Functional and Topological Data Analysis Course

Functional Data Analysis of the Visual Acuity and Signals Changes Across Selected Diagnosis

Dataset:

A Comprehensive Dataset of Pattern Electroretinograms (PERG) for Ocular Electrophysiology Research

Professor : Alessandra Micheletti

Shojaat Joodi Bigdilo
14088A

July 2024

1. Introduction

Ocular diseases and disorders significantly impact the quality of life, and accurate diagnostic techniques are essential for effective treatment and management. In this comprehensive project, we investigate how functional data analysis (FDA) can be applied to pattern electroretinograms (PERG) to enhance our understanding and diagnostic capabilities of various ocular conditions. Our goal is to explore the functional relationship between PERG signals and different ocular diseases using the PERG-IOBA dataset.

Pattern electroretinogram (PERG) is a critical tool in ophthalmic electrophysiology, offering valuable insights into the functioning of the central retina and retinal ganglion cells. By analyzing the electrical responses elicited by visual stimuli, PERG helps distinguish between macular and optic nerve diseases, which is often challenging through standard clinical examinations. As advancements in electrophysiological signal processing and analysis emerge, FDA provides a robust framework to interpret these complex signals and uncover deeper insights into ocular health.

1.1 PERG Signals in Ocular Diagnostics

PERG signals offer a non-invasive and objective measure of retinal function. Different ocular conditions, such as macular dystrophy, retinitis pigmentosa, and optic neuropathies, exhibit distinct PERG signal patterns. Understanding these patterns is crucial for early diagnosis and monitoring of disease progression. FDA allows for the detailed analysis of these signals, capturing the underlying functional relationships and variability across different conditions.

1.2 Functional Data Analysis (FDA) in Electrophysiology

Functional data analysis is a set of statistical techniques used to analyze data that can be considered as functions. In the context of PERG signals, FDA can handle the time-series nature of the data, allowing for the extraction of meaningful features and the modeling of temporal dynamics. By applying FDA to PERG data, we aim to enhance the accuracy of diagnostic models, identify characteristic signal patterns for various diseases, and provide a comprehensive tool for ophthalmic research.

2. Methodology

This study employs functional data analysis (FDA) to investigate the relationship between pattern electroretinogram (PERG) signals in various ocular diseases with their Visual acuity. Our methodology includes several critical steps designed to tackle the complexities of FDA. We are using the "fda" R library, created by James Ramsay et al. to facilitate our analysis.

2.1 Preprocessing Data

Before analysis, preprocessing of the PERG data is essential to ensure its quality and consistency. This involves multiple steps such as data cleaning to eliminate outliers and errors, noise reduction to enhance signal clarity, and temporal alignment to maintain consistency across observations.

To perform the analysis with the "fda" library in R, the raw PERG data must be converted into functional data objects. The preprocessing steps are as follows:

1. **Data Cleaning:** Identify and remove erroneous data points that could distort the analysis, including missing values.
2. **Noise Reduction:** Apply filtering methods to decrease noise in the PERG signals, preserving the accuracy of the data.
3. **Temporal Alignment:** Align all PERG signals correctly in time to enable accurate comparison across different recordings.

The preprocessing data is saved in CSV format for further manipulation. The CSV file is then imported into R Studio and transformed into a format recognizable by the "fda" library using the `Data2fd` function. This conversion step is crucial as it underpins all subsequent analyses.

2-2. Basis Function System

In functional data analysis (FDA), basis function systems are crucial for constructing functions that represent continuous data curves. James Ramsay's 2005 work extensively reviewed these systems, highlighting their importance and application. Basis functions act as the building blocks for representing data curves. Let ϕ_k denote the basis functions with $k=1, \dots, K$. Then, a function $x(t)$ can be expressed as:

$$x(t)=\sum c_k\phi_k(t)=c^T\phi(t)$$

This expression is known as the basis function expansion, where c represents the coefficients of the expansion. Two commonly used basis function systems in FDA are B-splines and Fourier basis functions.

B-splines: These are piecewise polynomial functions that provide local control over the curve shape.

Fourier Basis Functions: Derived from trigonometric sine and cosine functions, Fourier basis functions are ideal for analyzing periodic or cyclical data.

2-3. Smoothing

Smoothing techniques are applied to functional data to reduce noise and variability while preserving important features. This step is essential for enhancing the signal-to-noise ratio and improving data interpretability. Smoothing involves creating an approximate function that captures key patterns and filters out noise, resulting in a more uniform signal. Smoothing serves two primary purposes: extracting relevant information from data where smoothness is a reasonable assumption and providing analyses that are both flexible and robust.

2-4. Functional Regression

Functional regression models are powerful tools for examining the relationship between functional predictors and scalar response variables. These models expand traditional regression frameworks to accommodate functional data, enabling the estimation of functional coefficients that depict the influence of predictors on the response variable. Functional regression techniques consider the intrinsic variability and uncertainty in the functional predictors, yielding robust relationship estimates between variables.

A fundamental type of functional regression is the scalar-on-function regression, where the response variable y_i is scalar, and the predictor $x_i(t)$ is functional. The regression model can be expressed as: $yi = \alpha + \int x_i\,(t)\beta(t)dt + \epsilon_i$

In this context, y_i represents the average visual acuity of diagnosis$i$, and $x_i(t)$ is the smoothed functional form of Signal changes for country i.

2-6. Data Collection and Preprocessing

The Pattern Electroretinography (PERG) dataset was meticulously collected using the Optoelectronic Stimulator Vision Monitor MonPack 120, following ISCEV guidelines. Each participant had at least two signals recorded—one for each eye—while viewing a reversing checkerboard pattern.

The signals were captured at a high sampling rate of 1700 Hz over 150 milliseconds, with 255 observations per signal, and were averaged to enhance the signal-to-noise ratio.

The dataset includes 336 records, each with unique identifiers. Each CSV file contains a variable number of PERG signals, ranging from 2 to 10, measured in microvolts. Each record includes at least one PERG signal for each eye, labeled as RE_1 for the right eye and LE_1 for the left eye.

The `participants_info.csv` file provides clinical and demographic information, including unique record identifiers, dates, ages, sexes, up to three diagnoses, visual acuity for both eyes, and additional comments. This dataset offers a robust foundation for studying PERG signals in ocular electrophysiology research.

va_re: **Visual acuity** for the right eye, measured on the logMAR scale. Values above 0 indicate decreased visual acuity, while negative values indicate better-than-normal acuity. Missing values are marked as "NA".

va_le: Visual acuity for the left eye, measured similarly to the right eye.

diagnosis1: different diagnoses or medical conditions the participant may have.

For our study, we covers 24 type of diagnosis out of 54 and includes information from 24 participant, one participant for each type of diagnosis.

3. Results

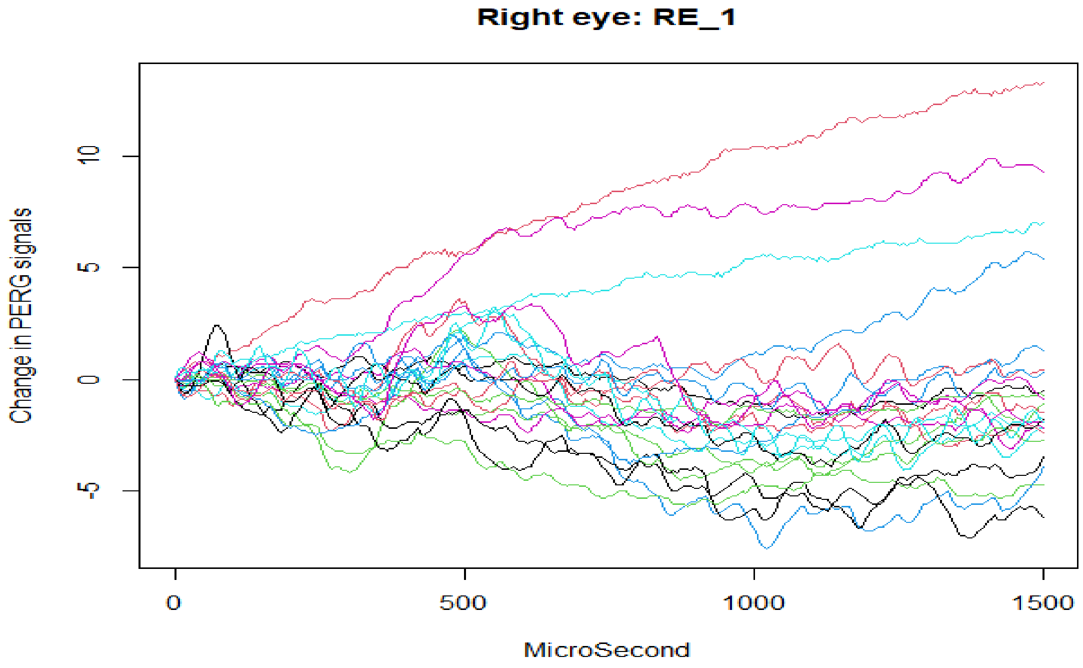
The results of our comprehensive analysis offer valuable insights into the dynamic relationship between PERG signals and visual acuity for both Right and Left eyes separately. Then we combined smoothed PERG signals for both Right and Left eyes and found relationship with visual acuity of Right eye and Left eye respectively. Finally, we combined smoothed PERG signals for both Right and Left eyes and smoothed value of visual acuity for both Right and Left eyes then found relationship with diagnosis type as a categorical variable by using Multinomial model (multinom) from “nnet” library. Diagnosis which analyzed are:

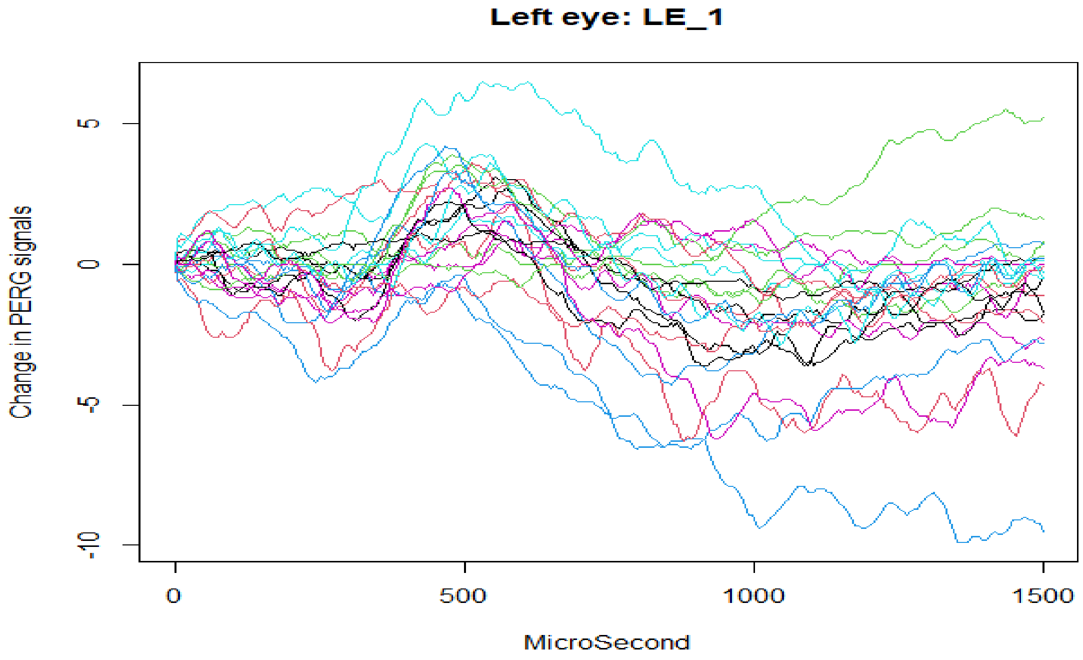
Diagnosis	
Paracentral acute middle maculopathy type 2	Bilateral optic nerve atrophy
Autoimmune retinopathy	Foveal hypoplasia
Inherited optic atrophy	Blue cone monochromatism
Chorioretinopathy Birdshot type	AlbinoidismA
Infectious neuritis	Orbital ischemia
Arterio-venous malformation in right thalamus	Optic neuropathy
Traumatic optic neuropathy	Sarcoidosis neuropathy
Cone-Rod dystrophy	Normal
Albinism	Central nervous system disorder
Bilateral Optic Atrophy	External ophthalmoplegia
Congenital Achromatopsia	Macular dystrophy
Anisometropic amblyopia	Acute macular neuroretinopathy

We can visualize the original trends of the PERG Signal change across the chosen Diagnosis types by plotting them during period 0 to 1499 microsecond.

Each colored line represents the PERG signals change trend for a specific diagnosis during period 0 to 1499 microsecond. Each signal captured for a duration of 150 milliseconds and 255 equally spaced observations per signal.

Pert of time interval: [0000   0006   0012   0018   0024   0030   . . . 1487 1493 1499]

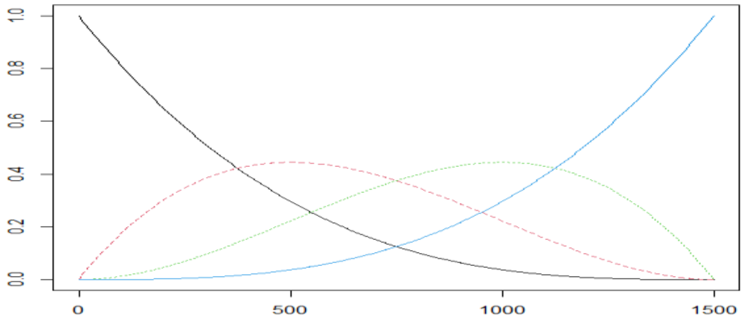




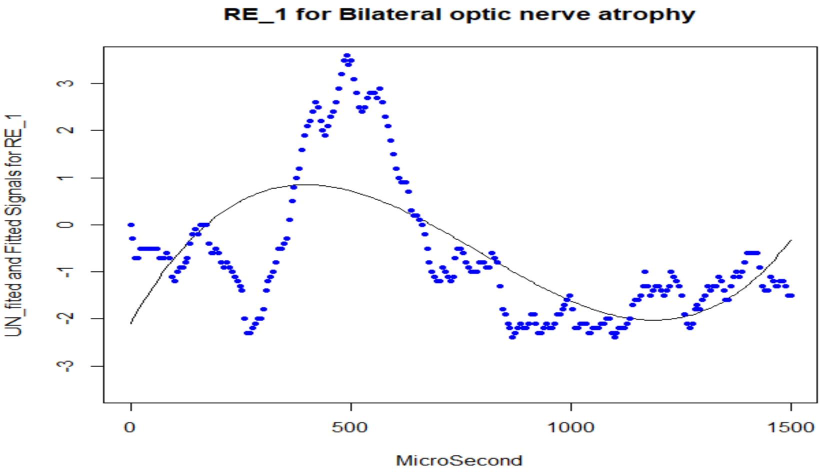
3-1. Smoothing Functional Data

Smoothing functional data involves transforming discrete and potentially noisy data points into continuous and smooth functions. This process is crucial for accurately representing and analyzing trends in the data. The key to effective smoothing lies in using roughness penalties and regularization techniques, which offer a powerful and flexible approach to achieve the desired smoothness. which can be optimized using criteria like GCV. This method ensures that the resulting functions are not only smooth but also meaningful representations of the underlying data trends.

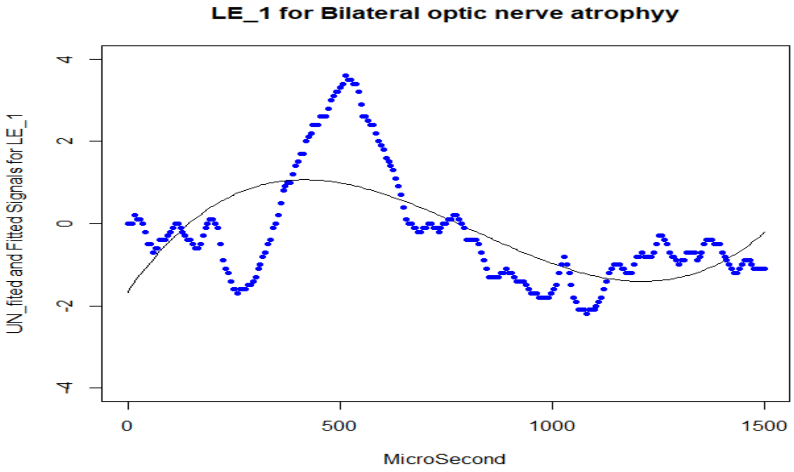
To smooth out sharp lines and better illustrate trends, we employ two primary methods: B-spline basis and Fourier basis. For this project, we utilized a B-spline smoothing curve of order 4 for the diagnosis type indexed as 2 .



Plotting "fitted data" and "actual data points" in one pane for Right eye, RE_1, for diagnosis type 'Bilateral optic nerve atrophy'. Black line shows the B-spline smoothing curve of order 4.

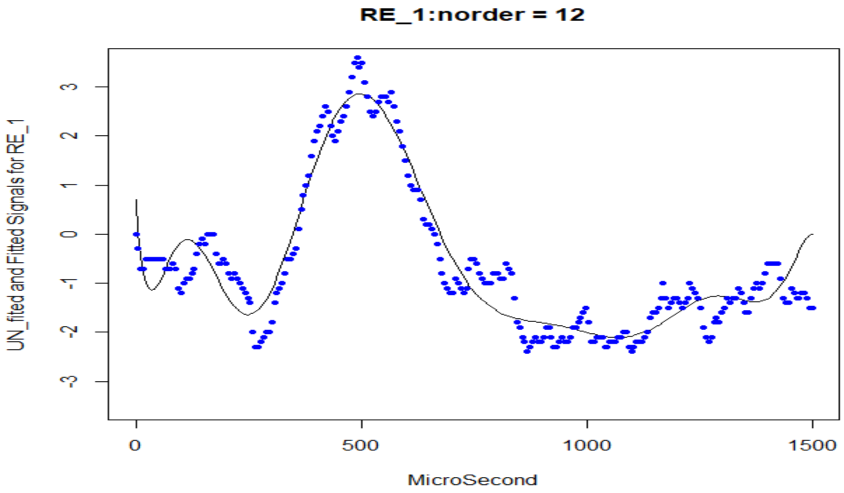


Plotting “fitted data” and “actual data points” in one pane for Left eye, LE_1, for diagnosis type 'Bilateral optic nerve atrophy’. Black line shows the B-spline smoothing curve of order 4.

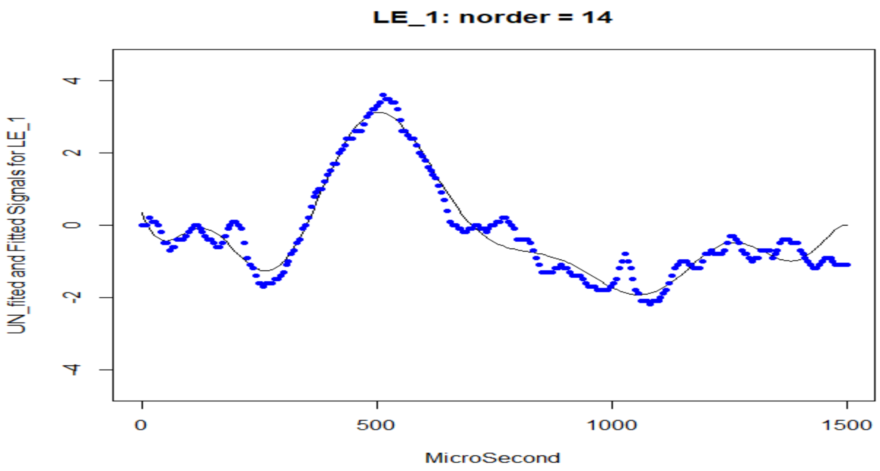


Smoothing functional data with “order”:

This section will present you with the impact of the number of orders on the fitting curve. We first defined the range of norder values to find the best oredr number. For doing this, we create a for loop to iterate over each norder value for diagnosis_levels = [2] ('Bilateral optic nerve atrophy’) and both RE_1 and LE_1 separately.



Result: norder from 3 to 12 analysed, and result shows that norder with 12 has a good result for smoothing signals of Right eye, RE_1.



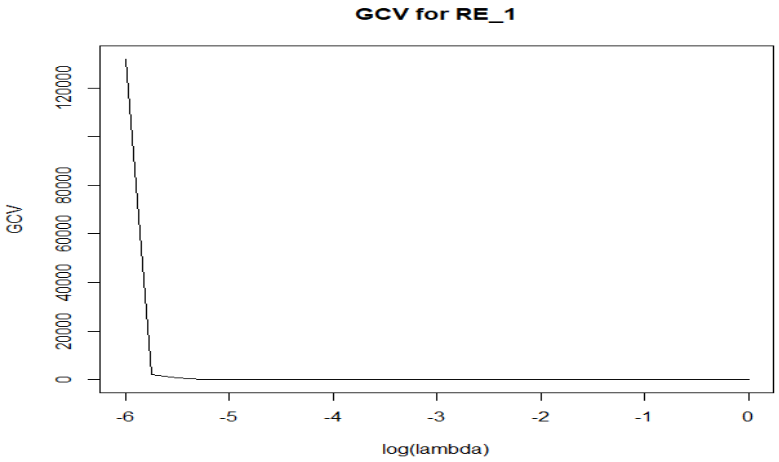
Result: norder from 3 to 14 analysed, and result shows that norder with 14 has a good result for smoothing signals of Left eye, LE_1.

B-spline with lambda penalty (λ):

We create a B-spline basis with lambda penalty term. But we need to find optimal λ penalty terms. Therefore, we used GCV (Generalized Cross Validation).

GCV (Generalized Cross Validation):

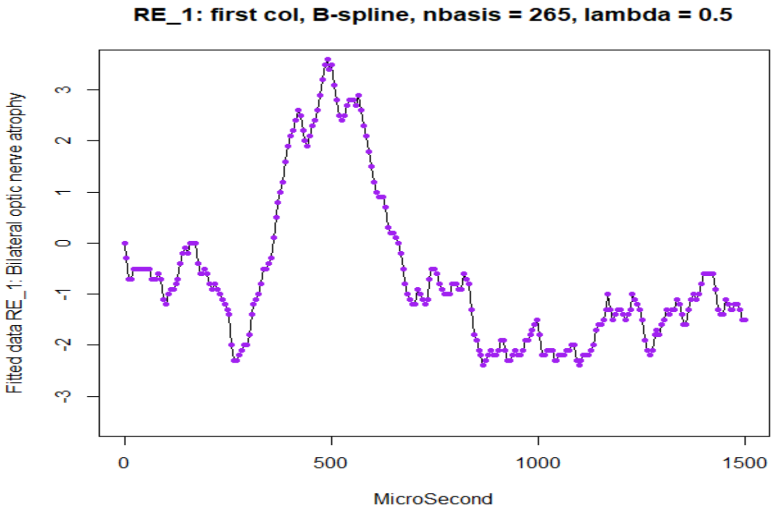
A method to find the optimal smoothing parameter by minimizing the prediction error. Let's see the result of GCV to see which values of lambda are optimal (or suboptimal).



The best value for lambda for RE_1 seems to be in between 0.0001 and 1.

Creating a B-spline basis using a derivative of order 4 and compute the smoothed fitting to the data, using penalty term = 0.5 for Signals of Right eye (RE_1).

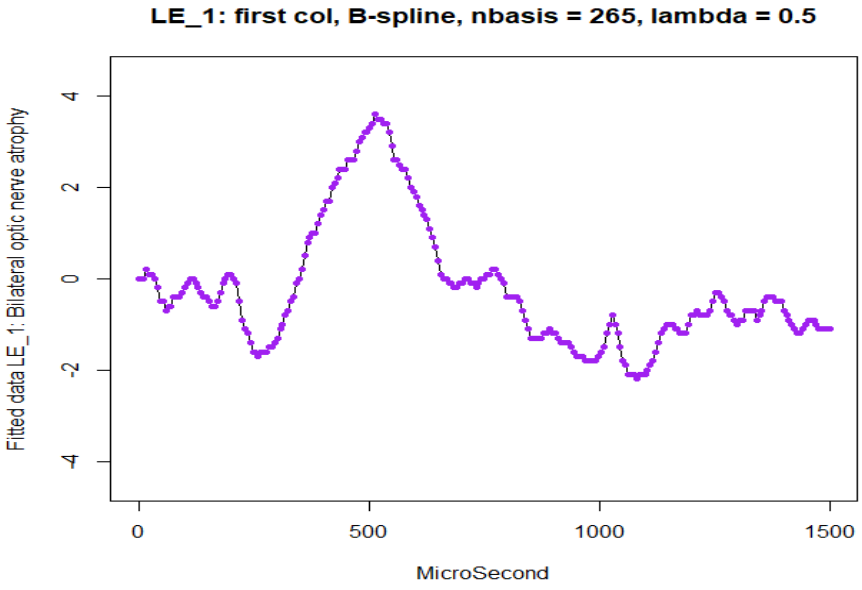
We used the norder = 12 and max number of basis which is nbasis = 255 + norder - 2 .



GCV (Generalized Cross Validation) for LE_1:

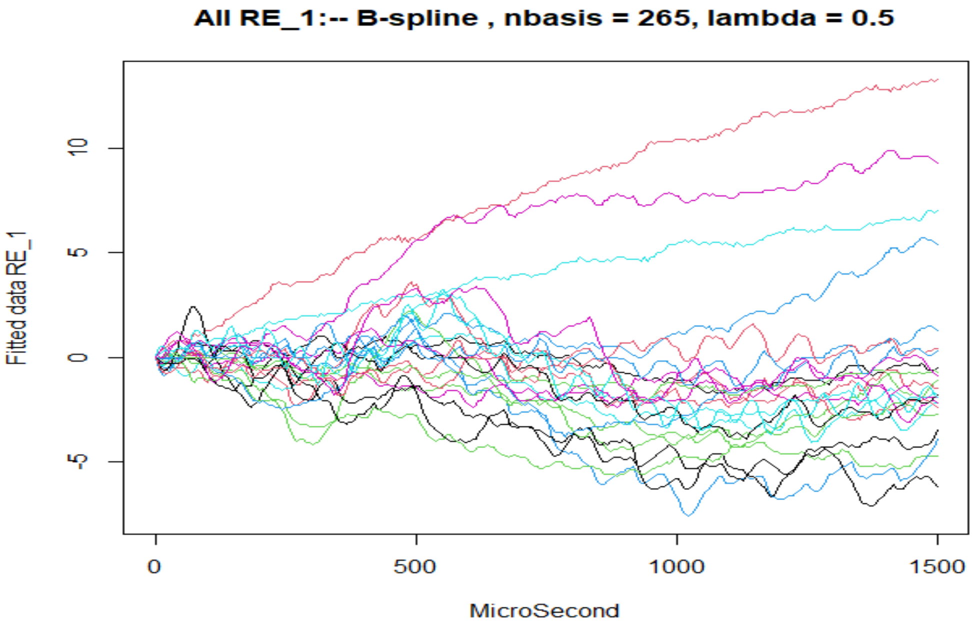
The best value for lambda for RE_1 seems to be in between 0.0001 and 1.

Creating a B-spline basis using a derivative of order 4 and compute the smoothed fitting to the data, using penalty term = 0.5 for Signals of Left eye (LE_1).

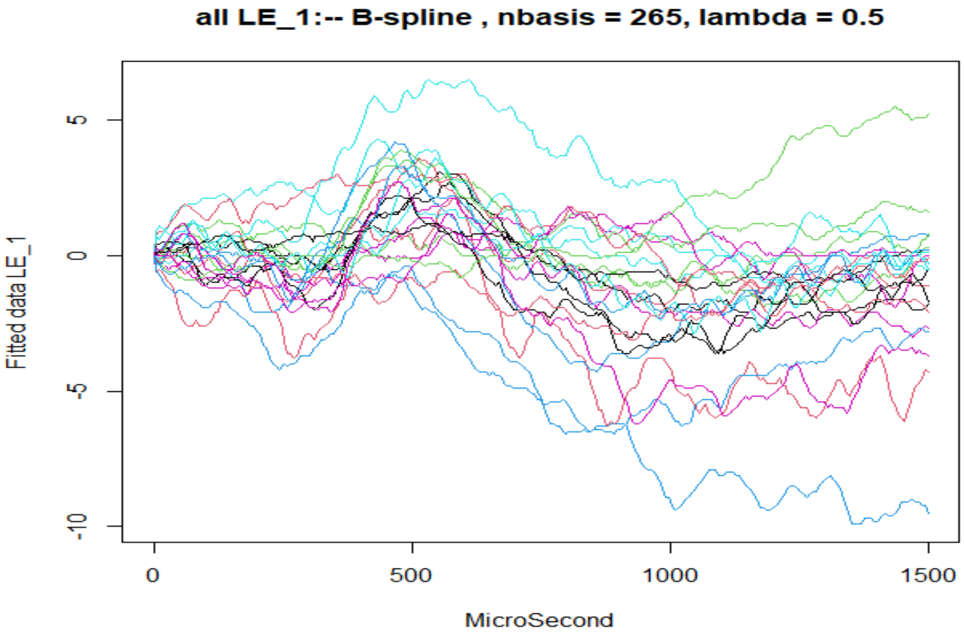


All RE_1 and LE_1 Data:

Applying B-spline basis with lambda penalty term, lambda = 0.5, and norder = 12 for all RE_1 Data:



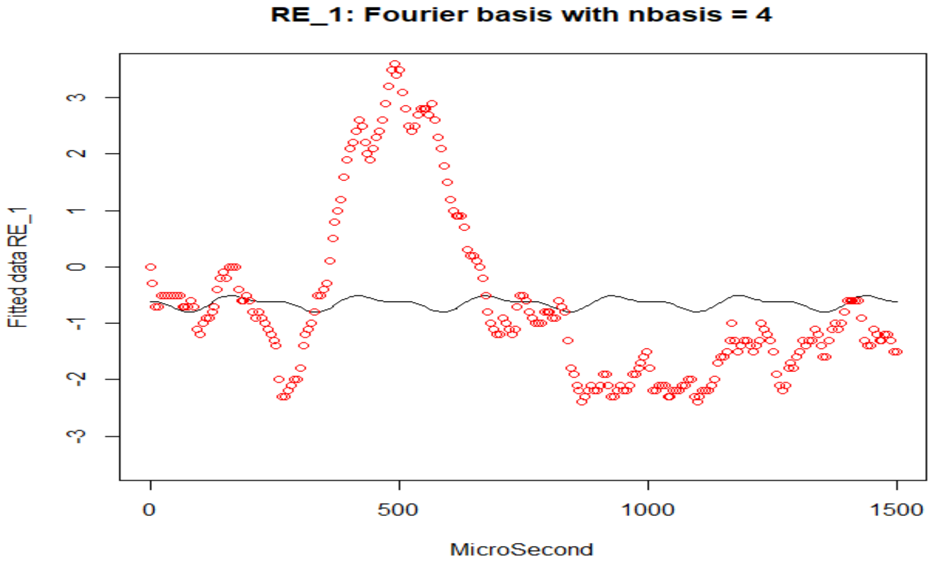
B-spline basis with lambda penalty term, lambda = 0.5, and norder = 12 for all **LE_1**:

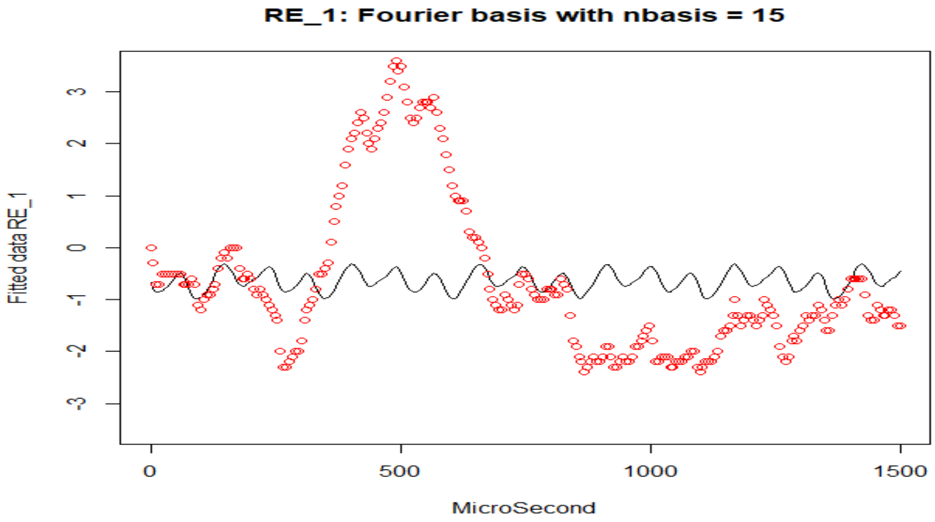


Fourier basis function:

Also, with the Fourier basis function we can do the same with the number of basis of 4 and 15.

Plotting “fitted data” and “actual data points” in one pane for Right eye, RE_1, for diagnosis type 'Bilateral optic nerve atrophy’. Black line shows the Fourier smoothing curve of basis 4 and 15.





Result: increasing the number of basis from 4 to 15 does not affect the result, and does not give a good smoothing curve.

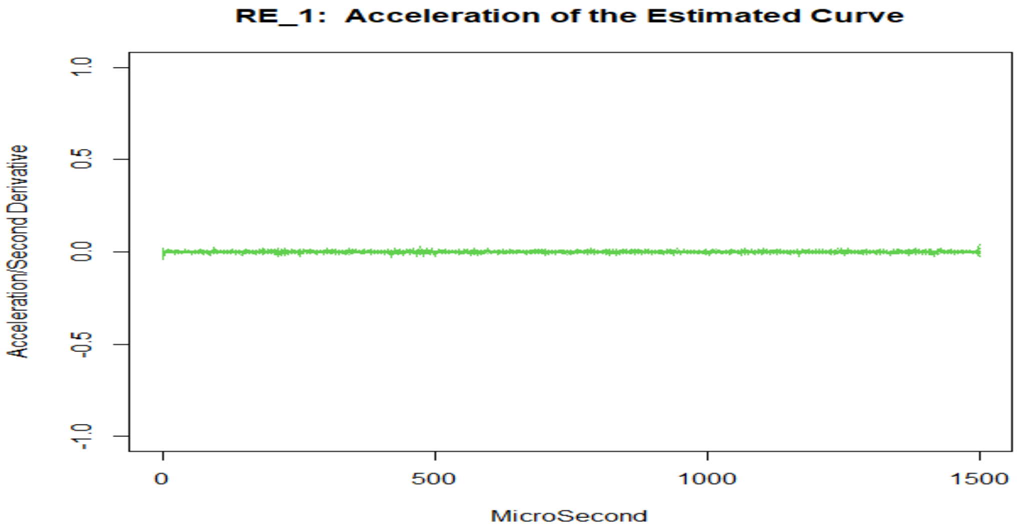
3-2. Feature (Landmark) Registration

In FDA (Functional Data Analysis), landmark registration is a technique used to align functional data by identifying key features or landmarks within each curve and then registering them based on these landmarks.

To assess the acceleration of the PERG signals, we computed the second derivative of the smoothed data for both the left eye (LE_1) and right eye (RE_1) signals. This analysis helps us understand the changes in the signals' curvature over time.

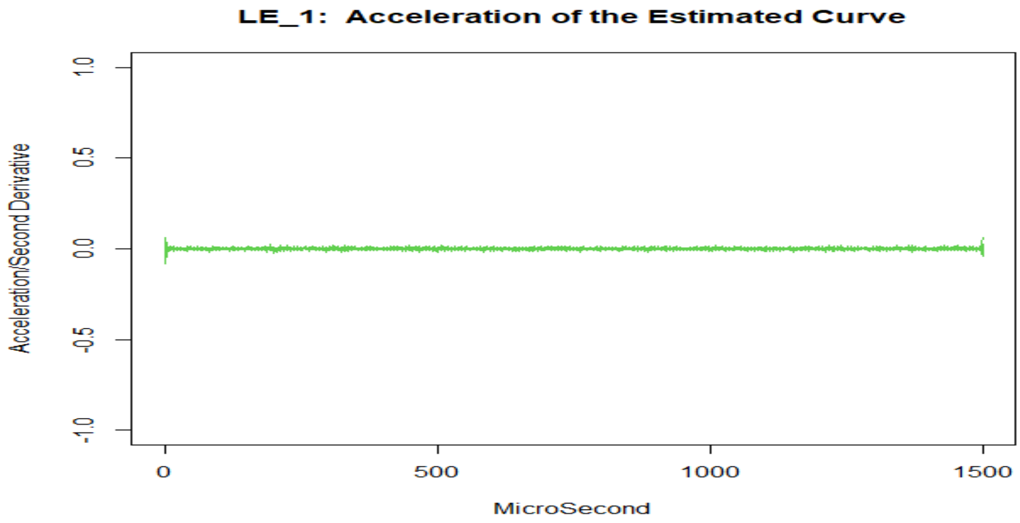
Acceleration of RE_1 Signals

The second derivative (acceleration) of the RE_1 signals was calculated and plotted. The green lines represent individual functional data points, while the black line represents the mean acceleration.



Acceleration of LE_1 Signals

Similarly, the second derivative of the LE_1 signals was calculated and plotted. Again, the green lines denote individual functional data points, and the black line indicates the mean acceleration.



The plots for both RE_1 and LE_1 signal showed no significant peaks or valleys. This suggests that there are no distinct landmarks or abrupt changes in the acceleration curves, indicating that landmark registration is not necessary for these datasets. The smooth nature of the curves confirms the consistent quality of the recorded signals.

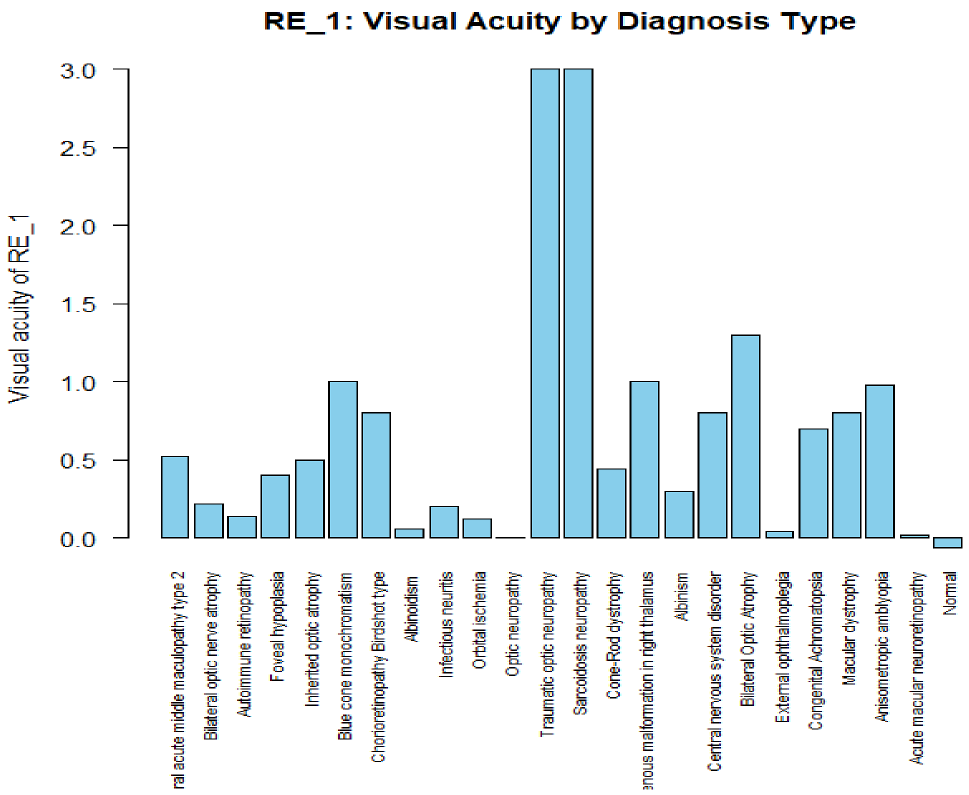
3-3 Functional Regression Analysis

Functional regression entails modeling the relationship between a response variable (y) and one or more predictor variables (x) represented as functional data. In this context, our response variable (y) is the Visual acuity of left and right eyes (va_re_logMar & va_le_logMar) and our predictor variables (x) is different based on models.

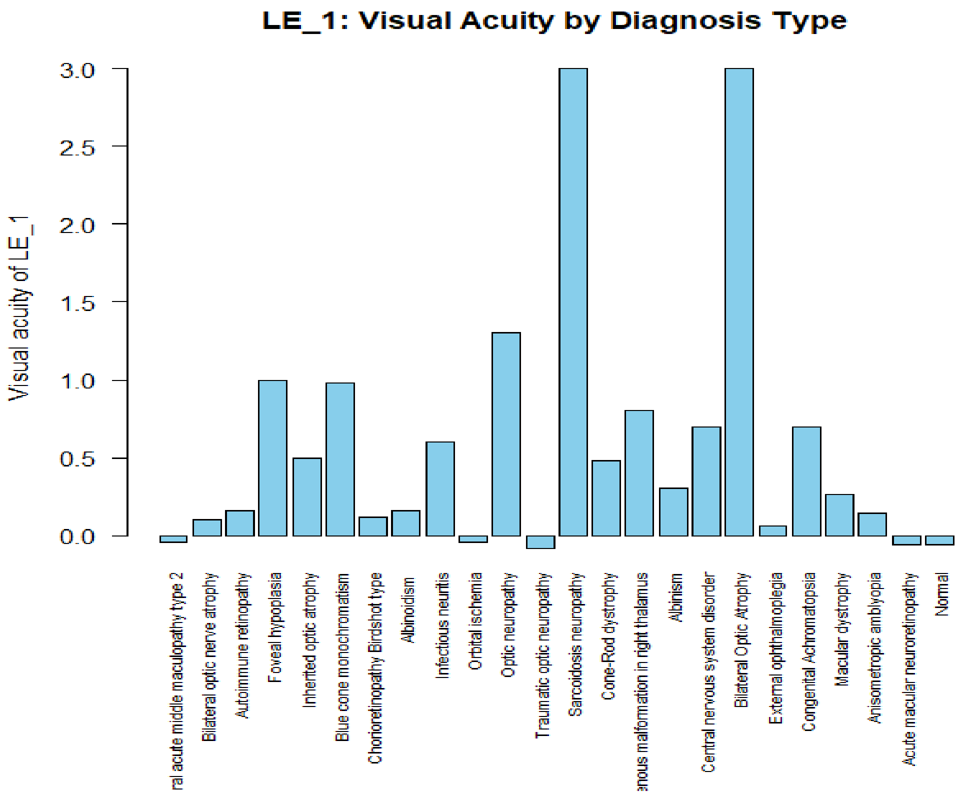
3.3.1. Functional Regression Analysis between RE_1 and va_re_logMar

“va_re”: "Visual acuity" for the right eye, measured on logMar (logarithm of the minimum angle of resolution) scale. A logMAR value of 0 denotes "normal" vision, while values above 0 indicate a decrease in visual acuity. Conversely, negative logMAR values indicate better-than-normal visual acuity.

“va_re”: "Visual acuity" for the Right eye:



RE_1: Visual Acuity by Diagnosis Type



LE_1: Visual Acuity by Diagnosis Type

The above bar charts display '**visual acuity' values** for different diagnosis types for the right eye (RE_1) and left eye (LE_1), measured on the logMAR scale. In the logMAR scale, a value of 0 indicates normal vision, positive values indicate decreased visual acuity, and negative values suggest better-than-normal visual acuity.

The bar charts show that for several diagnosis types, such as **'traumatic optic neuropathy'** for Right eye and **'Bilateral Optic Atrophy'** of Left eye, the logMAR values are high (around 3.0), indicating a significant **loss** of visual acuity. However, the logMAR value for left eye in **'traumatic optic neuropathy'** case is **negative** showing a good visual acuity.

For **'Sarcoidosis neuropathy'** the logMAR values are high around 3.0 for both eyes shows significant low visual acuity values.

The result of above plots shows that just **'normal'** diagnosis case has negative value for both eyes, which means va_re_logMAR values indicate a good visual acuity.

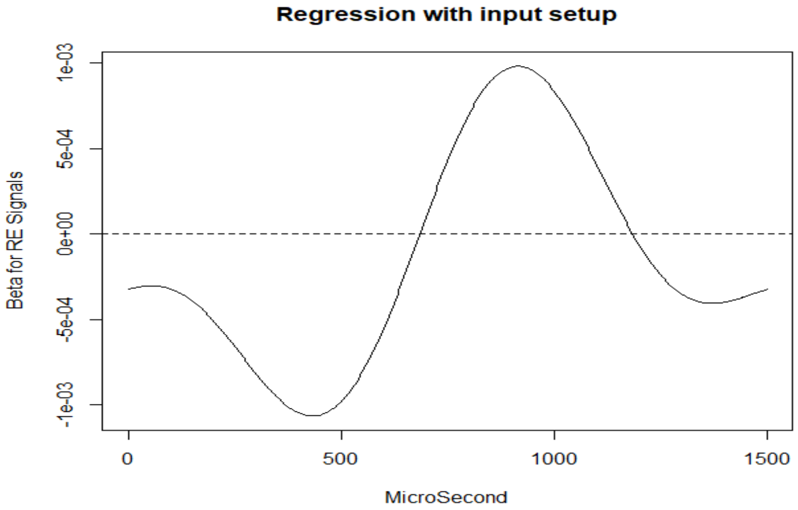
For other conditions like '**Autoimmune retinopathy**', the logMAR values are relatively lower, suggesting less severe visual problem. Additionally, the visual acuity values for many diagnoses are higher in the left eye (LE_1) compared to the right eye (RE_1), indicating that the right eye is often more affected by these conditions. These charts are useful for clinicians to understand the specific visual acuity impairments associated with various diagnoses and to plan appropriate interventions for each eye.

Relationship between RE_1 and va_re_logMar:

In this analysis we used 'va_re_logMar' as a 'response variable', and the "RE_1" signal as a "covariate'.

The following plot shows the result for Beta for "RE_1" signal:

In functional regression, the beta values are the regression coefficients that quantify the relationship between the functional predictor(s) and the response variable. Each beta value corresponds to a specific functional predictor, indicating both the magnitude and direction of its effect on the response variable. These beta values are estimated through regression analysis, providing valuable insights into how variations in the functional predictors influence the response variable.

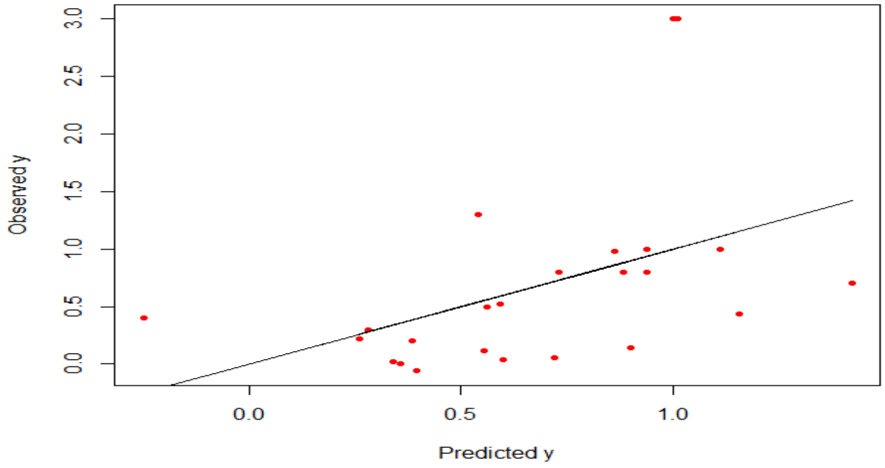


The y-axis is labeled "Beta for RE Signals," showing the regression coefficients (beta values) for the "RE_1" signal. The plot shows a curve that represents the beta values over time (in microseconds). The horizontal dashed line at y = 0 serves as a reference line, indicating where the beta value is zero.

- The curve starts below zero, indicating a negative beta value at the initial time points. This suggests that at these time points, the "RE_1" signal has a negative relationship with the response variable.
- As time progresses, the curve rises above zero, reaching a positive peak. This indicates that at this time, the "RE_1" signal has a strong positive influence on the response variable.
- After reaching the peak, the curve declines back towards zero, then drops below zero again, and finally rises slightly towards the end. This shows that the influence of the "RE_1" signal on the response variable changes over time, switching between negative and positive effects.
- The magnitude and sign of the beta values provide insights into how changes in the "RE_1" signal affect the response variable over time.

This plot is crucial for understanding the temporal dynamics of the relationship between the "RE_1" signal and the response variable, allowing researchers to identify periods where the signal has the most significant positive or negative impact.

The following scatter plot related to the evaluation of a functional regression model. shows the result for observed y-values and predicted y-values separately for "RE_1" signal:



The x-axis is labeled "Predicted y," which represents the values predicted by the functional regression model. The y-axis is labeled "Observed y," which represents the actual observed values of the response variable.

- The plot consists of red dots, each representing a pair of predicted and observed values.
- A solid black line is drawn through the points, representing the line of equality where predicted values equal observed values.

Squared Multiple Correlation (RSQ1): 0.2036262

This value indicates the proportion of variance in the observed data that is explained by the predicted data from the model. In this case, about 20.36% of the variance in the observed values is explained by the predicted values. The RSQ1 suggests that the model explains a moderate amount of the variance in the observed data

F-ratio (Fratio1): 0.9204902

The F-ratio is used to test the overall significance of the regression model. It compares the model with and without the predictors to see if the predictors provide a better fit to the data than a model with no predictors.

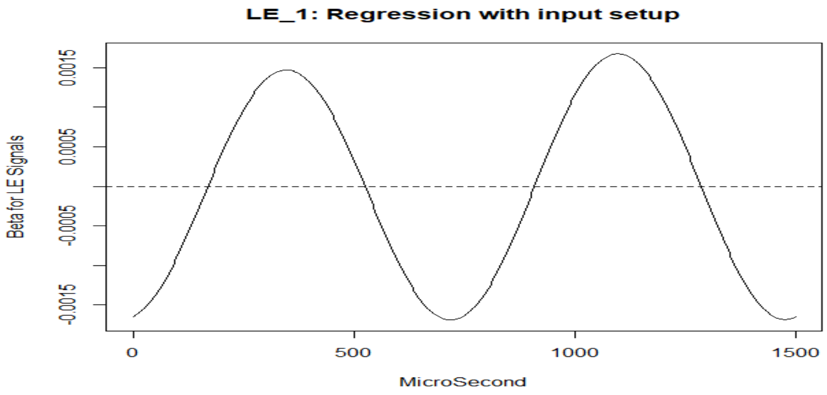
P-value: 0.4902265

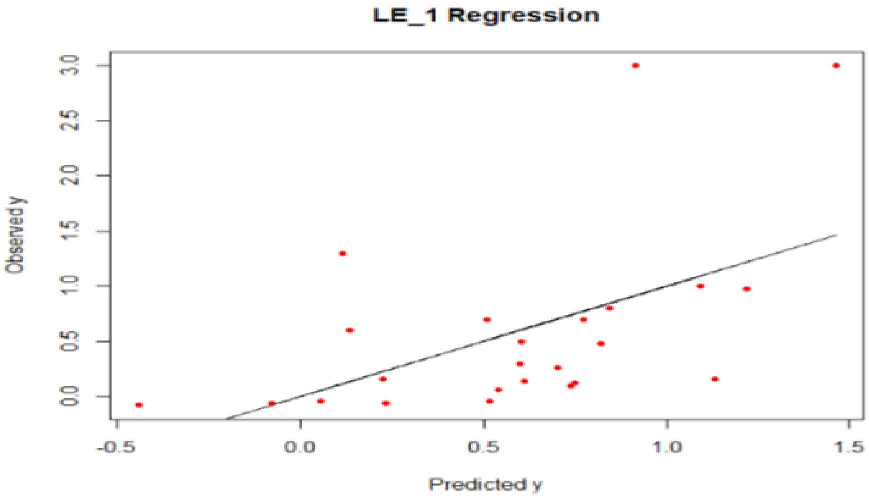
The p-value is used to test the null hypothesis (H0) that all regression coefficients are identically 0, meaning the predictors have no effect on the response variable.
In this case, the p-value is approximately 0.4902, which is greater than the common significance level of 0.05. This suggests that there is not enough evidence to reject the null hypothesis, indicating that the model's predictors may not have a significant impact on the response variable.

The F-ratio of 0.9204902 and the corresponding p-value of 0.4902265 indicate that the regression model is not statistically significant, meaning the predictors do not significantly improve the model's fit to the data. Therefore, the functional regression model has limited explanatory power and is not statistically significant in explaining the variability in the observed data.

3.3.2. Functional Regression Analysis between RE_1 and va_re_logMar

We will use 'va_le_logMar' as a 'response variable', and the "LE_1" signal as a "covariate'.





Assessing the "quality of fit" for LE_1 and va_le_logMar:

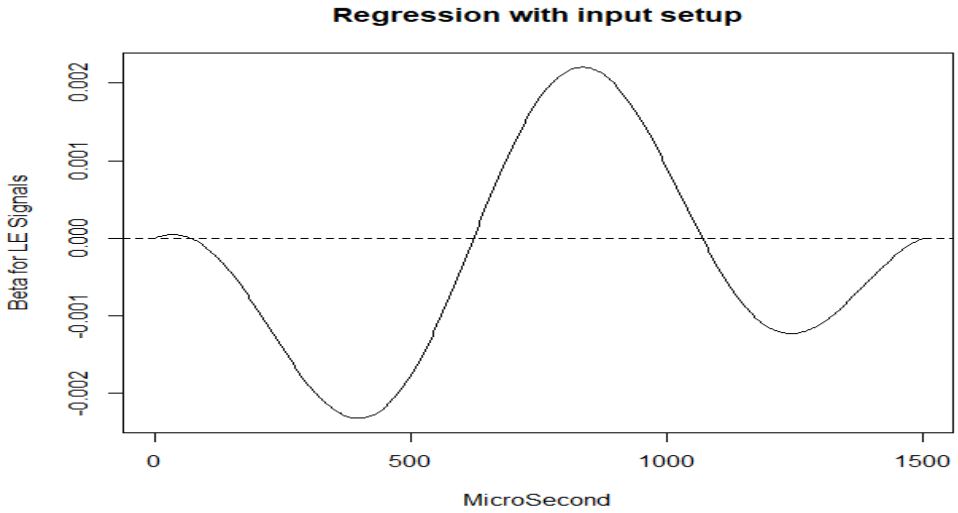
RSQ1 = 0.2804152 , Fratio1 = 1.402885 , P-value: 0.2702047.

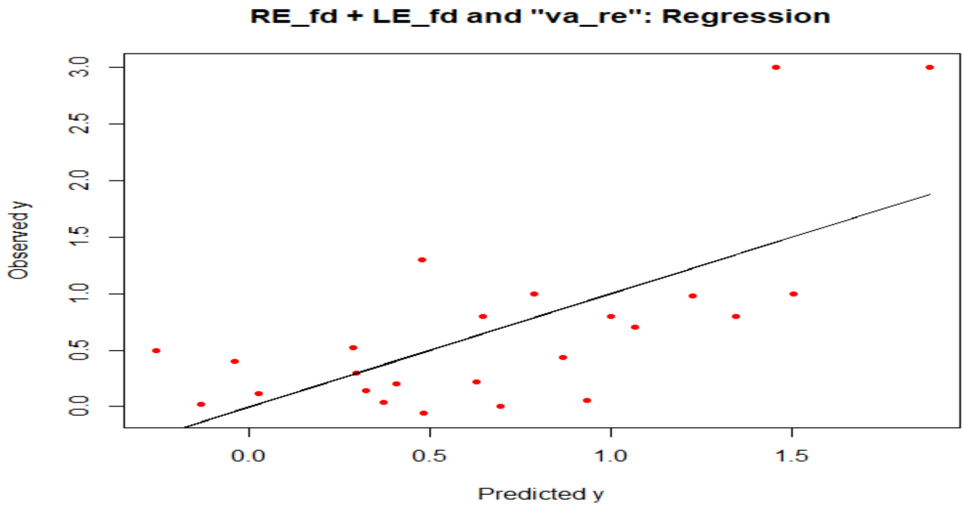
H0: all coefficients are identically 0. This suggests that there is not enough evidence to reject the null hypothesis, indicating that the model's predictors may not have a significant impact on the response variable.

3.3.3. Functional Regression Analysis between RE_1 + LE_1 and va_re_logMar

Multiple functional regression:

model_func_regress_visual <- fRegress(va_re_logMar ~ RE_fd + LE_fd)





Assessing the "quality of fit" for LE_1 + RE_1 and va_re_logMar:

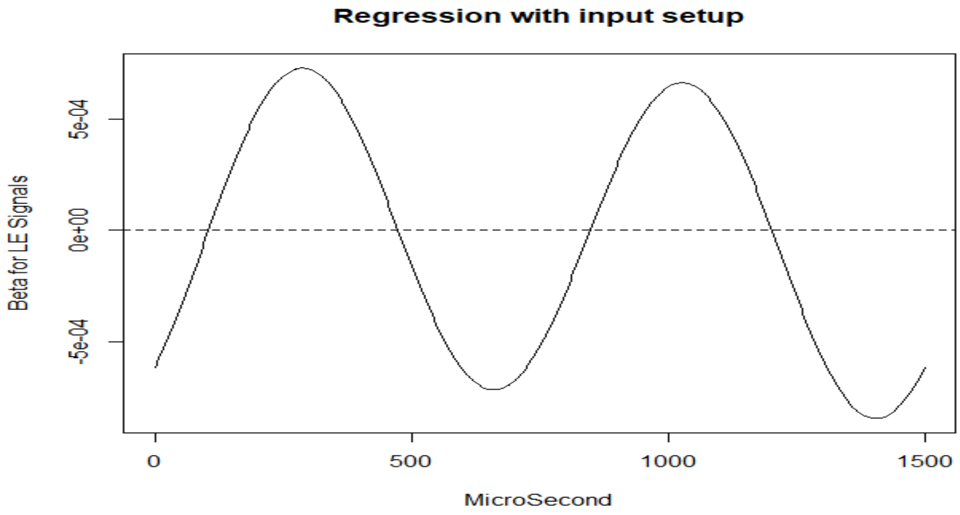
We can now compute the squared multiple "correlation" , the F-ratio, and p-value:

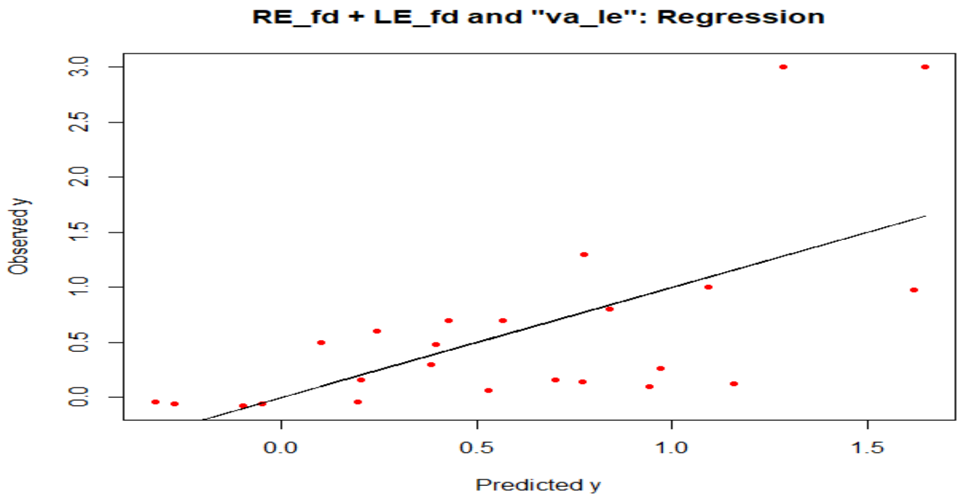
RSQ1 = 0.4639109 , Fratio1 = 3.115301, P-value: 0.03

H0: all coefficients are identically 0. This suggests that there is enough evidence to reject the null hypothesis, indicating that the model's predictors (LE_1 + RE_1) have a significant impact on the response variable (va_re_logMar).

3.3.4. Functional Regression Analysis between RE_1 + LE_1 and va_le_logMar

model_func_regress_visual <- fRegress(va_le_logMar ~ RE_fd + LE_fd)





Assessing the "quality of fit" for LE_1 + RE_1 and va_le_logMar:

The provided scatter plot titled "RE_fd + LE_fd and 'va_le': Regression" displays the relationship between predicted and observed values for a regression model involving right eye (RE_1), left eye (LE_1) functional data, and the visual acuity of the left eye (va_le_logMar). The red dots represent individual data points, and the solid black line is the line of best fit.

Key statistics:

- RSQ1 (R-squared) = 0.4266645, indicating that approximately 42.67% of the variance in observed values is explained by the model.
- Fratio1 = 2.679046, with a P-value of 0.055, suggesting marginal significance.

Interpretation: The P-value (0.055) is slightly above the common significance level of 0.05, providing borderline evidence to reject the null hypothesis (H0) that all coefficients are zero. This implies the predictors (LE_1 + RE_1) have a marginally significant impact on the response variable (va_le_logMar).

3.3.5. Functional Regression Analysis with Categorical response variable

We smooth the functional covariates for all predictor variables (x) with B-spline basis with 10 basis. The predictor variables (x) in this case are all RE_1, LE_1, va_re_logMar, and va_le_logMar and our response variable is diagnosis type (24 diagnosis types).

We perform functional PCA on each functional covariate in order to avoid high dimention errors. Therefore, we extract the scores (principal components) for each functional covariate (predictors) with nharm = 3, and combine the principal components of all predictors into a single dataframe. Finally, we Fit the **multinomial logistic regression** model. We used two fda.usc and nnet Library.

Response variable: diagnosis type (24 diagnosis types).

Predictors: RE_1 Signal, LE_1 Signal, va_re_logMar, and va_le_logMar

Combine the principal components into a single data frame:

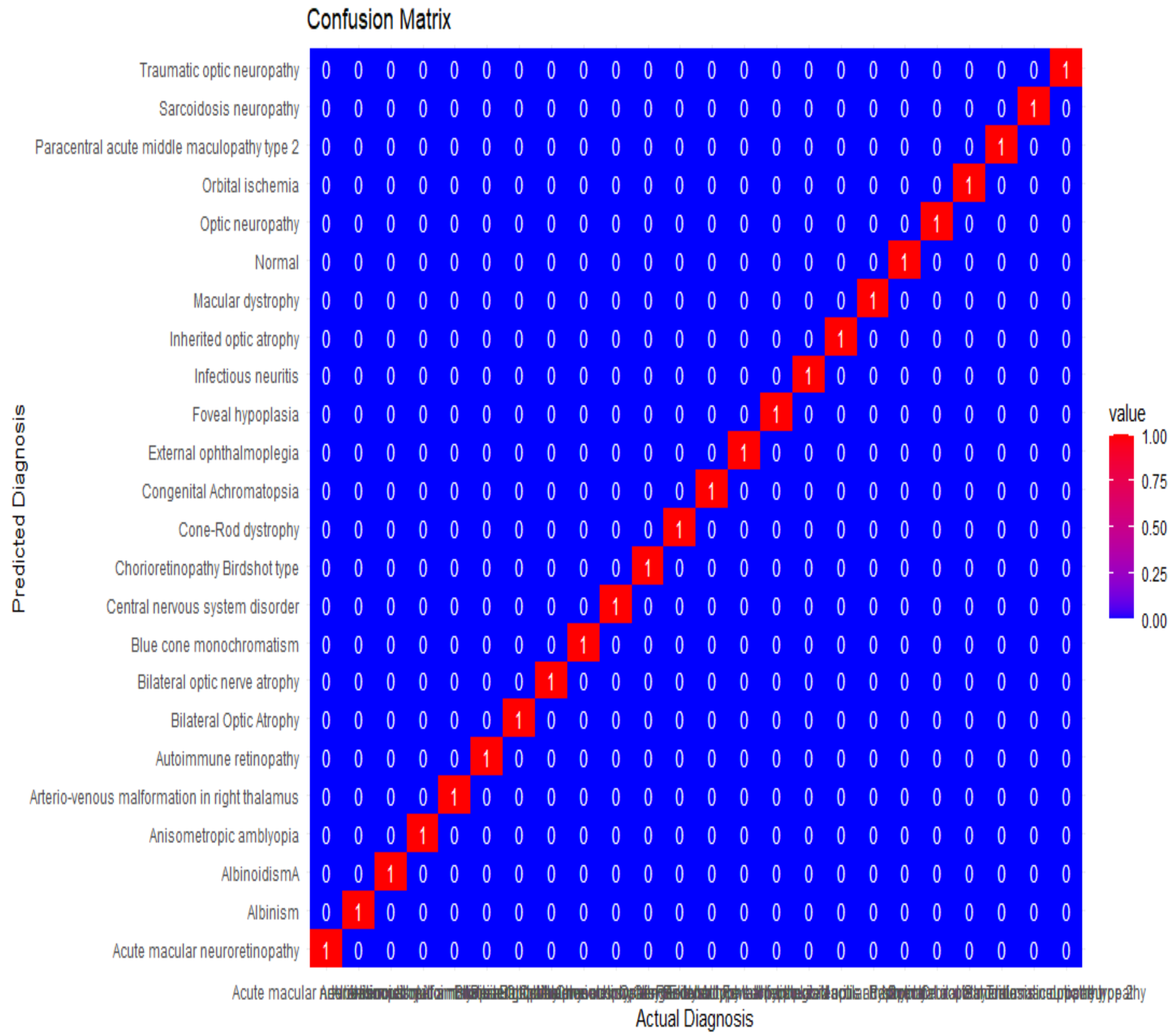
combined_data <- cbind(va_re_scores, va_le_scores, RE_scores, LE_scores)

Fit the multinomial logistic regression model:

multinom_model <- multinom(diagnosis_levels ~ . , data = combined_data)

Result:

The value of Accuracy is equal to 1.



The image shows a **confusion matrix** for a diagnostic classification model. Each cell represents the number of instances where the predicted diagnosis (rows) matches the actual diagnosis (columns). The diagonal cells, highlighted in red, indicate correct predictions, with a value of 1, showing perfect classification for those cases. The blue cells with a value of 0 indicate no instances of incorrect predictions for those particular diagnosis pairs. This matrix effectively illustrates the model's accuracy in diagnosing various conditions, with all diagnoses correctly classified in this instance.

TEST result:

After **multinomial logistic regression** model trained, we used 24 new individual data to test the trained model. But the result was not good. Because of two reasons:

1. **Overfitting**: The model may be overfitted to the training data, capturing noise and specifics that do not generalize to new data.

2. **Small Sample data**: since the number of observations for each diagnosis was limited (one per each), so model is not trained well, therefore we need big dataset to train well fitted model.
3. **Data Imbalance or Issues**: There might be imbalances or issues in the data that affect model performance differently on training and test sets.

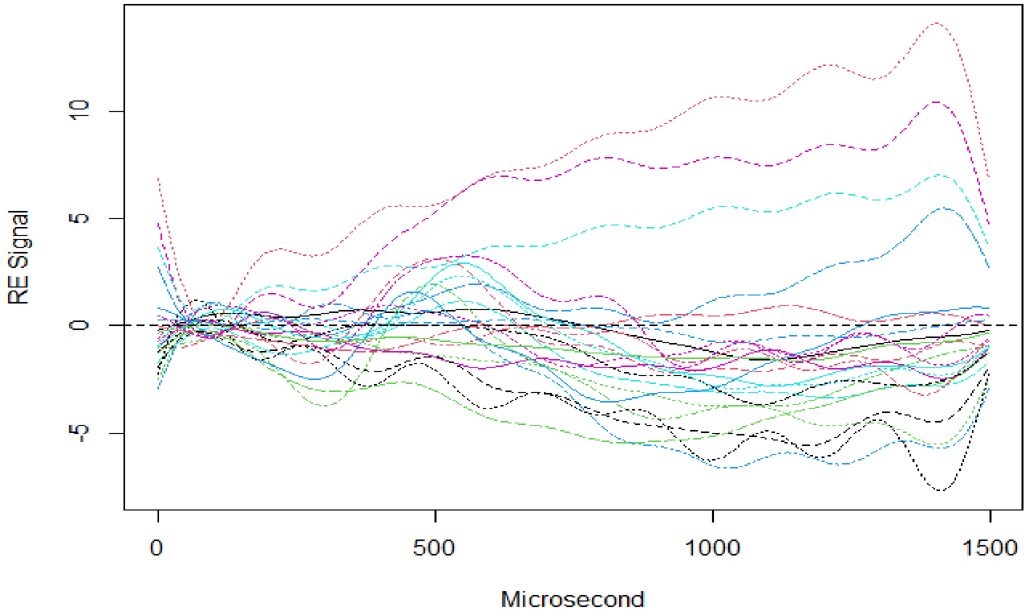
The following table shows the difference between the diagnosis type in both actual and predicted.

Index	Diagnosis_TEST	Diagnosis_Prediction	
1	Normal	Infectious neuritis	-
2	Macular dystrophy	Traumatic optic neuropathy	-
3	Macular dystrophy	Sarcoidosis neuropathy	-
4	Congenital achromatopsia	Paracentral acute middle maculopathy type 2	-
5	Congenital achromatopsia	Bilateral optic nerve atrophy	-
6	Bilateral optic nerve atrophy	Paracentral acute middle maculopathy type 2	-
7	Autoimmune retinopathy	Normal	-
8	Albinism	Bilateral optic nerve atrophy	-
9	Orbital ischemia	Normal	-
10	Autoimmune retinopathy	Cone-Rod dystrophy	-
11	Normal	Acute macular neuroretinopathy	-
12	Normal	Bilateral optic nerve atrophy	-
13	Normal	Acute macular neuroretinopathy	-
14	Normal	Infectious neuritis	-
15	Macular dystrophy	Traumatic optic neuropathy	-
16	Macular dystrophy	Sarcoidosis neuropathy	-
17	Congenital achromatopsia	Paracentral acute middle maculopathy type 2	-
18	Congenital achromatopsia	Bilateral optic nerve atrophy	-
19	Bilateral optic nerve atrophy	Paracentral acute middle maculopathy type 2	-
20	Autoimmune retinopathy	Normal	-
21	Albinism	Bilateral optic nerve atrophy	-
22	Orbital ischemia	Normal	-
23	Autoimmune retinopathy	Cone-Rod dystrophy	-
24	Normal	Acute macular neuroretinopathy	-

3.4. Depth Measures

In Functional Data Analysis (FDA), depth analysis is a method used to measure centrality and outlyingness of functional data. Depth measures are used to identify the center of a functional dataset, similar to how the median is used in scalar data.

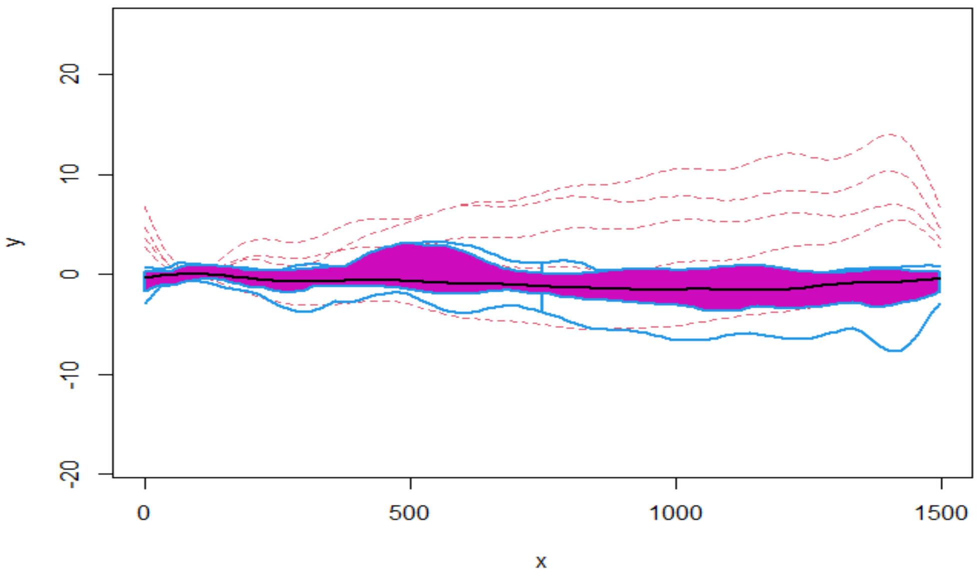
The following plot shows the smoothed fuctional data of RE signal.



Modified Band Depths (MBD):

Here we generate the boxplot using Modified Band Depths (MBD). Modified Band Depth (MBD) is a depth measure that determines the centrality of a functional curve within a group of curves. It improves upon the Band Depth (BD) measure, specifically designed to manage irregularities in functional data. MBD assesses the proportion of curves that contain a particular curve within a band defined by the set's maximum and minimum values. Higher MBD values signify greater centrality, while lower values indicate that the curve is more of an outlier.

The following plot is related to Modified Band Depths (MBD). The central black line represents the median curve, while the shaded purple area indicates the central region of the data, enclosed by the maximum and minimum values of the set. The blue lines represent other individual curves in the dataset, and the dashed red lines signify outlier curves. Higher MBD values are indicated by curves closer to the median, while lower values (or outliers) are further from the median, highlighted by the dashed red lines.



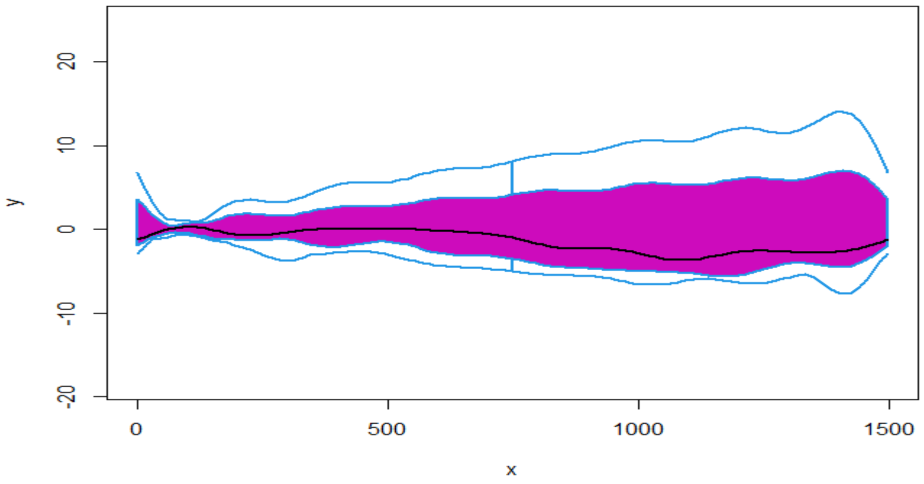
There are some outliers, the dashed red curves. Now let's re-compute the boxplot using the BD2 method, which uses two curves to compute the bands.

Band Depth (BD)

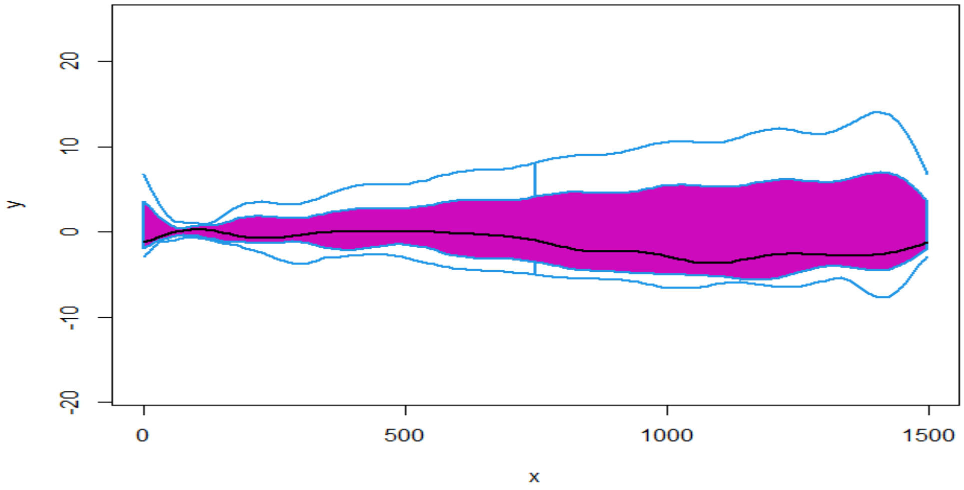
BD is a measure of depth that evaluates how central a functional curve is within a dataset. It measures the proportion of curves that enclose a particular curve within their convex hull. BD offers a reliable measure of centrality, especially useful for analyzing functional data with smooth, well-behaved curves. The shape remains consistent between both Mean Absolute Bending (MBD) and Bending (BD) figures. Furthermore, when a combined approach using both MBD and BD was attempted, the results were similar to those of the BD figure.

The following illustrates a plot related to Band Depth (BD). The central black line represents the median curve, while the shaded purple area shows the central region of the data, indicating high centrality. The blue lines represent individual curves within the dataset, showing the range of variability. This visualization highlights the centrality of functional curves, with the purple area representing curves enclosed by others within their convex hull.

As we see, in the following image the **outliers disappear**:



Let’s repeat using the method “Both”: it uses BD2 first and then uses MBD to break ties. It gives the same output of BD2.



Let’s check if the ‘**median curve**’ is the same with the three methods:

```
> b1$medcurve
Macular.dystrophy
        21
> b2$medcurve
External.ophthalmoplegia
            19
> b3$medcurve
External.ophthalmoplegia
            19
```

The result is not same. The result compares the 'median curve' identified by three different methods, labeled as `b1`, `b2`, and `b3`. The `median curve` represents the curve that is most central to the data set in each method.

- For method `b1`, the median curve is identified as "Macular dystrophy" with a value of 21.
- For methods `b2` and `b3`, the median curve is identified as "External ophthalmoplegia," both with a value of 19.

This indicates that methods `b2` and `b3` agree on the median curve being "External ophthalmoplegia" with the same value, while method `b1` identifies a different median curve, "Macular dystrophy," with a different value. This suggests some variation in the determination of the central curve depending on the method used.

We use the output of MBD method, which usually is the best, and extract the band depths.

Let's check, as an example, if the RE_1 signal can be considered as the others, using a Wilcoxon rank test applied to the depth measures. The Wilcoxon test in this case is testing H0: the means in the two groups are equal against all alternatives.

It is non-parametric, thus we don't need to make assumptions on the distribution of the depth measures.

Result:

```
> wilcox.test(D_Male,D_Female)
data: D_Male and D_Female
W = 98, p-value = 0.1339
alternative hypothesis: true location shift is not equal to 0.
```

The p-value is more than 0.05, meaning that there is not a significant difference between the RE_1 Signal values of Male and Female.

The Wilcoxon rank-sum test is a non-parametric statistical test used to determine if two independent samples come from populations with the same distribution. It is particularly useful when the assumptions of the t-test, such as normality and equal variances, are not satisfied. Instead of comparing means, the Wilcoxon rank-sum test compares the medians of the two samples by ranking all observations together and evaluating whether one group consistently has higher or lower values than the other.

Conclusion

In this project, we focused on analyzing the PERG signal data for both right and left eyes using various smoothing methods with different numbers of basis functions (nbasis) and orders (norder). The aim was to determine the optimal configuration for accurately capturing the functional characteristics of the data.

We experimented with different combinations of nbasis and norder to assess their impact on the smoothness and fit of the functional curves. By adjusting these parameters, we were able to observe how the smoothing methods influenced the representation of PERG signal over time and across different diagnosis types.

We explored various methods to analyze functional data, specifically focusing on the visual acuity and PERG signals of the right eye (RE_1) and left eye (LE_1) across different diagnosis types. We utilized Modified Band Depths (MBD) and Band Depth (BD) measures to quantify the centrality of functional curves, providing insights into the distribution and variability of visual acuity within the dataset.

Our analysis included a thorough examination of regression models to assess the relationship between functional predictors and visual acuity as a response variable. The regression results indicated that while the model explained a moderate amount of variance, the significance of the predictors was marginal. This suggested potential overfitting and highlighted the need for further refinement and validation of the model.

The confusion matrix initially suggested perfect classification, but subsequent testing revealed discrepancies, indicating overfitting to the training data and poor generalizability to new data. This emphasized the importance of robust model evaluation and the potential need for alternative modeling approaches or additional data preprocessing.

We also conducted a Wilcoxon rank-sum test to compare the depth measures of RE_1 signals between male and female groups. The non-parametric nature of this test allowed us to avoid assumptions about the distribution of the depth measures. The test result showed no significant difference between the groups, indicating similar centrality and distribution of visual acuity measures across genders.

In summary, this project provided valuable insights into the functional data analysis of visual acuity, highlighting the strengths and limitations of various methods. The findings underscore the importance of using appropriate depth measures, rigorous model evaluation, and the potential for further research to improve the robustness and generalizability of the models. These conclusions can inform future studies and clinical practices in assessing and understanding visual acuity and type of diagnosis across different conditions and populations.

References:

1. Course marerials

2. Joint Research Centre. "Impacts of climate change in agriculture in Europe."

3. Ramsay, J. O., & Silverman, B. W. (2002). Applied Functional Data Analysis: Methods of Applied Functional Data Analysis. Springer.

4. Ramsay, J. O., Hooker, G., & Graves, S. (2009). Functional Data Analysis with R and MATLAB. Springer.

5. https://rviews.rstudio.com/2021/05/04/functional-data-analysis-in-r/

6. https://cran.r-project.org/web/views/FunctionalData.html

Dataset link:

https://www.physionet.org/content/perg-ioba-dataset/1.0.0/csv/#files-panel