SAND2020-8138C

Explaining Neural Network Predictions for Functional Data Using Principal Component and Feature Importance

Katherine Goode, Daniel Ries, Joshua Zollweg

Sandia National Laboratories

Abstract

The shape of optical spectral-temporal signatures extracted from videos of explosions provides information for identifying characteristics of the explosive devices corresponding to the signatures. It is of interest to use machine learning algorithms such as neural networks to improve upon predictions made by the methods currently used in practice. Since this application lends itself to high consequence national security decisions, it is important to provide explanations for predictions made by the neural networks to garner confidence in the model. While work has been done to develop explainability methods for neural networks, not much of the work has focused on situations with functional data as the input variables. We demonstrate a technique that makes use of functional principal component analysis (fPCA) and permutation feature importance (PFI). fPCA is used to transform the signatures to create uncorrelated functional principal components (fPCs). Neural networks are trained using the fPCs as inputs to predict a characteristic of explosive devices, and PFI is applied to identify the fPCs important for the predictions. Visualizations are used to interpret the variability explained by the fPCs that are found to be important by PFI to determine the aspects of the signatures that are important for the neural network predictions.

1 Introduction

The predictive ability of neural networks have made them desirable tools in many applications including national security. However, the predictive ability of neural networks, and many machine learning algorithms, comes at the cost of interpretability due to the complicated nature of the underlying algorithms. The ability to interpret a model allows users to understand how the model makes predictions and assess the trustworthiness of the model. When it is not possible to directly interpret a model, there is still a need to provide indirect explanations for the predictions, which especially holds true in areas with high stakes decisions such as national security.

The identification of explosive device characteristics based on optical spectral-temporal signatures of explosions is an example of an area in national security where machine learning could improve predictive performance. Currently, the identification of explosive device characteristics is done using heuristic algorithms or direct subject matter expert (SME) review. While neural networks may provide a more accurate identification method for this application, it is imperative that a machine learning based method not only return high accuracy but an explanation for the prediction.

The optical spectral-temporal signatures used to identify the explosive device characteristics are functional data (Figure 1). That is, an observation corresponding to an explosive device is a function. While much research has been done relating to the explainability of neural networks [Hohman et al.,

2018, Montavon et al., 2017], little work has focused on explaining neural network predictions with inputs of a functional data nature. In this paper, we present an approach for explaining predictions made by neural networks with the optical spectral-temporal signatures from explosions as the model input and explosive device characteristics as the model output using functional principal components analysis (fPCA) and permutation feature importance (PFI).

fPCA is a common technique used in the analysis of functional data to understand the variability present in the functions [Ramsay and Silverman, 2005, Wang et al., 2015]. Similar to multivariate data principal components analysis (PCA), fPCA is a dimension reduction technique that transforms the observed data into functional principal components (fPCs). The fPCs are uncorrelated and ordered such that the first and last fPCs explain the largest and smallest amounts of variability, respectively.

PFI was originally developed by Breiman [2001] for random forests, and Fisher et al. [2018] generalized the method to any predictive model. The concept of PFI is to apply a trained model to the data (training or testing) with one feature randomly permuted, and if the predictions worsen significantly when the feature is randomly permuted, the feature is considered important for prediction. In particular, a loss function is used to compare the predictions from both the permuted and non-permuted data to the true response values. A difference between the two losses is computed and summed over all observations in the data. This process is performed for all features to identify the features with the largest positive loss (i.e. the features that are most important.) Note that a negative value of PFI indicates that a randomly permuted feature results in better predictions than the observed feature.

Our interest in using PFI is based on two reasons. First, it is important for the neural network explanations to be understood by both the data analysts, the scientists, and the government decisions makers. It is likely that the familiarity with neural networks decreases from the data analysts to decision makers who may have little understanding of machine learning models, but PFI is an easily understood explainability method. Second, PFI can be applied to any predictive models, so we could use it compare the feature importance from various machine learning models. However, in this paper, we focus on neural networks.

A disadvantage of PFI is that it is known to produce biased results with neural networks when there is correlation among the features [Hooker and Mentch, 2019]. As is natural with functional data, there is high correlation between individual time points in optical spectral-temporal signatures (Figure 1). To eliminate correlation in our model features, we transform the signatures using fPCA and use the uncorrelated fPCs as the features for training the neural network. As a result, we are able to apply PFI to identify the fPCs important for the prediction of an explosive device characteristic without any concern of bias in the PFI due to correlation of the features. We use visualizations of the fPCs to understand the variability explained by the important fPCs to connect the aspects of the signatures to the prediction of an explosive device characteristic. The visualizations of the fPCs are presented to an SME who is able to confirm that the aspects of the signatures important to the neural network for prediction are what is expected based on the signature generation mechanisms.

This paper is organized as follows. In Section 2, the simulated signatures used for analysis are described. Section 3 details the application of fPCA and PFI to explain the neural network predictions based on the signatures, and Section 4 presents the results. We discuss our conclusion in Section 5 about the ability to put trust in the neural network based on the explanations produced by this approach along with limitations and future research directions.

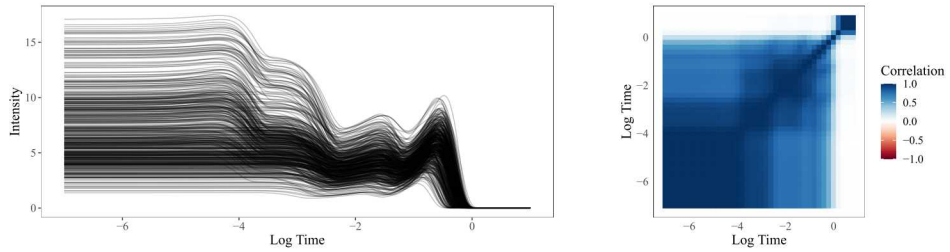

Figure 1: (Left) Example of optical spectral-temporal signatures from explosions. Each function represents a signature associated with an explosive device from a subset of 500 observations from the simulated data. (Right) Pearson correlations between intensity vectors at every 25th time point in the signatures.

2 Simulated Optical Spectral-Temporal Explosion Signatures

We consider a set of simulated signatures for the application of our explainability method. A subset of the simulated signatures is shown in Figure 1. A total of 10000 signatures are created with 1000 time points per signature. It is customary to consider the signatures on a log time scale since the events of interest occur within a short period of time.

The signatures are generated based on the scientific understanding of the relationship between three explosive device characteristics (Y_1, Y_2, and Y_3) and the corresponding signatures. Characteristics Y_1 and Y_2 are binary variables, and characteristic Y_3 is a continuous variable. The characteristics affect various aspects of the signatures including the intensity, location of peaks, and number of peaks. These effects are visible in Figure 2, which shows the point-wise functional means for the categories of Y_1 and Y_2 and quartile bins for Y_3 computed on the training data. In particular, Y_1 affects the intensity of the signature early on, the timing of the first peak, and the total number of peaks (3 or 4). Y_2 affects the intensity of the signature over the entire time. Y_3 also affects the intensity of the signature over the entire time, but in addition, Y_3 affects the timing of all peaks.

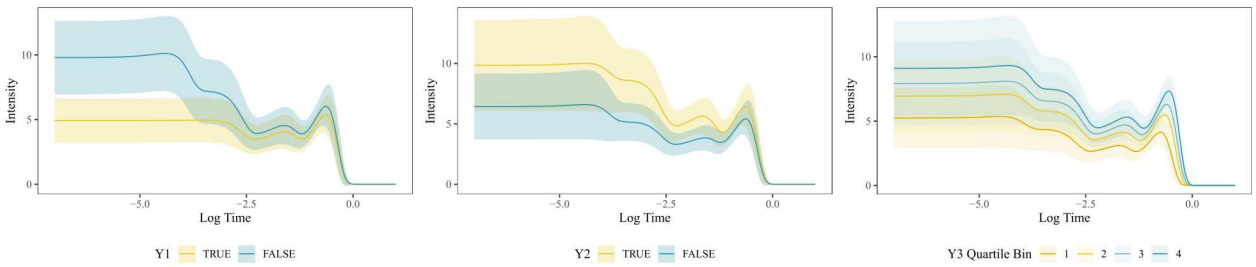

Figure 2: Pointwise functional means plus/minus pointwise one standard deviations for the two categories of Y_1 and Y_2 and quartile bins of Y_3.

3 Methods

The simulated data are randomly separated into training, testing, and validation sets containing 72.25% (7225), 15% (1500), and 12.75% (1275) of the signatures, respectively. Each of the 1000 time points in the signatures are treated as a feature, and fPCA is applied to convert the 1000 features to 1000 fPCs. The percent of variability explained by each fPC is assessed. The estimated

eigenfunctions from fPCA are used to transform the testing and validation data sets to fPCs. Note that in fPCA, the eigenfunctions are comparable to the eigenvectors in PCA.

A neural network is trained for each of the three explosive device characteristics. The 1000 training data fPCs are used as the features, and the corresponding vector of characteristics are used as the outputs. All models are fit using 3 layers with 50, 40, and 30 nodes, respectively. The transformed testing and validation features are used to assess model performance with the metrics of accuracy and F_1 for Y_1 and Y_2 and mean squared error (MSE) and R^2 for Y_3.

PFI is applied to the trained networks using 10 replications to account for random permutation variability. The most important fPCs identified by PFI are visualized to interpret the variability in the signatures explained by the fPCs. We consider three visualizations that convey the variability explained by the fPCs from different perspectives:

1. *Eigenfunction*: An eigenfunction possesses weights that correspond to the times on the original scale. The magnitude of a weight indicates the importance that a time plays in the variability captured by the corresponding fPC. Larger magnitudes indicate more importance. Thus, a plot of the eigenfunction identifies the modes of variability associated with the fPC. If all weights are positive (or negative), the eigenfunction represents a weighted average of times. If there are both positive and negative weights, the eigenfunction identifies that the fPC captures a contrast between the time intervals with positive and negative values.

2. *Point-wise functional mean plus/minus the eigenfunction*: Adding the eigenfunction weights (times the fPC standard deviation(s)) to the point-wise functional mean allows for a visualization of the principal component directions. That is, the point-wise functional mean plus/minus the eigenfunctions depicts the shapes of the functions with high/low fPC values.

3. *Signatures with extreme fPCs*: The observed signatures corresponding to the 50 highest and 50 lowest fPC values are identified and visualized along with the point-wise functional mean. The contrast in shapes of functions with high and low fPC values helps to identify the type of functional variability captured by the fPC.

fPCA and visualizations of the fPCs are performed using R 3.6.1 [R Core Team, 2019]. All visualizations are created using the R package ggplot2 (3.3.0) version [Wickham, 2016]. The neural networks and PFI are applied using Python 3.8.2 [Van Rossum and Drake, 2009] and the scikit-learn (0.22.1) package [Pedregosa et al., 2011].

4 Results

The first fPC explains 94% of the variability, and the first three fPCs combined explain 99% of the variability (Figure 3). All neural networks perform well (Figure 3). The fPCs identified as important by PFI are within the first 4 fPCs for all models (3). For Y_1, fPCs 1 and 2 are the most important, for Y_2, fPCs 1 and 3 are the most important, and for Y_3, fPC 2 is the most important with some importance for fPCs 1, 3, and 4. The fPCs greater than 10 had negligible feature importance. Since the first four fPCs are found to be the most important based on PFI, these will be the only fPCs considered for interpretation.

Figure 4 includes scatter plots of the relationships between the explosive device characteristics and the corresponding top one or two most important fPCs identified by PFI. In all three plots, there are clear separations between Y_1 and Y_2 categories. For example, fPC 2 versus fPC 1 provides a clear separation between the categories of Y_1 and a distinction of the categories of Y_2 within the Y_1 categories. The plot of Y_3 versus fPC 2 shows a weak negative relationship between Y_3 and fPC 2.

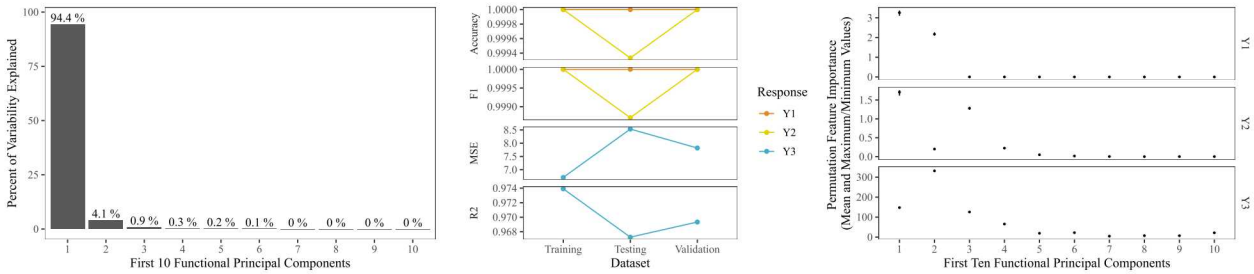

Figure 3: (Left) Percent of variation explained by the first 10 fPCs. (Middle) Performance metrics for the neural networks computed on the training, testing, and validation datasets. (Right) Mean (circle) and maximum/minimum (bars) PFI values for the first 10 fPCs computed from 10 replications. The other fPCs had negligable PFI values.

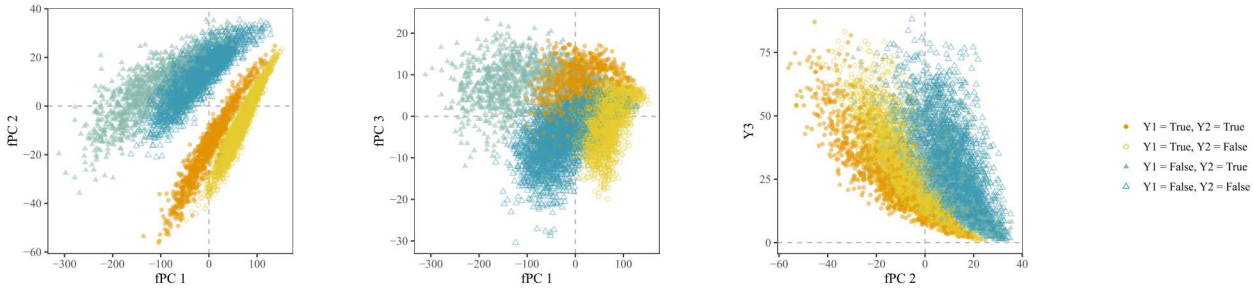

Figure 4: Scatter plots depicting relationships between key fPCs and response variables.

Visualizations for interpreting fPCs 1 to 4 are included in Figure 5. The interpretation of fPCs 3 and 4 are not as clear as fPCs 1 and 2 since fPCs 3 and 4 explain a smaller amount of variability. However, the variability explained by all fPCs that PFI identifies as important for prediction can be connected to the effects caused by the explosive device characteristics. The fPCs are interpreted as follows:

- *fPC 1*: The eigenfunction for fPC 1 makes it clear that fPC 1 is a weighted average over all times. The visualizations of the point-wise mean function plus/minus the eigenfunctions and the signatures corresponding to the 50 highest and lowest fPC 1 values indicate that fPC 1 captures a contrast between signatures with high intensity starting values, a large decrease in intensity over time, and four peaks and signatures with low starting values, a relatively contrast value over all times, and three peaks.

- *fPC 2*: The eigenfunction for fPC 2 makes it clear that fPC 2 explains a contrast between time points before and after -3.75. The visualizations of the point-wise mean function plus/minus the eigenfunctions and the signatures corresponding to the 50 highest and lowest fPC 2 values indicate that fPC 2 captures a contrast between signatures with high starting values and 3 peaks that occur after the mean function and signatures with lower starting values and four peaks before the mean function.

- *fPC 3*: The eigenfunction of fPC 3 indicates that the fPC explains a contrast in variability between the times of (-3.75, -1.75) and (-1.75, 0). The mean plus/minus eigenfunctions and signatures with extreme fPC 3 values suggest that fPC 3 captures the variability between signatures with lower values between the first time interval and a large fourth peak during the

second time interval and signatures with higher values during the first time interval and a small fourth peak during the second time interval.

- *fPC 4*: The eigenfunction for fPC 4 depicts that fPC 4 explains a contrast between the two time intervals of (-4,-2.5) and (-0.5,0) and the time interval of (-2.5,-0.5). The other two visualizations for fPC 4 indicate that fPC 4 captures the variability between signatures with a steep decrease during the time interval of (-4,-2.5) and dramatic third and fourth peaks during the time intervals of (-2.5,-0.5) and (-0.5,0), respectively and signatures with less intense peaks throughout the entire time interval of (-4,0).

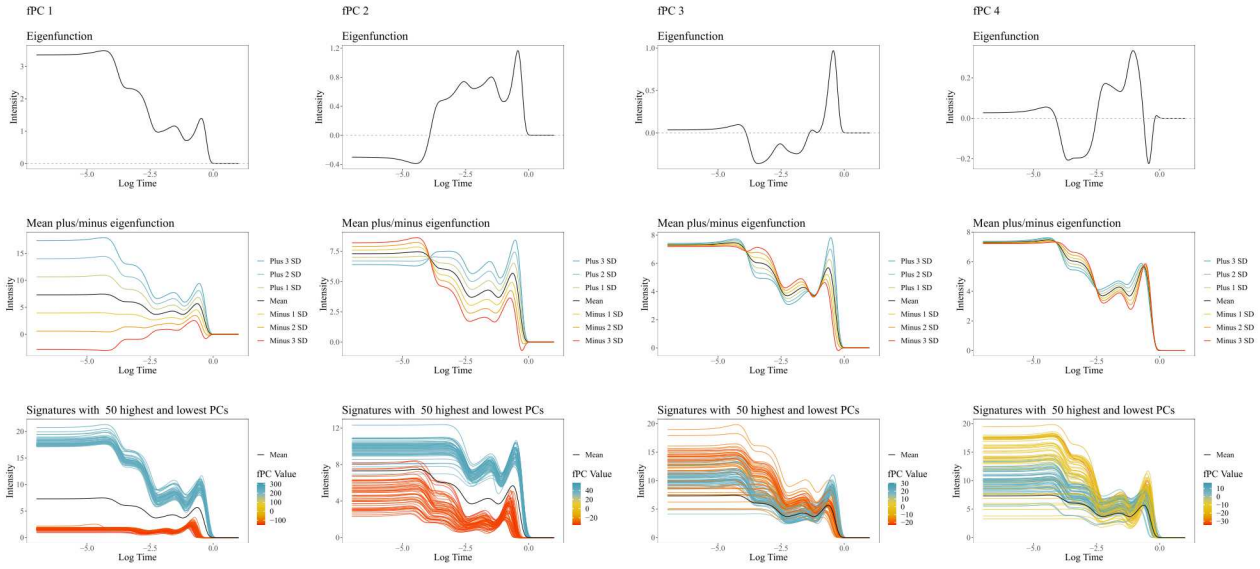

Figure 5: Visualizations of the first four fPCs.

To connect the interpretations of the first four fPCs to the explosive device characteristics, consider the functional means for the Y_1, Y_2, and Y_3 in Figure 2. PFI identified fPCs 1 and 2 as being important for predicting Y_1. This is reasonable since both fPCs capture a variability in functions with intensity during early times, different timings for peak 1, and the number of peaks (3 and 4). PFI found that fPCs 1 and 3 are important for predicting Y_2, which is reasonable since both fPCs capture a variability between signatures with high and low intensities across the entire time interval. PFI identified fPC 2 as being the most important for predicting Y_3, which is reasonable since fPC 2 captures the variability between signatures with high intensity values and peaks occurring after the mean function and signatures with low intensity values and peaks occurring before the mean function. PFI also identified fPCs 1, 3, and 4 as having some importance for predicting Y_3, and these fPCs pick up on smaller amounts of variability affected by Y_3 such as the intensity in certain regions and the intensity of the fourth peak. By sharing these findings with an SME, we are able to confirm that the fPCs PFI identifies capture the type of variability in the signatures that is important to the corresponding explosive device characteristics.

5 Discussion

The use of machine learning models could provide improved prediction of explosive device characteristics based on the optical spectral-temporal signatures from explosions in practice. While

high predictive accuracy is important for this application, it is also imperative that it is possible to explain how the model makes predictions. In this paper, we demonstrate a method for explaining predictions made by neural networks with functional data inputs. In particular, the transformation of the optical spectral-temporal signatures using fPCA permits the identification of fPCs important to prediction in a neural network for an explosive device characteristic using PFI, and visualizations for interpreting the variability captured by the important fPCs allows for the determination of the aspects of the signatures that are important for prediction. The validation from the SME of the meaningfulness of the fPCs identified by PFI allows us to be confident that the neural networks are using trustworthy aspects of the signatures to make predictions.

A limitation of this method is that the ability to explain a prediction made by the neural network is dependent on the ability to interpret the fPCs. In our example, PFI identifies the first four fPCs as important for predicting at least one of the characteristics, and it is possible to determine meaningful variation captured by these fPCs. However, if PFI identifies fPCs that are not able to be interpreted, it would not be possible to explain the aspects of a function that are important to the neural network for prediction. The data in this paper are simulated and possess less variability than is likely to be observed with real data. With noisier data, it is likely that more fPCs would be needed to explain a large amount of the variability in the data, which could lead to higher numbered fPCs being identified as important by PFI. These higher numbered fPCs may be more difficult to interpret.

Another aspect not considered in this paper is that fPCA accounts for amplitude variability (vertical variability) but does not account for phase variability (horizontal variability) in the functions. Joint functional principal component analysis (joint fPCA) is a method that can be applied after smoothing and aligning functional data that accounts for both amplitude and phase variability [Lee and Jung, 2016, Tucker et al., 2013]. It would be possible to adjust the procedure in this paper by substituting fPCA with smoothing, aligning, and applying joint fPCA to the signatures (Figure 6). With noisier signatures, accounting for phase variability will be important to capture the signals in the data.

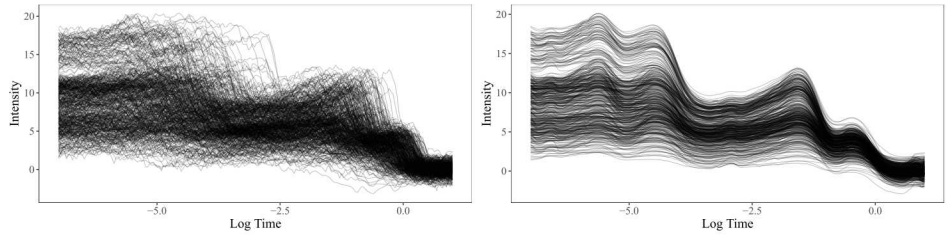

Figure 6: (Left) Examples of simulated optical spectral-temporal signatures from explosions with more variability. (Right) The signatures from the plot on the left after applying smoothing and alignment (using box filtering and time warping, repsetively, from the fdasrvf (1.9.3) R package [Tucker, 2020]).

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/a:1010933404324.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the "Rashomon" Perspective. 2018. URL https://arxiv.org/pdf/1801.01489v1.pdf.

Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *arXiv*, 2018.

Giles Hooker and Lucas Mentch. Please Stop Permuting Features: An Explanation and Alternatives. *arXiv*, 2019.

Sungwon Lee and Sungkyu Jung. Combined Analysis of Amplitude and Phase Variations in Functional Data. *arXiv*, 2016.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for Interpreting and Understanding Deep Neural Networks. 2017. doi: 10.1016/j.dsp.2017.10.011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, United States of America, 2005. ISBN 0-387-40080-X.

J. Derek Tucker. *fdasrvf: Elastic Functional Data Analysis*, 2020. URL https://CRAN.R-project.org/package=fdasrvf. R package version 1.9.3.

J. Derek Tucker, Wei Wu, and Anuj Srivastava. Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis*, 61:50–66, 2013. ISSN 0167-9473. doi: 10.1016/j.csda.2012.12.001.

Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Mueller. Review of Functional Data Analysis. 2015.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.