Shimmer:APQ5 - 5 point Amplitude Perturbation Quetient
- MDVP:APQ - 11 point Amplitude Perturbation Quetient
  → Higher shimmer value generally indicates irregulaties in loudness, i.e. can be a symptom of PD

4. Noise-To-Harmonics (noise to tonal components, indicating the clarity of the sound)
   - NHR - Noise To Harmonics Ratio
   - HNR - Harmonics to Noise Ratio
     → A high NHR and low HNR can indicate PD
5. Nonlinear Measures of Fundamental Frequency Variation
   - spread1
   - spread2
   - PPE (Pitch Period Entropy, stability of speech)
     → Higher values can indicate instability in the speech, i.e. PD
6. Other Variables (structure of the speech)
   - RPDE - randomness of pitch patterns
   - D2 - complexity of speech signal
   - DFA - signal fractal scaling component
     → Can indicate chaotic structure in the speech, i.e. PD

Because the dataset consists of many different measurements of e.g. jitter and shimmer, their respective correlations were analysed to infer whether we could already reduce redundant variables as well as the dimensionality of the data. Additionally, status is converted into a factor.

```
data <- read.table("parkinsons.data", sep = ",", header = TRUE)

# converting status into factor
data$status <- as.factor(data$status)

# jitter variables
cor(data[, c(5:9)])
```

```
##                  MDVP.Jitter... MDVP.Jitter.Abs.  MDVP.RAP  MDVP.PPQ Jitter.DDP
## MDVP.Jitter...        1.0000000        0.9357140 0.9902756 0.9742564  0.9902762
## MDVP.Jitter.Abs.      0.9357140        1.0000000 0.9229110 0.8977779  0.9229130
## MDVP.RAP              0.9902756        0.9229110 1.0000000 0.9573169  0.9999996
## MDVP.PPQ              0.9742564        0.8977779 0.9573169 1.0000000  0.9573192
## Jitter.DDP            0.9902762        0.9229130 0.9999996 0.9573192  1.0000000
```

As can be seen, all the variables are very highly correlated. Because of this, only one of the jitter variables was kept for analysis: MDVP.Jitter. This variable was chosen because the mean correlation in its column is the highest. Therefore, all the jitter variables correlate the most with MDVP.Jitter.

The same procedure was repeated for the shimmer variables:

```
# shimmer variables
cor(data[, c(10:15)])
```

```
##                  MDVP.Shimmer MDVP.Shimmer.dB. Shimmer.APQ3 Shimmer.APQ5
## MDVP.Shimmer        1.0000000        0.9872578    0.9876251    0.9828354
## MDVP.Shimmer.dB.    0.9872578        1.0000000    0.9631981    0.9737506
## Shimmer.APQ3        0.9876251        0.9631981    1.0000000    0.9600698
## Shimmer.APQ5        0.9828354        0.9737506    0.9600698    1.0000000
## MDVP.APQ            0.9500829        0.9609767    0.8966445    0.9491461
## Shimmer.DDA         0.9876257        0.9632017    1.0000000    0.9600716
##                   MDVP.APQ Shimmer.DDA
## MDVP.Shimmer     0.9500829   0.9876257
## MDVP.Shimmer.dB. 0.9609767   0.9632017
## Shimmer.APQ3     0.8966445   1.0000000
## Shimmer.APQ5     0.9491461   0.9600716
## MDVP.APQ         1.0000000   0.8966468
## Shimmer.DDA      0.8966468   1.0000000
```

We can see that the variables are again very highly correlated. Taking the variable with the highest mean: MDVP.Shimmer. Now we can remove the redundant variables from the dataset and create another dataset that only contains numerical values for EFA. Therefore, the name and status columns are removed from that set.

```
# removing redundant variables
filtered_data = data[, c(-(6:9), -(11:15))]
test_data = filtered_data[, c(-1, -9)]
dim(test_data)
```

```
## [1] 195  13
```

We are left with 13 variables and 195 observations.

Methodology

The methods in this project consist of Exploratory Factory Analysis and Classification Analysis. The information in this section is taken from the book *Methods of Multivariate Analysis* (Rencher and Christensen 2012) and lecture slides.

Exploratory Factory Analysis

EFA tries to find underlying constructs, factors, that can generate the original variables in the data set through their correlations or covariances. The aim is to reduce the redundancy among the dataset by generating fewer factors than variables in the original set and express y_i as linear combinations of these factors. The model can be expressed as

$$\mathbf{y} - \mu = \Lambda\mathbf{f} + \epsilon$$

where

- y is the observed data
- μ is the mean vector of the observed data
- Λ contains the loadings that show how each y_i depends on the factors
- f is the vector of the underlying constructs, factors
- ϵ is the random error

We can estimate the loadings through two main methods: Principal Factor Analysis or Maximum Likelihood Factor Analysis. This project uses the Maximum Likelihood method (MLE) because it can be used in hypothesis testing for the covariance structure

$$H_0 : \Sigma = \Lambda\Lambda' + \Psi$$

where

- Λ is as defined above
- Ψ is $cov(\epsilon)$ where the variances are of each ϵ_i are on the diagonal and every other cell is 0.

The p-value corresponding to this hypothesis testing can be found in the printout. Because the p-value is the probability of a perfect fit between the source data and number of factors, we want a larger number (Magazine 2017).

The result of EFA is not unique, and can be multiplied by an orthogonal matrix $\mathbf{G}$ ($\mathbf{GG}' = \mathbf{I}$) to rotate it and improve interpretation or structure as follows

$$\mathbf{y} = \Lambda\mathbf{GG}'\mathbf{f} + \epsilon = \Lambda_*\mathbf{f}_* + \epsilon$$

The process of finding the best possible $\mathbf{G}$ is called rotation. There are different methods of that will be tried in order to find the best rotation for this dataset. In addition, we need to determine how many factors to use in the model. Both of these will be assessed by the cumulative variance; we want to use the model with the highest cumulative variance.

In R, the analysis can be performed with the function *factanal()* that takes in either the dataset, covariance or correlation matrix, the number of factors, the method chosen for the estimation of loadings and rotation. Additionally, we include the parameter *scores* to obtain the factor scores used for the classification analysis and this result was saved in a separate variable. These scores come from the conditional distribution of $\mathbf{f}$ given $\mathbf{y}$.

In this project, the whole dataset was used as input for the model since it is rather small. The *factanal* call was repeated for different rotations and number of factors to obtain the highest cumulative variance.

```
results_efa <- factanal(test_data, factors = 6, method = "mle",
  rotation = "varimax", scores = "regression", n.obs = 195)

factor_scores <- results_efa$scores
```

Classification Analysis

The goal of classification analysis is to describe group separation through discriminant functions (discriminant analysis) and then assign observations to these groups (classification). In discriminant analysis, a linear combination transforms each observation vector into a scalar after which the means of the two groups are calculated.

$$z = \mathbf{a}'y \text{ and } \overline{z}_i = \mathbf{a}'\overline{y}_i$$

We seek to find $\mathbf{a}$ that maximises standardizes the difference between $\overline{z}_1$ and $\overline{z}_2$. The maximum is known to occur at

$$\mathbf{a} = S_{pl}^{-1}(\overline{y}_1 - \overline{y}_2)$$

where

$$S_{pl} = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)S_1 + (n_2 - 1)S_2).$$

Then to classify the observations into the two groups, we calculate the discriminant scores of each observation and then the means of the groups as defined above. To make the classification, we check whether the z is closer to $\overline{z}_1$ or $\overline{z}_2$.

$$z > \frac{1}{2}(\overline{z}_1 + \overline{z}_2)$$

If z is larger as shown above, it belongs to group 1 and vice versa.

Additionally, because we want to test how the classification analysis actually performs on unseen data, we split the data into test and training groups. In this project, the model was trained with 80% of the data and tested with the rest. In R, the package *mclust* provides a function to perform both discriminant and classification analysis. The parameters of *MclustDA* are the training set and the classes of those observations, the type of the model and method of discriminant analysis. Here we have chosen EDDA (Eigenvalue Decomposition Discriminant Analysis) as model type and VVV (quadratic discriminant analysis) as the special case. VVV was chosen because it assumes that each class have their own covariance matrices, which is beneficial here as the different groups may have very different covariances.

```
library(mclust)
```

```
## Warning: package 'mclust' was built under R version 4.4.2
```

```
## Package 'mclust' version 6.1.1
## Type 'citation("mclust")' for citing this R package in publications.
```

```
# splitting the data
classification_data <- data.frame(factor_scores, status = filtered_data$status)
nc <- ncol(classification_data)
nr <- nrow(classification_data)

# seed set for reproducibility
set.seed(42)
train_indices <- sample(1:nr, size = 0.8 * nr)

train <- classification_data[train_indices, -nc]
train_class <- as.factor(classification_data$status[train_indices])

test <- classification_data[-train_indices, -nc]
test_class <- as.factor(classification_data$status[-train_indices])

# producing model
results_clas <- MclustDA(train, train_class,
                         modelType = "EDDA", modelNames = "VVV")
```

Results

EFA

Let's first look through the results of the Exploratory Factory Analysis. The chosen rotation was varimax and the number of factors in the final model 6. This reached a cumulative variance between all the 6 factors of 80.5% and it was the highest obtained with different numbers of factors and other rotations (promax and none)

```
results_efa
```

```
##
## Call:
## factanal(x = test_data, factors = 6, n.obs = 195, scores = "regression",     rotation
##
## Uniquenesses:
##    MDVP.Fo.Hz.   MDVP.Fhi.Hz.   MDVP.Flo.Hz. MDVP.Jitter...   MDVP.Shimmer
##          0.005          0.761          0.534          0.023          0.223
##            NHR            HNR           RPDE            DFA        spread1
##          0.075          0.005          0.350          0.019          0.024
##        spread2             D2            PPE
##          0.476          0.005          0.036
##
## Loadings:
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## MDVP.Fo.Hz.             -0.156   0.918  -0.331   0.110
## MDVP.Fhi.Hz.    0.127           0.319  -0.327
## MDVP.Flo.Hz.            -0.188   0.641                  -0.106
## MDVP.Jitter...  0.916   0.354
## MDVP.Shimmer    0.668   0.314           0.139   0.227   0.400
## NHR             0.913   0.156          -0.178   0.153   0.101
## HNR            -0.610  -0.277                  -0.306  -0.670
## RPDE            0.221   0.342  -0.442  -0.173           0.508
## DFA                     0.137  -0.100   0.971
## spread1         0.407   0.806  -0.298           0.167   0.199
## spread2         0.166   0.557  -0.185           0.371
## D2              0.273   0.310          -0.161   0.880   0.130
## PPE             0.443   0.792  -0.221   0.154   0.150   0.213
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## SS loadings      3.027   2.204   1.747   1.305   1.164   1.018
## Proportion Var   0.233   0.170   0.134   0.100   0.090   0.078
## Cumulative Var   0.233   0.402   0.537   0.637   0.727   0.805
##
```

```
## Test of the hypothesis that 6 factors are sufficient.
## The chi square statistic is 42.54 on 15 degrees of freedom.
## The p-value is 0.000185
```

Let's interpret the results. Uniqueness corresponds to Ψ, proportion that is not explained by the variables, and $1 - \Psi$ gives the communality: the proportion of the variance that the variable contributed to the factors (Magazine 2017). Therefore, we can see that the variables with high uniqueness, such as MDVP.Fhi.Hz, MDVP.Flo.Hz and spread2, are poorly explained by the 6 factors. For example, only 23.9% of the variation in MDVP.Fhi.Hz was retained with the factors. In contrast, the factors, MDVP.Fo.Hz, MDVP.Jitter, NHR, HNR, DFA, spread1, D2 and PPE, with low uniqueness are well explained by the factors in the model. For example, 99.5% of MDVP.Fo.Hz is explained with the factors. Therefore, the model is not capturing all the variation in the variables and is thus not a perfect fit for the data.

As mentioned in the methodology section, loadings represent how each variable depends on the factors. We can see that the first factor accounts largely for MDVP.Jitter, MDVP.Shimmer, NHR and HNR (in the opposite direction of NHR). This indicates that the first factor captures a lot of the variability in the sound wave: amplitude and period of the wave and the clarity of the sound. The second factor has the highest scores for spread1, spread2 and PPE which means that the factor accounts for the variation in frequency and thus the stability and regularity of the speech. The third factor has the highest scores for MDVP.Fo.Hz and MDVP.Flo.Hz, so the factor accounts frequency (pitch) of voice. For factor 4 and 5 the highest values are for DFA (factor 4) and D2 (factor 5). Therefore, these two factors largely account for the complexity in speech. Finally, factor 6 has largest scores in HNR and RPDE which could be interpreted as relating to the randomness of pitch and the amount of noise. All together, we can see that the factors identify similarities in the measurements that closely resemble those groups explained in the section Data. The proportion of variance explains how each factor contributes to the cumulative variance of the model. We can see that even the final factor adds almost 8% to the cumulative proportion of variance explained.

Finally, because we used MLE, the printout also includes a p-value. Because this p-value is so small, we cannot reject the null hypothesis. This means that the model does not perfectly fit the data and that more factors could be needed.

Classification Analysis

Moving on to the classification analysis based on the factor scores obtained in EFA.

```
summary(results_clas, newdata = test, newclass = test_class)
```

```
## ----------------------------------------------
## Gaussian finite mixture model for classification
## ----------------------------------------------
```

```
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -1277.249 156 55 -2832.241
##
## Classes   n     % Model G
##       0  37 23.72   VVV 1
##       1 119 76.28   VVV 1
##
## Training confusion matrix:
##      Predicted
## Class   0   1
##     0  29   8
##     1  14 105
## Classification error = 0.141
## Brier score          = 0.0907
##
## Test confusion matrix:
##      Predicted
## Class  0  1
##     0  9  2
##     1  6 22
## Classification error = 0.2051
## Brier score          = 0.1467
```

The results indicate that the training set had 156 observations of which 119 were of observations of subjects diagnosed with PD. After performing discriminant analysis based on the training set, the parameters found were used for classifying the training set. The training confusion matrix shows how the model classified the observations. Out of 37 observations that belonged to subjects without PD, the model predicted 29 (78.4%) correctly. In comparison, out of 119 with PD, the model predicted 105 (88.2%) correctly. This gives the classification error of 14.1% where lower scores indicate better performance. Therefore, there is a lot of room for improvement in the score. This score could indicate that the data in itself has a lot of variety. This would be a reasonable interpretation because auditory issues in PD appear as the disease progresses and the time since diagnosis between subjects in the dataset varied between 0 and 28 years.

With unseen data, the classification error is higher, 20.5%. This indicates that the model is slightly overfitting the training data and does not perform as well with new data. The model successfully classified 9/11 (81.8%) correctly into the group of subjects without PD and 22/28 (78.6%) into the PD group. The brier score supports our findings as it is higher for the unseen data.

Conclusion

The project aimed to identify latent variables in Parkinson's data through EFA and use these results as input in Classification Analysis to classify observations into those with Parkinson's disease and those without.

The EFA model successfully accounted for 80.5% of the variance in the dataset with 6 factors. We could see that similar variables were grouped together by the factors where the first factor highly accounted for jitter, shimmer and noise in the voice, second and third to pitch and frequency, fourth and fifth to complexity and finally the sixth factor HNR so the amount of noise in the voice, indicating speech degradation. While the proportion of variance explained by the model is rather high, the low p-value and uniqueness scores indicate that the model was not the perfect fit. I could not add more factors to improve the performance of the model, so alternative methods are needed to address this. For example, the data could be analysed more carefully with an expert in PD and auditory problems in PD before EFA to further reduce variables and identify the necessary ones. Although plenty of research was done, I was not able to perfectly grasp the meanings and contributions of all the variables in the dataset.

Classification Analysis successfully classified 78.6% of the subjects with PD correctly and 81.8% for those without. While this is a satisfactory score, it leaves room for a lot of improvement and suggests overfitting the training data. Because the input for the Classification Analysis was the factor score found through EFA, the classification model did not attain all the information of the data. For example, some features that would have improved the performance may have been lost. Furthermore, since the time since diagnosis for those with PD varied between 0 and 28 years, that could also account largely for the variability and be a reason why both the factors and the classification analysis where not able to perform perfectly. This could be included in the dataset and thus provide better results for both EFA and classification analysis. Additionally, ovefitting could be addressed by using k-fold validation in place of just separating the data into training and test sets.

In conclusion, the models performed reasonably well but will need adjustments for better performance. These include contextualising the variables and their meaning in the data set, adding a variable for time since diagnosis and using k-fold validation to avoid overfitting.

Sources

Little, Max. 2007. "Parkinsons." UCI Machine Learning Repository.

Little, Max, Patrick Mcsharry, Eric Hunter, Jennifer Spielman, and Lorraine Ramig. 2009. "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease." *IEEE Transactions on Bio-Medical Engineering* 56 (April): 1015. https://doi.org/10.1109/TBME.2008.2005954.

Magazine, Visual Studio. 2017. "Revealing the Secrets of r Factor Statistics." *Visual Studio Magazine*. https://visualstudiomagazine.com/articles/2017/03/01/revealing-secrets-r-factor-statistics.aspx.

Michael J. Fox Foundation. 2024. "Speech and Swallowing Problems in Parkinson's." https://www.michaeljfox.org/symptoms/speech-swallowing-problems#:~:text=Many%20people%20with%20PD%20speak,rapidly%2C%20even%20stuttering%20or%20stammering.

Rencher, Alvin C., and William F. Christensen. 2012. *Methods of Multivariate Analysis*. 3rd ed. Wiley Series in Probability and Statistics. Hoboken, N.J: Wiley.