|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | For office use only | | | T1 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | T2 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | T3 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | T4 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | |  | | --- | | Team Control Number **96584** | |  | | Problem Chosen **C** | | |  |  | | --- | --- | | For office use only | | | F1 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | F2 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | F3 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | F4 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |   **Summary** |

# Introduction

## 1.1 Background

Parkinson's disease (PD) is a chronic central nervous system disease that affects the patient's ability to move. In the early stage, the most obvious symptoms are shaking, rigidity, slowness of movement and difficulty with walking. As time goes by, thinking and behavioral problems may occur. Give that those symptoms are easily confused with symptoms of aging, it requires an efficient method to identify this disease.

The Unified Parkinson’s Disease Rating Scale (UPDRS) is the most commonly used scale in the clinical study of Parkinson's disease and provides a reliable measure of disease severity. However, the UPDRS is highly examiner-dependent and less reliable with nonmovement trained neurologists. As an objective, noninvasive and easy way, computerized spiral analysis can automatedly detect PD. All the patients need to do is drawing spirals in different conditions and the data collected and analyzed by computer will show the possibility of PD.

## 1.2 Restatement of the Problem

We are required to analyze and compare the attached data and images and give the differences between the patient and non-patient in the record data and the classification and classification criteria. The problem is analyzed into three parts:

* Analyze the data information and estimate an ideal spiral.
* Extract features which distinguishes patients and non-patients.
* Explore a classification and recognition methods to detect Parkinson’s disease.

## 1.3 Results

After processing the data, we employ the PCA (Principal components analysis) method to build a comprehensive evaluation model. From four indexes which are smoothness, pressure, velocity and the frequency of repeated points, we can obtain the overall score and distinguish patients from non-patients numerically. In other words, what differentiate patients from non-patients are the scores of different indexes. In the first and second situation where ID is 0 and 1, we conclude that the final scores of most patients are above -2 while majority of the non-patients are below -2.

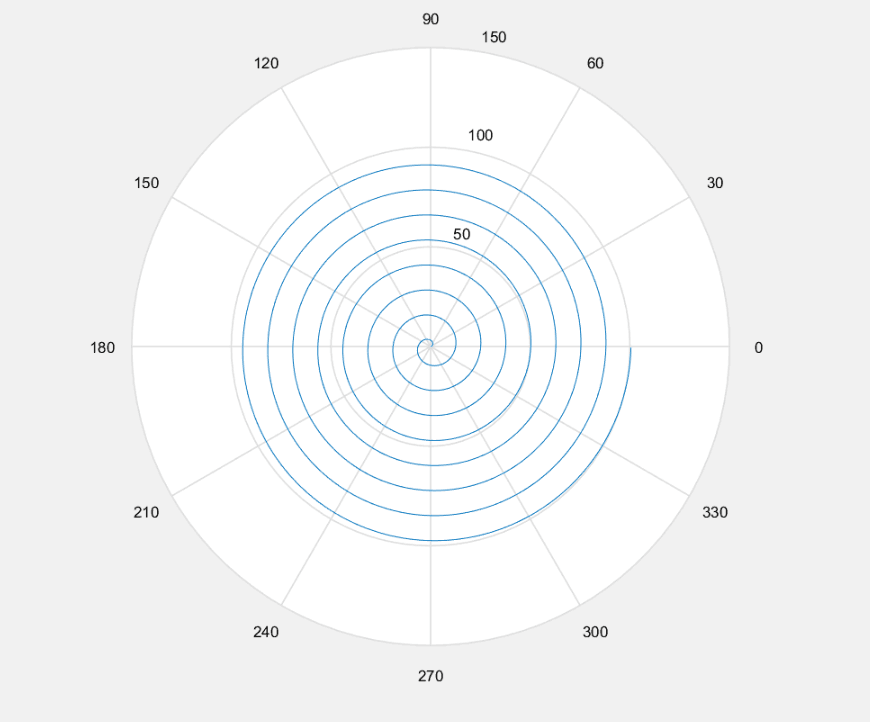
# The Description of the Problem

## 2.1 How do we define an ideal spiral?

First, we consider a spiral as a two-dimensional image, which means an ideal spiral does not contain information in three dimensions like pressure or the dynamic information like velocity. Subtracting these two features, we define an ideal spiral with two other features we extracting from the given data, smoothness and frequency of repeated points. The latter condition is easy to meet. Therefore, our requirement of an ideal spiral is smoothness. That leads to us to the Archimedean spiral. It can be depicted with polar equation as



It can be drawn in polar coordinates to illustrate its smoothness.



**Figure 1**

## 2.2 How do we differentiate patients from non-patients?

The most obvious symptoms of PD are shaking, rigidity, slowness of movement and difficulty with walking. We exact four indexes from data to measure those symptoms, smoothness, pressure, velocity and frequency of repeated points. The details of these four indexes are fully discussed in III. Models. Because the data in the attachment is given without units, we treat it as numbers instead of physical quality. Besides, there is a negative correlation between possibility of PD and scores of four indexes. In the light of this, we can differentiate patients from non-patients by comparing their overall scores.

# Model

## 3.1 Data preprocessing

### 3.1.1 Assumption

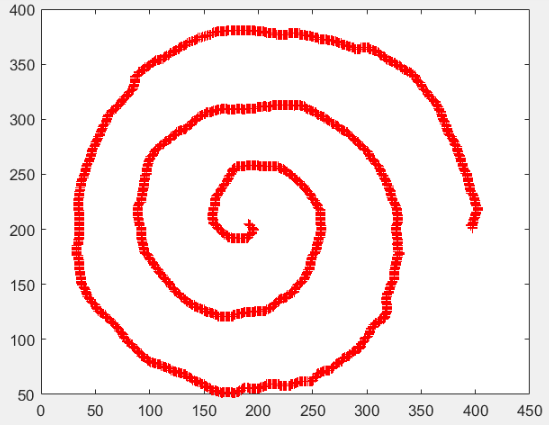
* The z-axis data has little impact on our classification results. (Most Z-axis data is 0).
* Majority of the data in the attachment are authentic

### 3.1.2 Symbols and definition:

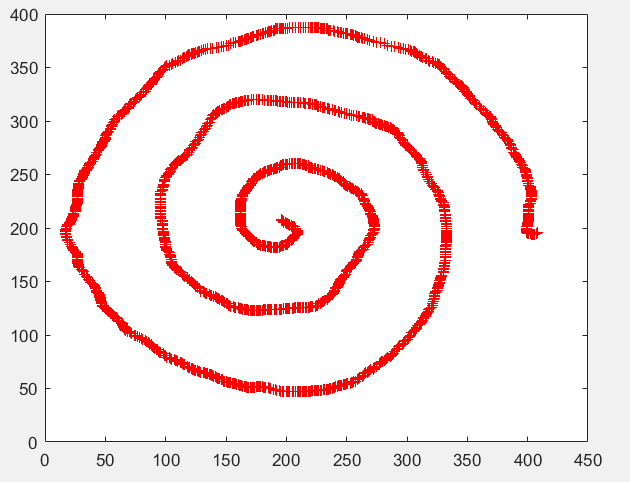
|  |  |
| --- | --- |
| Symbol | Definition |
| S | smoothness |
| P | pressure |
| V | velocity |
| F | frequency of repeated points |
| Cr | difference of the polar diameter |
| Cx | difference of the x |
| Cy | difference of the y |
| Acr | the average value of absolute value of the Cr |
| c | the difference value between Cr and Acr |
| Sp | the square of Ar |
| Avz | the average value of absolute value of the difference between V and Z |

### 3.1.3 Extract features

#### Smoothness



**The spiral of a patient (ID = 0)**



**The spiral of a patient (ID = 0)**

From these two pictures drawn by a non-patient and a patient, we can see that the spiral of the non-patient is smoother than patient. We highlight the area where the spiral curve is rougher by enclosing it with blue rectangles in the spiral of the patient. These obvious difference between the spiral of the non-patient and patient can be defined as whether it’s smooth. When it comes to the dynamic test where the ID is 1, this kind of difference still exists.

The smoothness can be mathematically depicted as how close its polar radius in polar coordinates remains to its mean. We measure this index with its variance. In short, this index can be depicted as

……

Where

S is the smoothness index

is the polar radius of each point (centered at original point)

is the difference between adjacent points

is the mean of the

#### velocity

When Alzheimer’s patients draw spirals, their hands shake frequently. Therefore, speed of their spiral often mutates. That is to say, their speed transition is rougher than non-patients, which indicates the variance of their speed is larger than that of non-patients.

We calculate the algebraic distance between adjacent points.

……

……

Where

And we calculate the variance of all data of patients and non-patient. Some results (I choose 10 data) are as follows:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1408.1 | 1589. | 1989.7 | 1385.2 | 1482.1 | 1597.5 | 1831.3 | 1647.2 | 1494.2 | 1603.4 |

**10 patients**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 729.5 | 442.4 | 587.3 | 931.7 | 724.4 | 1079.4 | 988.6 | 1018.1 | 1100.7 | 1010.5 |

**10 non-patients**

From the data above, the variance of patients and non-patient differ greatly, which means this feature can reflect the difference between patients and non-patient well.

#### Standard deviation of pressure

Previous studies have found that patients with Parkinson's disease cannot control their strength well. Therefore, we will find the difference by comparing the difference between the patients’ and nonpatients’ pen power changes.

Since the pressure values ​​collected for each individual are constantly changing, it is difficult to make a clear quantitative comparison between different data. In this regard, we take the degree of change in the pressure data of each tester, that is to say, take the standard deviation of the pressure value as an index.

#### Frequency of repeated points

Since Parkinson's disease is a central nervous system deformity disease and its main symptoms are limb stiffness, tremor and slow movement. These symptoms have appreciative effects on the test results.

For example, if the action is sluggish, a large number of repeated points will appear in the measured coordinate data. Even with a sluggishness of 0.1s, there will be nearly 10 coincidences, and the longer the time, the more will repeated points appear. The trembling of the limbs may change the orbit of the curve and consequently the intersection occurs. Because the frequency of intersections is relatively low, we treat the two as Frequency of overlap. The frequency data is directly processed by the computer. The formula and processing results of the frequency are as follows.

## 3.2 Comprehensive evaluation Model

### 3.2.1 Principal component analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. It is a powerful tool when it comes to dimensionality reduction, which makes PCA a regular procedure in multidimensional data processing. However, we also use PCA in clustering in a specific way.

### 3.2.2 data processing

By preprocessing the given data, we have obtained a matrix of four features.

|  |  |  |  |
| --- | --- | --- | --- |
| smoothness | velocity | pressure | frequency of  repeated points |
| 1.4512 | 831.255 | 145.5041 | 0.857434 |
| 0.8124 | 1223 | 166.7957 | 0.278366 |
| 1.1612 | 1613.7 | 66.852 | 0.765708 |
| 0.537 | 805.7778 | 117.8877 | 0.648475 |
| 0.6394 | 2572.5 | 136.622 | 0.544619 |
| 0.5045 | 1139.5 | 62.8349 | 0.665619 |
| 0.6145 | 2404.6 | 153.7102 | 0.608897 |
| 0.5981 | 839.5185 | 123.24 | 0.673514 |
| 0.8953 | 842.2979 | 116.8845 | 0.624948 |
| 1.0751 | 1736.8 | 125.8346 | 0.718076 |
| 1.4951 | 970.4201 | 149.9644 | 0.907003 |
| …… | …… | …… | …… |

**Table 1 ID=0**

1. Step 1: Data standardization

Because the data in the attachment is given without units, numbers of each index differ greatly. Therefore, we standardize the data to avoid this disturbance. The process can be depicted as



Where

 is the standardized value

 is the original value

 is the mean value of original data

 is the variance of original data

The processed data is illustrated in Table 2

|  |  |  |  |
| --- | --- | --- | --- |
| smoothness | velocity | pressure | frequency of  repeated points |
| 1.253 | -0.9093 | 1.3582 | 1.3665 |
| -0.1519 | -0.0575 | 1.9923 | -1.3163 |
| 0.6152 | 0.792 | -0.9841 | 0.9415 |
| -0.7576 | -0.9647 | 0.5358 | 0.3984 |
| -0.5324 | 2.8767 | 1.0937 | -0.0828 |
| -0.8291 | -0.2391 | -1.1037 | 0.4778 |
| -0.5872 | 2.5117 | 1.6026 | 0.215 |
| -0.6233 | -0.8913 | 0.6952 | 0.5144 |
| 0.0304 | -0.8853 | 0.5059 | 0.2894 |
| 0.4258 | 1.0597 | 0.7725 | 0.7208 |
| 1.3495 | -0.6067 | 1.4911 | 1.5961 |
| …… | …… | …… | …… |

**Table 2 ID=0**

1. Step 2: calculate eigenvalue and eigenvectors

The results are displayed in Table 3&4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Y1 | Y2 | Y3 | Y4 |
| S | 0.1594 | 0.8547 | -0.1733 | -0.4627 |
| P | 0.5503 | 0.0386 | 0.8325 | -0.0509 |
| V | 0.6279 | 0.1617 | -0.3824 | 0.6583 |
| F | 0.5267 | -0.4918 | -0.3616 | -0.5916 |

Table 3 eigenvectors

In the table

S is smoothness P is pressure

V is velocity F is frequency of repeated points

Y1,,Y2,Y3,Y4 are principal components

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Y1 | Y2 | Y3 | Y4 |
| λ | 1.6525 | 1.1373 | 0.7164 | 0.4937 |

Table 4 eigenvalue

In the table

λ is the eigenvalue

Step 3: choose principle components and calculate the comprehensive index

Calculate the contribution rate of indexes.



Where

λ1,λ2,λ3 are corresponding eigenvalue of Y1,Y2,Y3.

The former three principle components have carried over 80 percent of original information. Consequently, we choose the former three components as our basis of final score.

1. Step 3:Calculate the score of each index



Putting in corresponding value of original index, we get a score matrix.

1. Step 3:Calculate the comprehensive index

We make use of the contribution rate of each component we choose and recalculate them again as their weight in the new index.



Where

λ1,λ2,λ3 are corresponding eigenvalue

W1,W2,W3 are corresponding eigenvalue

In different situations where ID = 0 or 1, we draw two scatter diagrams(Figure 2&3) to see the boundary between patients and non-patients on the data.

The orange line divided the former 25 patients’ scores and 15 non-patients’ scores into two parts. The red line is set at the value of -2.

When ID is 0

With the assistance of the two lines, it’s obvious to notice that final scores of 24(25 all) patients are above -2. Meanwhile scores of 10(15 all) non-patients are below -2. This means our model is reliable to detect patients but likely to make mistakes when judging non-patients.

When ID is 1

Similarly, scores of 24 patients (25 all) are above -2 and 7 (15 all) non-patients are below -2. It demonstrates that in the dynamic spiral tests, using our model are much more likely to identify non-patients as patients.

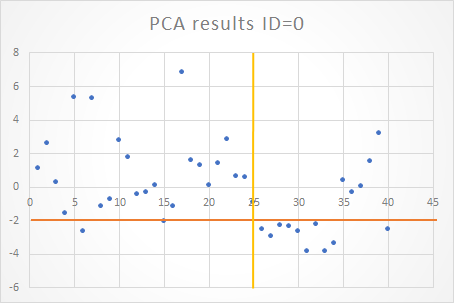


Figure 2

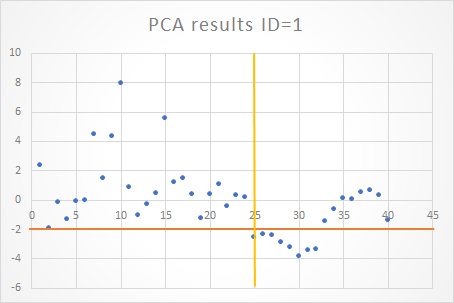


Figure 3

# Discussion

Our spiral analyses use smoothness, velocity, pressure, and frequency of repeated points as characteristics. W change continuous statistics to discrete statistics. This is a significant step, because these continuous data cannot be used directly to go with Principal Component Analysis. The limitation of our model is the limited data of patients and non-patient. For the further study, we need to collect tremendous reliable data to test our model and improve our model.

In conclusion, our spiral analysis can be a convenient and simple diagnosis tool in differing non-patient and patients.

# Future work

## 4.1 Make full use of the data

During the processing of the data, we extracted four eigenvalues, namely the smoothness, velocity, pressure, and frequency of the overlapping points. Although these four eigenvalues ​​can better reflect the difference between the sample data of healthy people and patients, the process of extracting data inevitably loses some information, making the use of data insufficient. For further use of data, we have the following points:

(1) Including the parameters such as the z value that are not fully utilized into the discriminating basis, and improving the accuracy of the discrimination.

In the modeling process, we generally extract features that have a greater impact and discard the less effective factors. But in contrast, supplementing the impact of various factors will help improve the accuracy of the identification. And in our modeling process, we also have difficulty in determining that the amount discarded is a less influential factor. Therefore, fully considering the influence of various factors will make our model more perfect.

(2) extract more difference features and reduce data loss

The situation reflected by two or three eigenvalues ​​is often insufficient. To make the data more detailed, it is necessary to make comprehensive considerations from multiple angles.

(3) Compare the accuracy of other algorithm discriminants, and consider using other classification models for identification.

Due to the increase in the number of features, it may be difficult to reflect the difference between different samples by using the comprehensive score in the principal component analysis as a basis for discrimination. Finding the best solution by comparing the accuracy of other algorithms would be a good choice.

## 4.2 Another model

The detection problem of Parkinson's disease can be classified as a pattern recognition problem, that is, the data of the detected tester is identified to determine whether it is sick. Therefore, the model can also be solved by the BP neural network algorithm. The BP network continuously adjusts the weights and thresholds of the network through backpropagation to minimize the sum of squared errors of the network, thereby classifying the input data.

Since the raw data of each person detected is constantly changing, it is difficult to make quantitative comparisons. Therefore, in this algorithm, we also need to extract the features first. After extracting the four main difference features as described above, we divided the data into two categories, a training set and a test set. After preprocessing the data, with the Matlab neural network toolbox, we input the matrix composed of feature components of the training set, and the BP network will get the data classification basis after the training data. It will be able to discriminate which type of sample data the test set belongs to, that is, whether it is sick or healthy, so that we have a new identification method.