Functional Data Analysis of Spectometric Data

Ketrin Hristova
13209A

Functional and Topological Data Analysis

Abstract

This project explores the application of Functional Data Analysis (FDA) on spectrometric data from meat samples to investigate the relationship between absorbance spectra and chemical composition (fat, water, and protein content). By employing smoothing, functional principal component analysis (fPCA), and other FDA methods, the project aims to identify key spectral features, reduce dimensionality, and develop predictive models that can accurately assess meat quality based on spectrometric data.

1 Introduction

Functional Data Analysis (FDA) is a statistical approach designed to analyze data that can be represented as functions, curves, or shapes. In the context of this project, FDA techniques are applied to spectrometric data from meat samples to investigate how the absorbance spectra across various wavelengths relate to the chemical composition of the meat, specifically focusing on fat, water, and protein content.

The Tecator dataset, available in the `fda.usc` package and originally part of a larger dataset from Carnegie Mellon University, includes 215 finely chopped meat samples with spectrometric measurements and corresponding chemical compositions. For each sample, the dataset provides a spectrometric curve that represents the absorbance measured at 100 wavelengths ranging from 850 to 1050 nm. The samples are categorized into two classes based on fat content—"small" (¡20%) and "large"—as defined by Ferraty and Vieu (2006). This dataset offers a functional data structure where each observation consists of a wavelength (args) and its corresponding absorbance value (vals), allowing for a detailed functional representation of each meat sample's spectral profile. Understanding these profiles is crucial for developing methods to accurately assess the quality of meat based on its chemical constituents.

2 Methodology

2.1 Basis Function Systems in FDA

In functional data analysis (FDA), basis function systems are essential for constructing representations of continuous data curves. Basis functions, denoted as ϕ_k for $k = 1, \ldots, K$, serve as functional building blocks. A function $x(t)$ can be expressed as:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{c}'^{\boldsymbol{\phi}(t)}$$

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

B-splines

B-splines are piece-wise polynomial functions that provide local control over the shape of the curve. They are defined recursively using the Cox-de Boor recursion formula:

$$\phi_{k,1}(t) = \begin{cases} 1 & \text{if } t_k \leq t < t_{k+1} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{k,i}(t) = \frac{t - t_k}{t_{k+i-1} - t_k} \phi_{k,i-1}(t) + \frac{t_{k+i} - t}{t_{k+i} - t_{k+1}} \phi_{k+1,i-1}(t)$$

Here, the knots are denoted by $t_0, t_1, \ldots, t_{m+k}$. This recursion formula defines each B-spline basis function in terms of its lower-order counterparts.

Fourier Basis Functions

Fourier basis functions, derived from trigonometric sine and cosine functions, are particularly useful for analyzing periodic or cyclical data. The Fourier basis functions are defined as:

$$\phi_1(t) = 1$$

$$\phi_2(t) = \sin(\omega t)$$

$$\phi_3(t) = \cos(\omega t)$$

where $\omega = \frac{2\pi}{T}$ is the fundamental frequency for the period T.
These basis functions are ideal for representing data with repeating patterns and can capture periodic behavior effectively.

Smoothing Techniques

Smoothing techniques are employed in functional data analysis to reduce noise and variability while preserving essential features. This step is crucial for enhancing the signal-to-noise ratio and improving data interpretability. In essence, smoothing involves creating an approximate function that captures key patterns while filtering out noise and intricate structures. During this process, individual data points within a signal are adjusted to reduce discrepancies between neighboring points, typically caused by noise, resulting in a more uniform signal.

Smoothing serves two primary purposes in data analysis: it facilitates the extraction of relevant information when the assumption of smoothing is justified, and it enables analyses that are both flexible and robust. Various algorithms are used in smoothing to achieve these objectives.

Smoothing aims to provide a broad depiction of gradual changes in value, with less emphasis on precise data point matching, contrasting with the meticulous pursuit of an optimal fit in curve fitting. Moreover, smoothing methods often include a tuning parameter to regulate the degree of smoothing, whereas curve fitting adjusts multiple parameters to achieve the best fit.

The general idea can be summarized as:

$$\hat{c} = \arg\min_{c} \left\{ \sum (y_i - x(t_j))^2 + \lambda \int [Lx(t)]^2 dt \right\}$$

Since we have $x(t) = \phi'(t)c$, then we can write:

$$\hat{c} = \arg\min_{c} \left\{ \sum (y_i - \phi'(t_j)c)^2 + \lambda c' \left[\int L\phi(t) L\phi'(t) dt \right] c \right\}$$

Here, λ is the penalty parameter and $\lambda c' \left[\int L\phi(t) L\phi'(t) dt \right] c$ is the penalty term.

Generalized Cross-Validation (GCV)

Generalized Cross-Validation (GCV) is a statistical method used to select the optimal smoothing parameter in nonparametric regression, such as B-splines and other smoothing techniques. The smoothing parameter, λ, controls the balance between the fidelity of the model to the data and the smoothness of the estimated function.

In functional data analysis, overfitting occurs when λ is too low, capturing noise in the data, while oversmoothing happens when λ is too high, leading to an overly simplistic model. GCV addresses this by providing an efficient, computationally inexpensive alternative to traditional cross-validation methods.

GCV minimizes a criterion that balances the model's fit against its complexity. The GCV score is computed as:

$$\text{GCV}(\lambda) = \frac{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}_\lambda(x_i) \right)^2}{\left(\frac{1}{n} \text{tr} \left(\mathbf{I} - \mathbf{S}_\lambda \right) \right)^2}$$

where n is the number of observations, y_i are the observed values, $\hat{f}_\lambda(x_i)$ are the fitted values, and $\mathbf{S}_\lambda$ is the smoothing matrix. The numerator represents the model's mean squared error, and the denominator penalizes the model's complexity.

GCV is particularly useful in functional data analysis because it automatically adjusts for model complexity and is computationally efficient, requiring only a single fit of the model. This makes GCV a valuable tool for selecting the optimal smoothing parameter, ensuring that the resulting models are both accurate and interpretable.

Practical Application

Before discussing the development and application of smoothing techniques, an introduction and study of the variability and distribution in meat contents were conducted. Below are shown the histograms of three different chemical components: fat, water, and protein content.

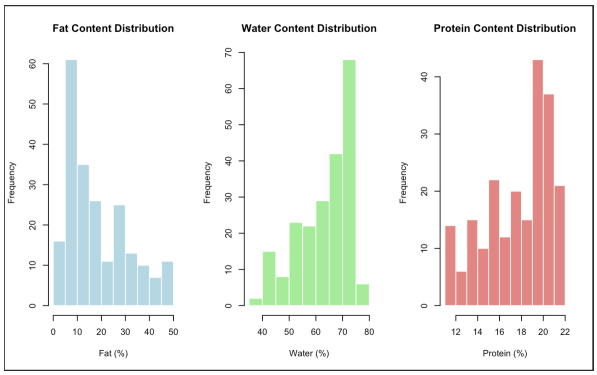

Figure 1: Histograms of the chemical components: fat, water, and protein content.

- **Fat Content**: Most samples exhibit low fat levels, typical of many biological tissues.
- **Water Content**: The majority of samples are high in water, reflecting the common characteristic of biological tissues.
- **Protein Content**: There is a notable variation, with a significant number of samples showing relatively high protein content.

Next, different smoothing techniques were applied to spectrometric curves to reduce noise and capture the underlying structure of the data.

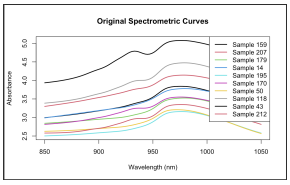

Figure 2: * Original Spectrometric Curves

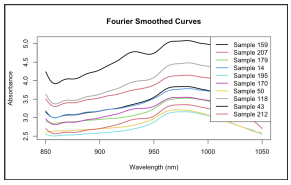

Figure 3: * Fourier Smoothed Curves

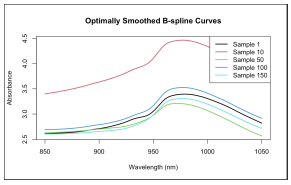

Figure 4: * B-spline Smoothed Curves

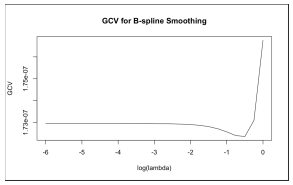

Figure 5: * GCV for B-spline Smoothing

Figure 6: Comparison of Smoothing Techniques Applied to Spectrometric Curves

The B-spline smoothed curves effectively capture the general shape and trends of the original absorbance data, reducing noise while preserving essential features. In contrast, the Fourier smoothed curves, while also effective, may introduce artifacts, particularly near the boundaries or regions with rapid changes. This difference arises because Fourier basis functions are global, meaning each function spans the entire range of data, potentially leading to less effective handling of localized features.

To further optimize the B-spline smoothing, Generalized Cross-Validation (GCV) was employed. GCV is a method used to select the optimal smoothing parameter (lambda) by balancing the trade-off between the smoothness of the curve and its fidelity to the data. The GCV curve indicates that the B-spline smoothing is robust across a range of lambda values, as the GCV score remains relatively flat before sharply increasing at higher values. This increase suggests that over-smoothing, due to excessive regularization, deteriorates the fit, while very low lambda values may lead to under-smoothing. Using the optimal lambda identified through GCV, the B-spline smoothing was applied to the spectrometric data, yielding the curves. These curves represent the best possible fit, balancing smoothness and accuracy.

In conclusion, the B-spline smoothing method, especially when optimized using GCV, is well-suited for analyzing the Tecator dataset's spectrometric data. Compared to Fourier smoothing, B-splines offer greater flexibility in capturing localized variations in the data, making them more appropriate for this type of analysis.

2.2 Functional Principal Component Analysis

The goal of Principal Component Analysis (PCA) is to uncover the primary modes of variation within the dataset and evaluate their significance. Similar to multivariate statistics, the eigenvalues obtained from the variance-covariance function indicate the importance of these principal components.
In Functional PCA, each eigenvalue is associated with an eigenfunction rather than an eigenvector, which describes the key components of variation. Rotating these functions can provide a clearer understanding of the dominant modes of variation in the functional data while preserving the total common variation.
The primary equation involves the expression for the probe score variance related to a probe weight ξ. It is defined as the maximum value of the sum of squared probe scores, subject to the constraint that the squared integral of the weight function ξ over its domain equals 1. Mathematically, it is expressed as:

$$\mu = \max\left\{\sum \rho_{\xi}^{2}(x_i)\right\} \quad \text{subject to} \quad \int \xi^{2}(t)\,dt = 1$$

Additionally, constructing eigenvalue/eigenfunction pairs through eigen-analysis and determining the optimal number of harmonics to use in PCA involves visual inspection of eigenvalue plots, known as scree plots. This process also touches on the interpretability of PCA results and considers the potential for using alternative optimal basis systems.

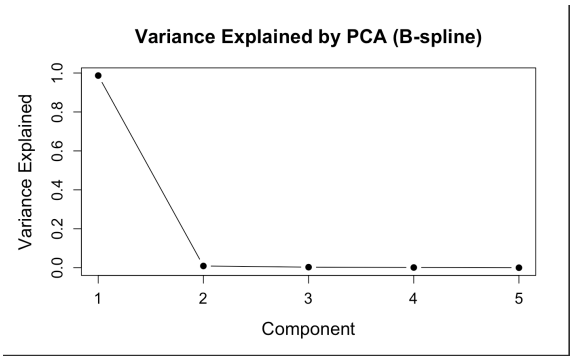

Figure 7: Variance Explained by PCA (B-spline)

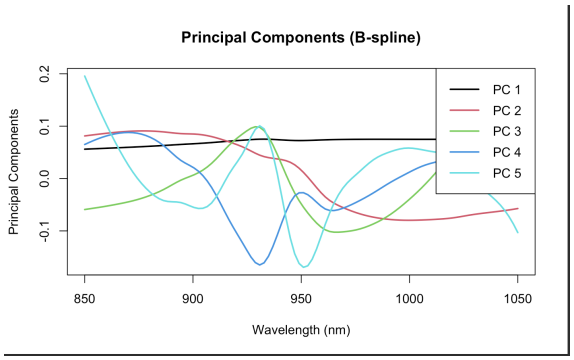

Figure 8: Principal Components (B-spline)

Figure 9: FPCA Analysis using B-spline Basis

The first plot displays the proportion of variance explained by each of the first five principal components (PCs) derived from the B-spline smoothed data. The first principal component (PC1) accounts for nearly all the variance in the data, as indicated by a variance explained of approximately 0.987 (98.7The subsequent components (PC2,

PC3, etc.) explain very little variance, indicating that the majority of the data's variability can be captured by the first component alone. This suggests that the underlying structure of the dataset is dominated by a single major trend or pattern.

The second plot shows the first five principal component functions for the B-spline smoothed data across different wavelengths. The shapes of the principal component functions indicate how each component contributes to variations in the data. PC1 (black line) shows a relatively flat trend, implying that it represents the overall level of absorbance across all wavelengths. The other components (PC2, PC3, etc.) show more oscillatory behavior, capturing finer details and variations in the absorbance data that PC1 does not account for. However, since these components explain very little variance, their influence on the overall data structure is minimal.

The PCA results on the B-spline smoothed Tecator dataset reveal that the first principal component captures almost all the variance in the data. This suggests that a single dominant factor (e.g., an overall trend across all wavelengths) largely drives the variability in the dataset. The remaining principal components capture only minor variations and are less significant in explaining the data's structure. The B-spline smoothing has likely enhanced the underlying signal, making it easier to capture the primary trend with just the first principal component.

2.3 Depth Measures

In functional data analysis, depth measures are used to determine how central or outlying a particular function is within a sample of functions. The idea is similar to identifying central tendencies or outliers in traditional univariate or multivariate data. The Modified Band Depth (MBD) and the second band depth (BD2) methods are two common techniques for calculating depth measures in functional data.

MBD (Modified Band Depth): This method assesses the centrality of a curve by considering how often it lies within the band formed by other curves. The higher the frequency with which a curve lies within these bands, the higher its depth measure, implying it is more central.

BD2 (Second Band Depth): Similar to MBD, this method looks at how frequently a curve lies within the central region formed by other curves. However, it uses a different approach to define and assess these bands.

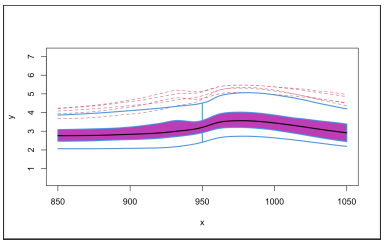

Figure 10: MBD Method

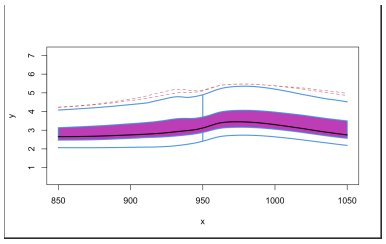

Figure 11: BD2 Method

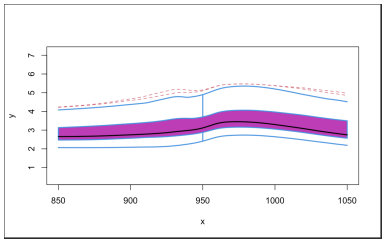

Figure 12: Combined MBD and BD2

Figure 13: Comparison of Depth Measures in the Tecator Dataset

The first plot shows the curves with the MBD depth measure applied. The central black curve represents the median curve, which is the most central according to the MBD method. The blue lines represent the central region, within which most of the data lies. The red dashed lines indicate the outer limits, beyond which the data is considered more outlying. The purple shaded area indicates the interquartile range (IQR) of the depth values, emphasizing the spread and central tendency of the data.

The second plot displays the curves with the BD2 depth measure applied. The structure is similar to the MBD plot, with the central curve (black), central region (blue), and outer limits (red dashed lines). The BD2 method typically yields slightly different centrality measures than MBD, which may result in different identifications of the median and the range of central curves.

The analysis using MBD and BD2 methods on the Tecator dataset provides insights into the centrality and variability of the absorbance spectra. Both methods generally agree on the identification of central curves but may differ slightly in the extent and interpretation of centrality. The depth measures reveal that most curves exhibit moderate centrality, with some notable outliers that lie further from the central band.

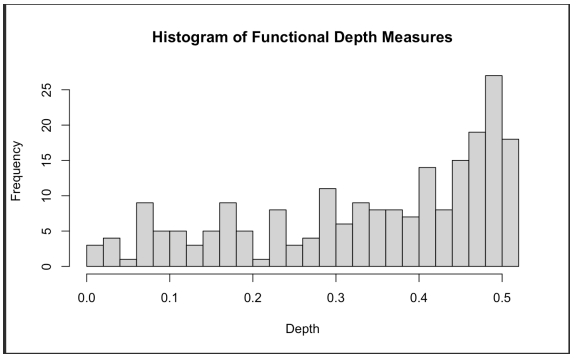

Figure 14: Histogram of Functional Depth Measures

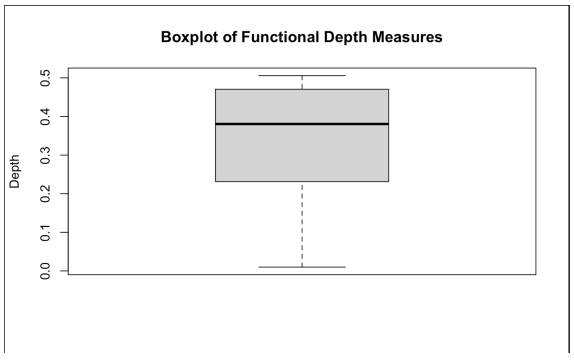

Figure 15: Boxplot of Functional Depth Measures

Figure 16: Distribution and Summary of Functional Depth Measures

The histogram illustrates the distribution of depth values across the dataset, highlighting the frequency of various levels of centrality among the sample curves. A bimodal or widely spread distribution could indicate the coexistence of both central and outlying curves within the dataset. The boxplot offers a concise visual summary of the depth measures, showcasing the median, interquartile range (IQR), and potential outliers. A large IQR or the presence of outliers suggests notable variation in the centrality of the curves within the dataset, indicating that some curves may deviate significantly from the central tendency.

The depth measures are crucial for understanding the structure and variability of the functional data. In the Tecator dataset, both MBD and BD2 methods highlight the presence of a strong central tendency with some variability, but no extreme outliers. The absence of extreme outliers and the presence of a well-defined central region suggest that the Tecator dataset is relatively homogeneous, with most curves following a similar pattern.

2.4 Functional Linear Regression

Functional Linear Regression (FLR) extends classical linear regression to cases where the predictors are functions rather than scalars. In the **scalar-on-function** model, a scalar response y_i is regressed on a functional predictor $X_i(t)$, represented as:

$$y_i = \alpha + \int_{\mathcal{T}} X_i(t)\beta(t)\,dt + \epsilon_i$$

Here, $\beta(t)$ is the coefficient function that indicates the influence of different parts of the functional predictor on the response.

To estimate $\beta(t)$, both $X_i(t)$ and $\beta(t)$ are typically expanded using basis functions, such as Fourier or B-splines, allowing the problem to be solved with standard linear regression techniques.

FLR is particularly useful in situations with high-dimensional functional data, such as in chemometrics or biostatistics, where it captures complex relationships between the functional predictors and the response variable. This approach enables a detailed understanding of how the shape of the functional predictor influences the outcome, offering insights that traditional regression models may miss.

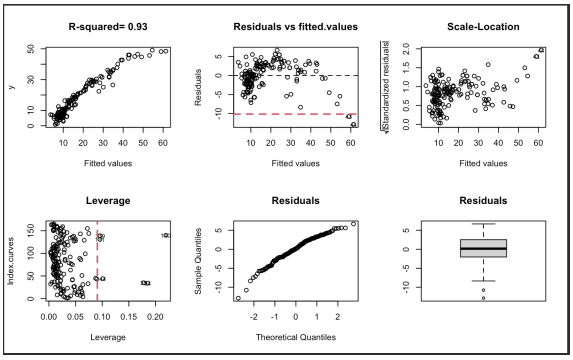

Figure 17: Model Diagnostic Plots

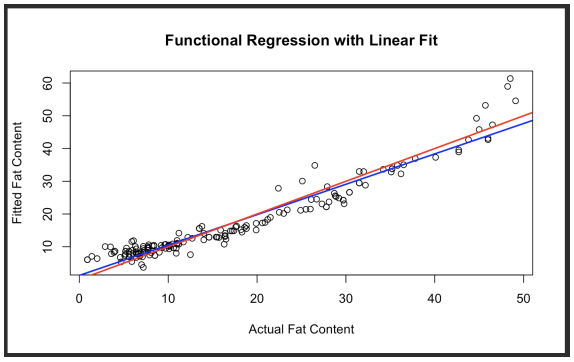

Figure 18: Functional Regression with Linear Fit

Figure 19: Diagnostic and Fitted vs Actual Plots for Functional Regression Model

The goal of this section is to predict the fat content in meat samples using the first derivative of the absorbance spectra as the functional predictor. The R-squared value of 0.93 indicates that 93% of the variance in the fat content is explained by the model, suggesting a strong relationship between the functional predictor and the response variable. the residuals are scattered randomly around the zero line, with no apparent patterns, indicating that the model fits the data well. The spread-location plot reveals that the residuals' variance is fairly consistent across the range of fitted values, though some increase in variability is observed at higher fitted values, suggesting slight heteroscedasticity. Most observations have low leverage, indicating that no single data point has an undue influence on the model. The Q-Q plot shows that the residuals closely follow the normal distribution line, suggesting that the residuals are approximately normally distributed. The boxplot further supports this, showing a fairly symmetric distribution with a few outliers.

The second plot shows that the points closely follow the 45-degree reference line (red), indicating that the model's predictions align closely with the observed fat content values. This plot supports the high R-squared value obtained in the model. The linear regression line (blue) is also close to the 45-degree reference line, further confirming the model's accuracy. However, slight deviations at higher fat content values suggest that the model may slightly underpredict or overpredict at the extremes.

The FLR model applied to the Tecator dataset demonstrates strong predictive performance, with the first derivative of the absorbance spectra effectively predicting the fat content in meat samples. The diagnostic plots and high R-squared value suggest that the model fits the data well, with minimal issues related to non-linearity or heteroscedasticity. Overall, the FLR model proves to be a robust tool for analyzing this dataset.

3 Conclusion

In this report, we have explored various functional data analysis techniques to analyze the Tecator dataset. This foundational analysis started with the application of smoothing techniques to the spectrometric curves, specifically comparing B-spline and Fourier smoothing methods.

The B-spline smoothing, particularly when optimized using Generalized Cross-Validation (GCV), demonstrated superior performance in capturing the underlying structure of the spectrometric data while effectively reducing noise. This method proved to be more flexible in handling localized variations in the data compared to the global nature of Fourier smoothing, making it better suited for the analysis of the Tecator dataset.

Following the smoothing procedures, Functional Principal Component Analysis (FPCA) was applied to the smoothed data to identify the primary modes of variation. The FPCA results revealed that the first principal component captured nearly all the variance in the dataset, indicating a dominant trend across the entire spectrum. This insight simplifies the complexity of the data and underscores the effectiveness of the B-spline smoothing in enhancing the primary signal.

We then examined the centrality and variability of the functional data using depth measures, specifically the Modified Band Depth (MBD) and second band depth (BD2) methods. These methods provided valuable insights into the distribution of the spectrometric curves, identifying central tendencies and potential outliers within the dataset. The depth measures confirmed the relative homogeneity of the dataset, with most curves adhering to a similar pattern.

Finally, the Functional Linear Regression (FLR) model was employed to predict fat content based on the first derivative of the absorbance spectra. The FLR model demonstrated strong predictive accuracy, as evidenced by a high R-squared value of 0.93. The diagnostic plots indicated a well-fitting model with minimal issues, such as non-linearity or heteroscedasticity. The model's predictions closely aligned with the observed fat content values, further validating its effectiveness.