National Laboratories
SAND2020-11748C

Explaining Neural Networks with Functional Data Using PCA and Feature Importance

PRESENTED BY

Katherine Goode, Daniel Ries, and Joshua Zollweg

Presented at AAAI FSS-20: Artificial Intelligence in Government and Public Sector
November 14, 2020

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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-NA-0003525. SAND NO. XXX

Presentation Structure

1. Motivating national security example and explainability
2. Our approach
3. Application to national security example
4. Concluding thoughts

Motivating National Security Example

Objective: Identify explosive device characteristics using optical spectral-temporal signatures from videos of explosions

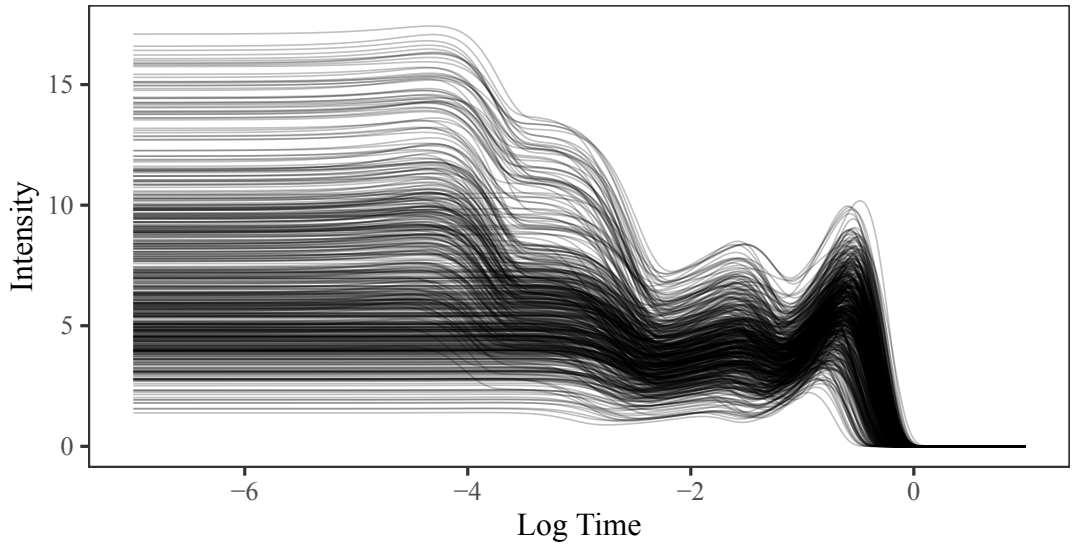

Figure: Example simulated explosion spectral-temporal signatures.

Motivating National Security Example

Current methods

- Heuristic algorithms
- Subject matter expert direct review

Applying maching learning

- Interest in using machine to improve predictions
- Important model qualities
  - Accurate predictions
  - Uncertainty quantification
  - ***Understanding of predictive process***

Explainability

Advantages of interpretability

- Understand model prediction making process
- Assess the model

Disadvantage of machine learning

- Predictive ability of machine learning models make them desirable tools
- Often comes at cost of interpretability

Explaining black-box models

- Alternative way to explain non-interpretable model predictions
- Important with sensitive applications (e.g., national security)

Explainability

Functional Data

- Each observations is a function
- Spectral-temporal signatures are an example
- Collection of functional data easy with modern technology

Functional Data and Explainability

- Many explainability methods proposed [1; 2; 3; 4; 5]
- Methods not focused on functional data
- Would like to account for the functional nature of the data

Explainability

A naive approach for explainability with functional data

- Use each time point as a feature in the model to train a model
- Compute feature importance
- Does not account for structure and correlation in functional data

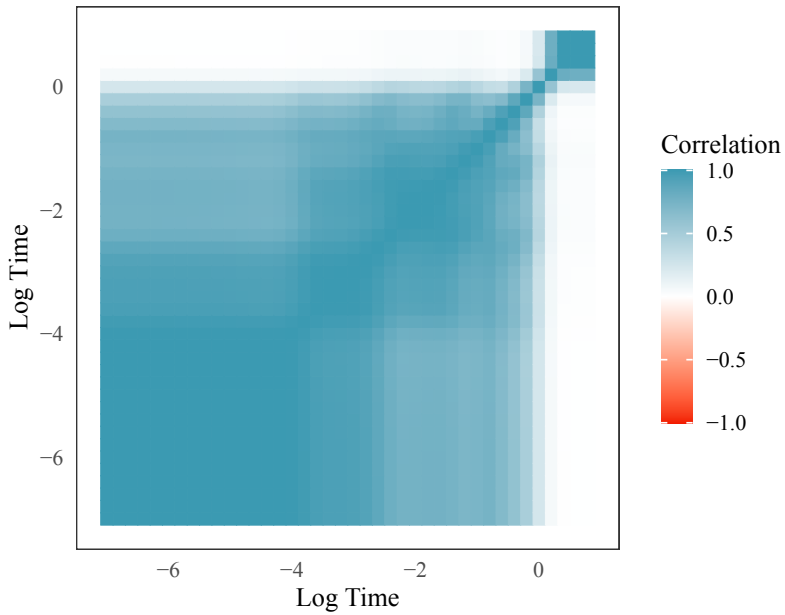

Figure: Correlation between every 25th time in simulated spectral-temporal signature data.

Our Approach

Overview

Combine techniques of

- **functional principal component analysis** (fPCA) [6; 7] and
- **permutation feature importance** (PFI) [8]

in conjunction with

- **visualizations of functional principal components**

to train and explain a model in a way that ***accounts for functional nature of data***

Our Approach

Background on fPCA

- fPCA is essentially PCA with functional data
- Transforms original functions in a way that provides nice properties:
  - Independent features
  - First few features capture majority of variation in original data
- Eigenvectors are now ”eigenfunctions”

Background on PFI

- Originally developed for random forests [9] and generalized by Fisher, Rudin, and Dominici [8]
- Procedure:
  - Permute a feature
  - Determine how model predictions are affected
  - Repeat for all other features
  - Repeat to account for random variation
- Features that decrease model performance when permuted are considered important
- Biased when features are correlated [10; 11; 12]

Our Approach

Procedure

1. Transform signatures using fPCA
   - Removes correlation between features and captures functional aspect of data
2. Train machine learning model using fPCs
3. Apply PFI to identify important fPCs
   - No concern of bias in PFI due to correlation
   - PFI applicable to any predictive model
4. Visualize and interpret important fPCs
   - Interpret variability explained
   - Identify functional characteristics important for prediction

Application to National Security Example

Simulated data

- SMEs simulated 10,000 signatures with 1,000 time points
- 3 explosive device characteristics
- Randomly divided into training (72.25%), testing (15%), and validation (12.75%) sets

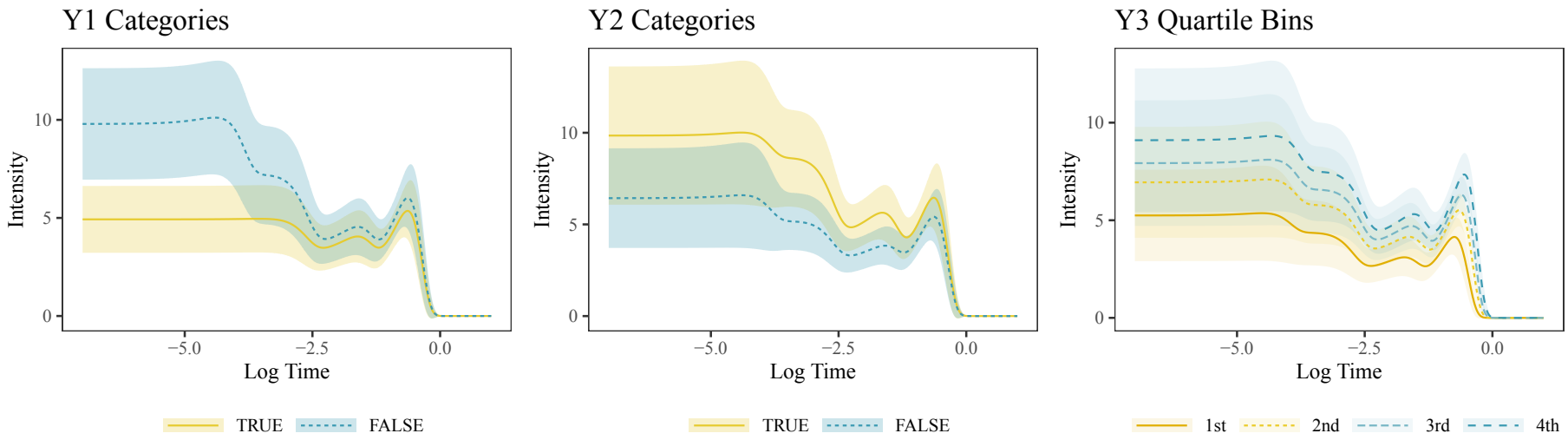

Figure: Pointwise functional means and standard deviations of explosive device characteristics.

Application to National Security Example

Step 1: Transform signatures using fPCA

- fPCA applied to convert 1,000 features to 1,000 fPCs
- Eigenfunctions used to transform testing and validation data sets to fPCs

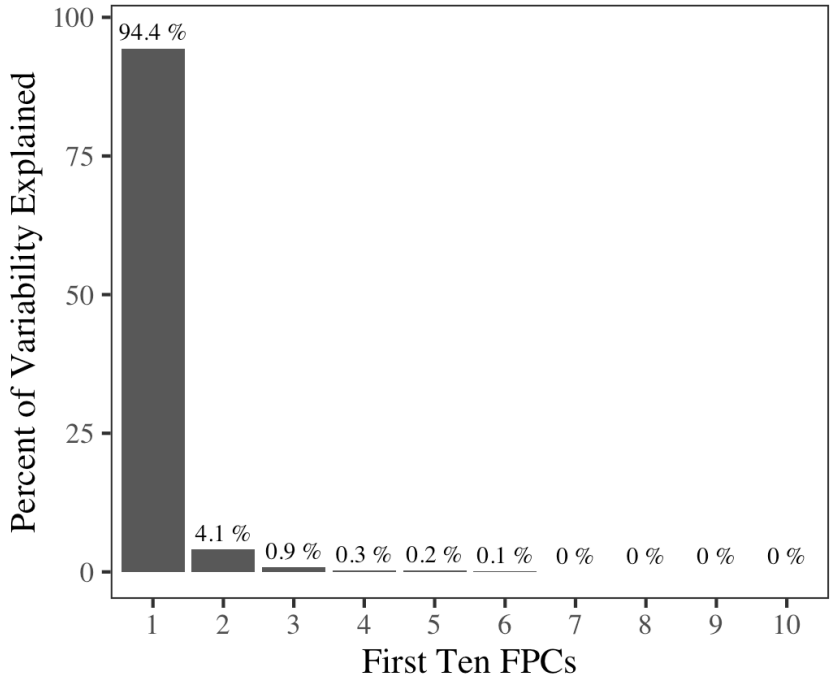

Figure: Percent variation explained by first 10 fPCs.

Application to National Security Example

Step 2: Train machine learning model using fPCs

- Neural network trained for each explosive device characteristic
- All 1,000 fPCs used as features
- Model structure: 3 layers with 50, 40, and 30 nodes, respectively

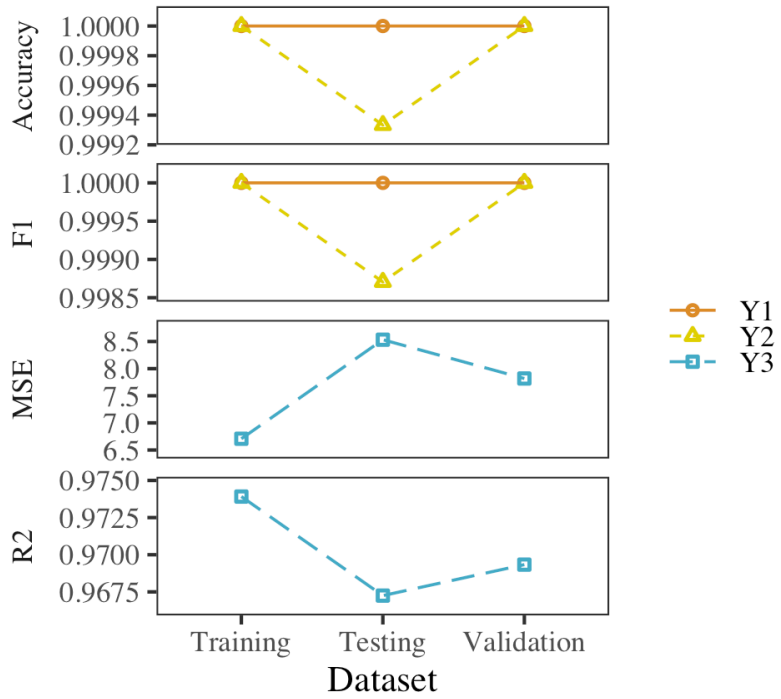

Figure: Model performance metrics.

Application to National Security Example

Step 3: Apply PFI to identify important fPCs

- PFI applied to trained networks
- 10 replications used to account for random permutation variability

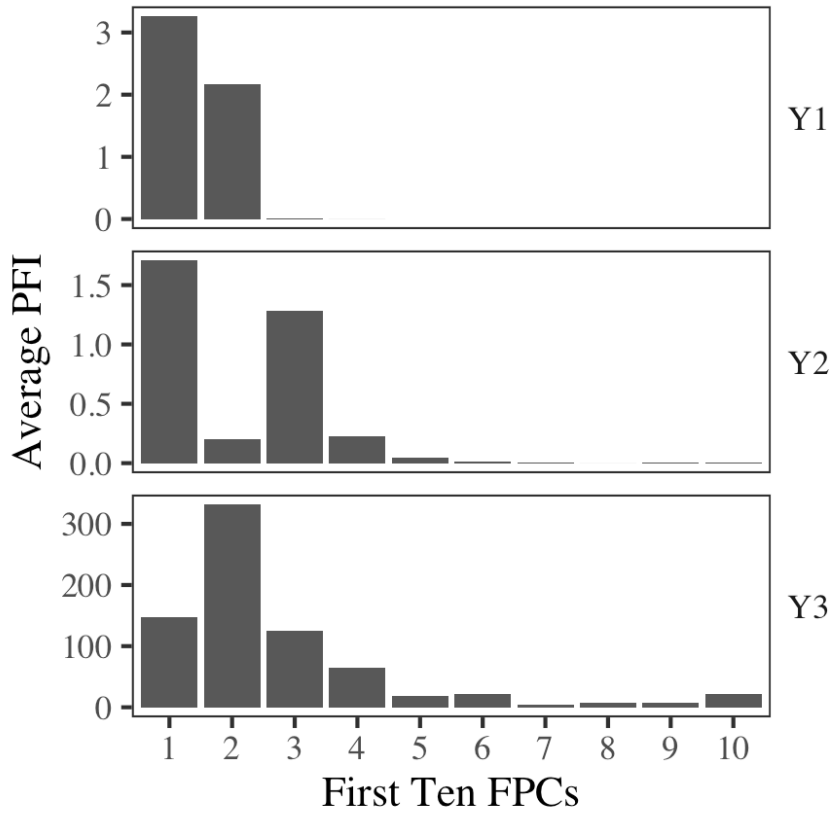

Figure: PFI for first 10 fPCS for each explosive device characteristic neural network.

Application to National Security Example

fPCs important for discrimination identified by PFI

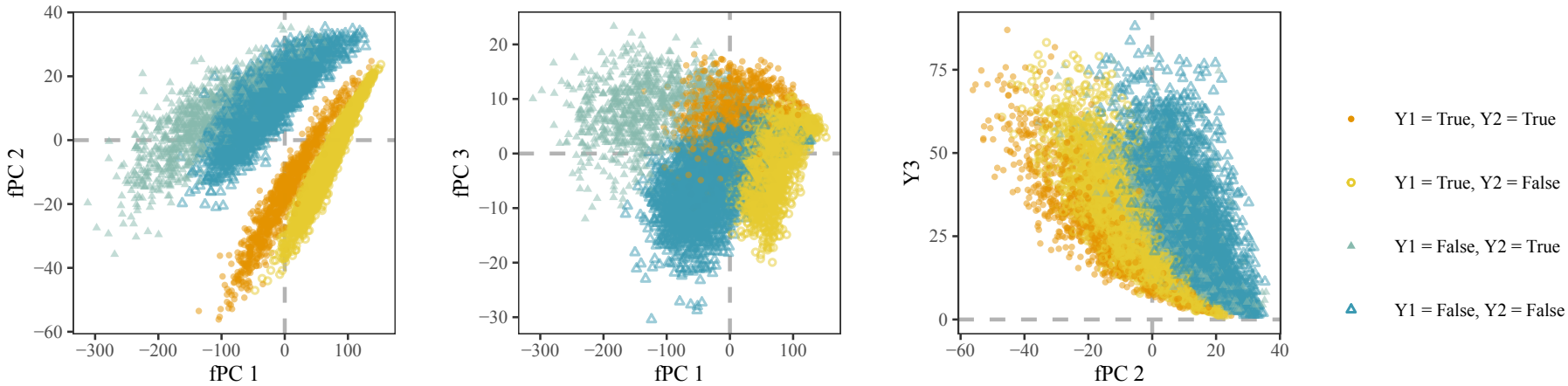

Figure: Relationships between device characteristic levels and important fPCs.

Application to National Security Example

Step 4: Visualize and interpret important fPCs

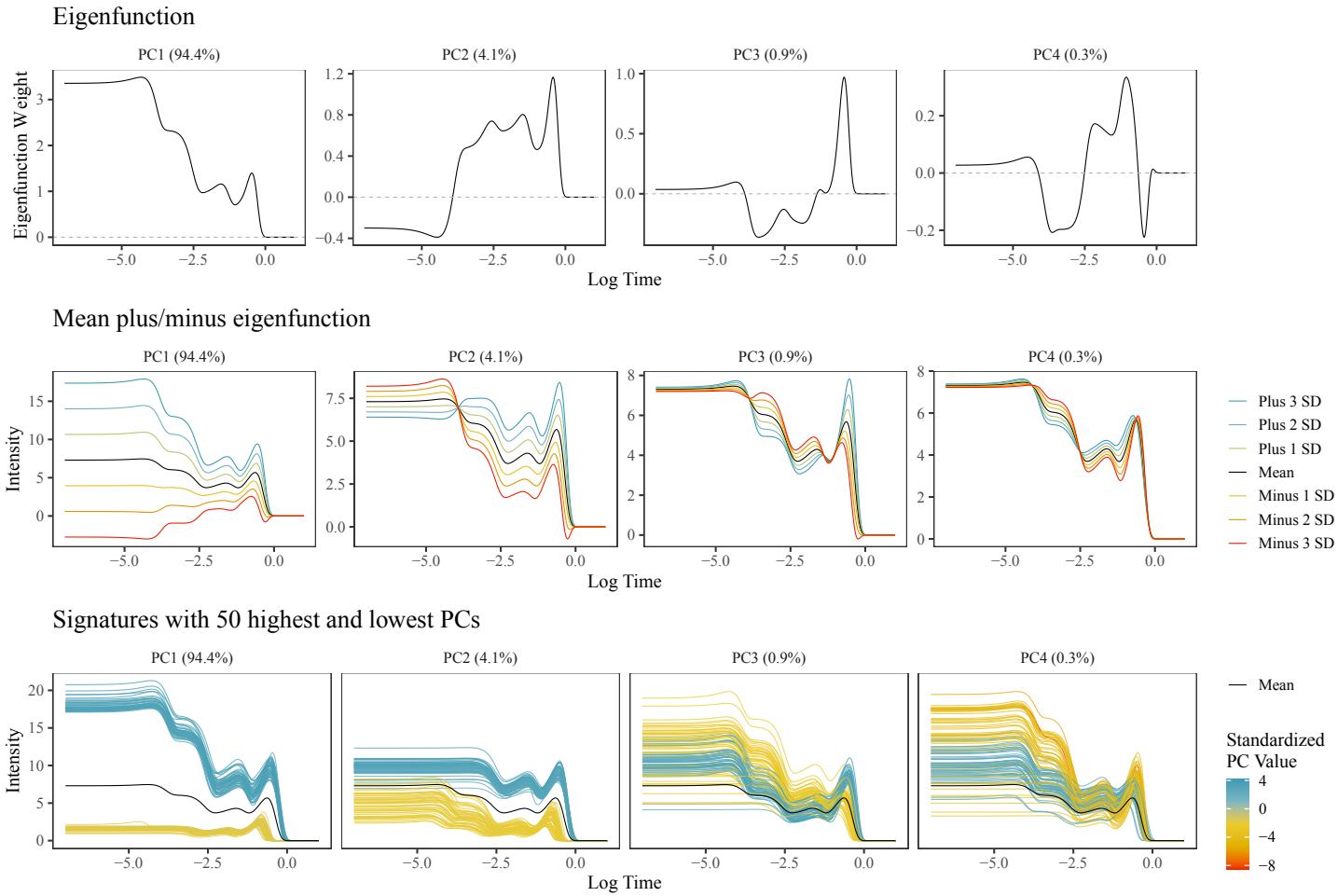

Figure: Visualizations of important fPCS for interpretation.

Concluding Thoughts

Advantages

- Approach provides insight into model predictions while accounting for nature of functional data
- Able to share findings with an SME who confirmed
- Information about functional characteristics important for prediction could be distilled and shared with decision makers

Limitation

- Difficult to interpret higher numbered fPCs

Concluding Thoughts

Future work

- Change to using joint fPCA [13; 14]
  - Accounts for horizontal and vertical variability of functions
- Application to non-simulated data
- Adjust PFI to account for uncertainty
  - For example, what if model is a Bayesian neural network?

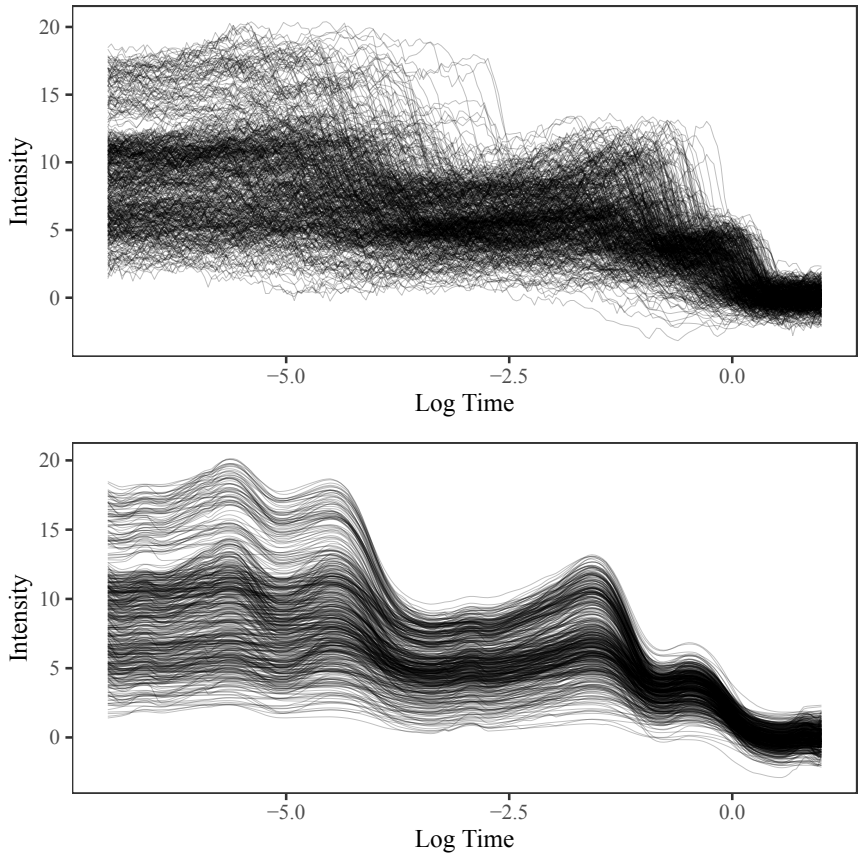

Figure: Simulated optical spectral-temporal signatures from explosions with more variability (top) and signatures after smoothing and alignment (bottom).

References

[1] L. H. Gilpin, D. Bau, B. Z. Yuan, et al. “Explaining Explanations: An Overview of Interpretability of Machine Learning”. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2018, pp. 80-89. ISBN: 9781538650912. DOI: 10.1109/dsaa.2018.00018. URL: https://ieeexplore.ieee.org/document/8631448.

[2] R. Guidotti, A. Monreale, S. Ruggieri, et al. “A Survey of Methods for Explaining Black Box Models”. In: *ACM Comput. Surv.* 51.5 (Aug. 2018). ISSN: 0360-0300. DOI: 10.1145/3236009. URL: https://doi.org/10.1145/3236009.

[3] F. M. Hohman, M. Kahng, R. Pienta, et al. “Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers”. In: *IEEE transactions on visualization and computer graphics* (Jun. 2018). ISSN: 1077-2626. DOI: 10.1109/tvcg.2018.2843369. URL: https://europepmc.org/articles/PMC6703958.

[4] C. Molnar. *Interpretable Machine Learning*. lulu.com, 2019. ISBN: 0244768528. URL: https://christophm.github.io/interpretable-ml-book/.

[5] G. Montavon, W. Samek, and K. Müller. “Methods for Interpreting and Understanding Deep Neural Networks”. In: *Digital Signal Processing* 73 (2018), pp. 1-15. DOI: 10.1016/j.dsp.2017.10.011. eprint: 1706.07979. URL: https://doi.org/10.1016/j.dsp.2017.10.011.

References

[6] J. Ramsay and B. Silverman. *Functional Data Analysis*. United States of America: Springer, 2005. ISBN: 0-387-40080-X.

[7] J. Wang, J. Chiou, and H. Müller. "Functional Data Analysis". In: *Annual Review of Statistics and Its Application* 3.1 (2016), pp. 257-295. DOI: 10.1146/annurev-statistics-041715-033624. URL: https://doi.org/10.1146/annurev-statistics-041715-033624.

[8] A. Fisher, C. Rudin, and F. Dominici. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously". In: *Journal of Machine Learning Research* 20.177 (2019), pp. 1-81. URL: http://jmlr.org/papers/v20/18-760.html.

[9] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5-32. ISSN: 0885-6125. DOI: 10.1023/a:1010933404324.

[10] G. Hooker and L. Mentch. "Please Stop Permuting Features: An Explanation and Alternatives". In: *arXiv preprint* (2019). arXiv: 1905.03151 [stat.ME].

References

[11] K. K. Nicodemus, J. D. Malley, C. Strobl, et al. “The behaviour of random forest permutation-based variable importance measures under predictor correlation”. In: *BMC Bioinformatics* 11.1 (2010), p. 110. DOI: 10.1186/1471-2105-11-110.

[12] C. Strobl, A. Boulesteix, A. Zeileis, et al. “Bias in random forest variable importance measures: Illustrations, sources and a solution”. In: *BMC Bioinformatics* 8.1 (2007), p. 25. DOI: 10.1186/1471-2105-8-25.

[13] S. Lee and S. Jung. “Combined Analysis of Amplitude and Phase Variations in Functional Data”. In: *arXiv* (2017). eprint: 1603.01775.

[14] J. D. Tucker, W. Wu, and A. Srivastava. “Generative models for functional data using phase and amplitude separation”. In: *Computational Statistics & Data Analysis* 61 (2013), pp. 50-66. ISSN: 0167-9473. DOI: 10.1016/j.csda.2012.12.001.

Thank you!