SAND2021-13774C

Sandia National Laboratories

Exceptional service in the national interest

Machine Learning Explainability for Functional Data

Gaining Insight into Black-Box Models with Functional Data as Inputs

Katherine Goode, PhD

Postdoc in Statistical Sciences Group (SNL Org 5573)

UC Davis/Sandia Symposium

November 3, 2021

SAND XXX

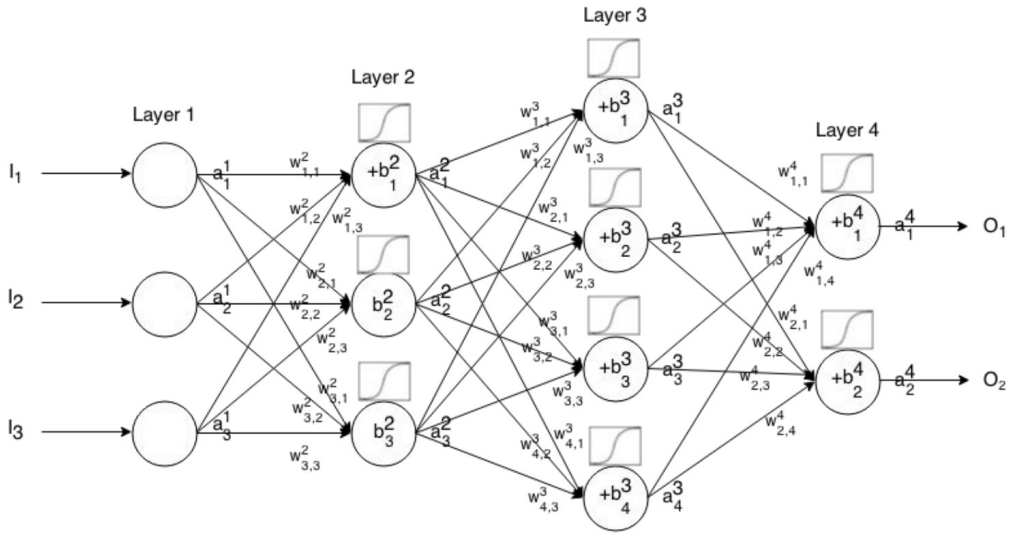
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

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.

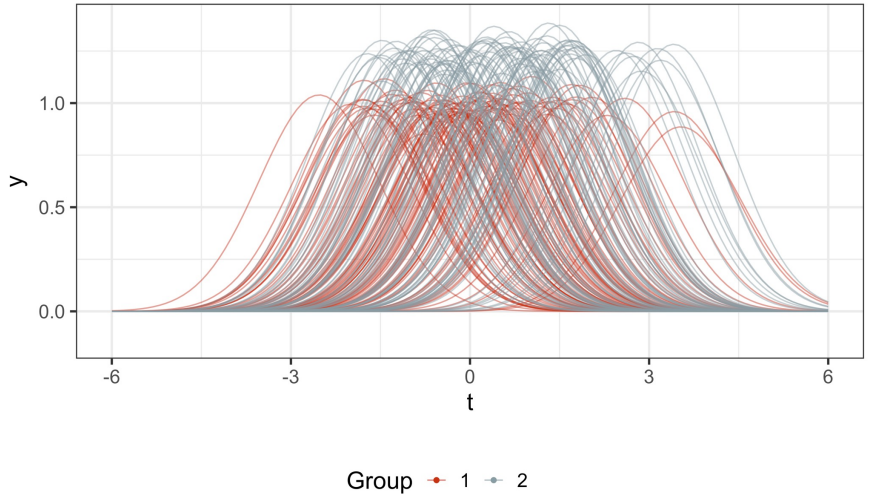
Background

Explainability with Machine Learning Models

- Non-interpretable ("black-box") models may provide great predictions, but in the context of the application, how to…
  - **Understand and explain** the predictions?
  - **Motivate and assess** the model?
- Explosion of research in "**explainability**"
  - Try to gain insight into black-box models
- Explainability **methods**
  - Feature importance
  - Surrogate models
  - Visualization techniques
  - Interpretable machine learning models
  - Etc.



Explainability with Machine Learning Models with *Functional Data Inputs*



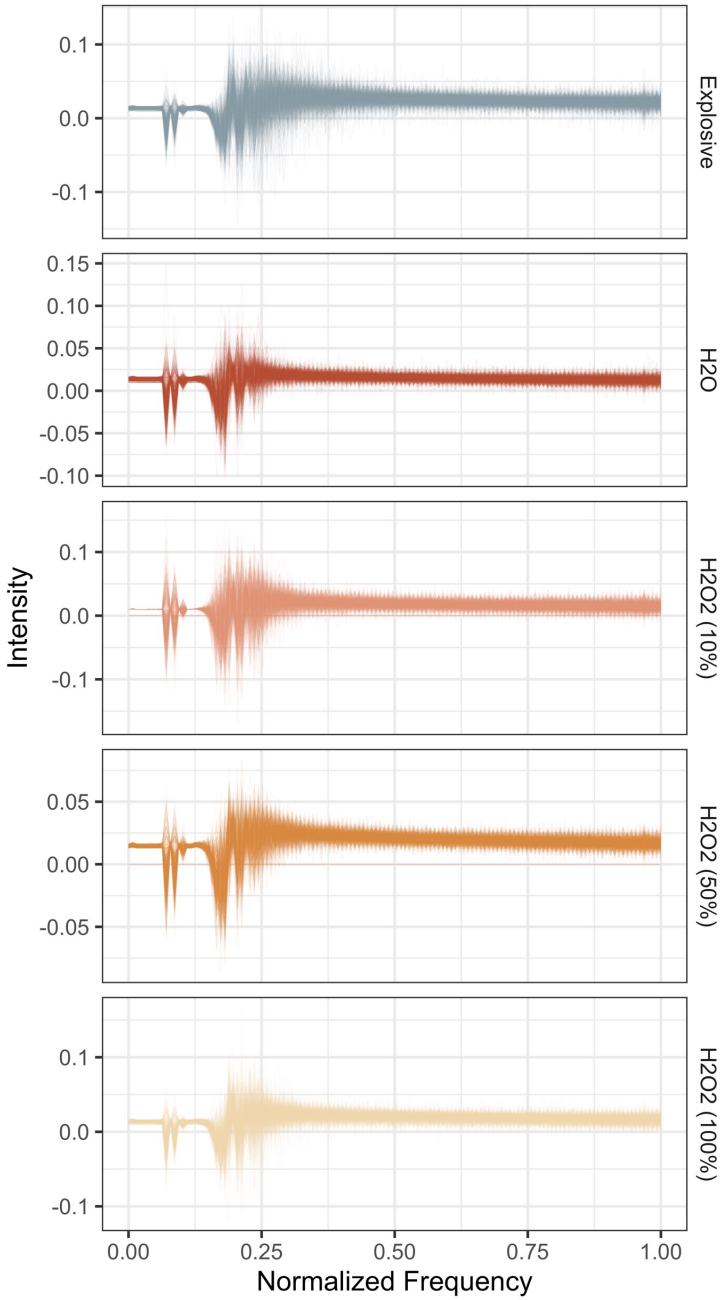
- **Functional data**
  - Each observation is a curve
  - National security example: H-CT scans of materials
- **Cross-sectional approach**
  - Each time/frequency/etc. is a feature in model and existing explainability methods applied
  - Does NOT account for correlation
  - Correlation shown to negatively affect explainability methods
- **Functional data approaches**
  - Little previous research

My Research at Sandia

Joint work with Dr. Daniel Ries (SNL), Dr. J. Derek Tucker (SNL),
and Dr. Heike Hofmann (Iowa State University)

Research Objectives

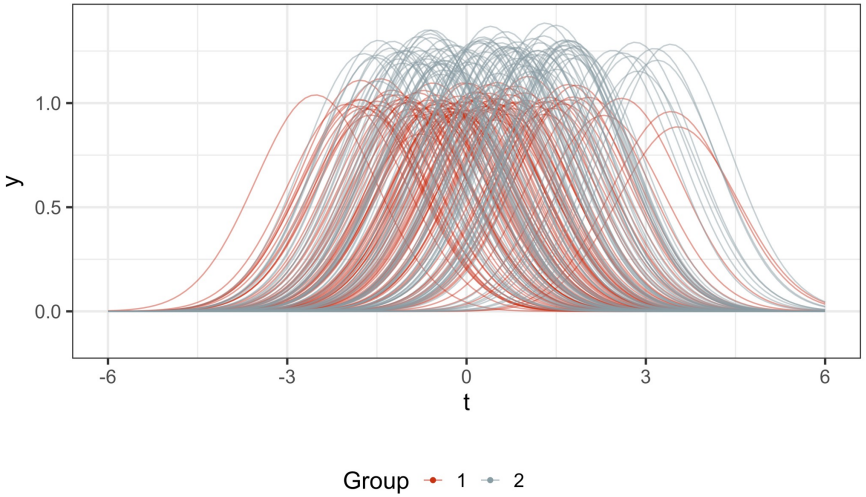
- **Gain insight into black-box models with functional data inputs**
  - Understand how predictions are made
  - Assess the model
  - Motivate model to scientists and decision makers
- **Motivated by prediction applications with functional data at SNL**
  - Example: H-CT scans of materials
  - Explainability important with national security examples



My Previous Work: VEESA Pipeline

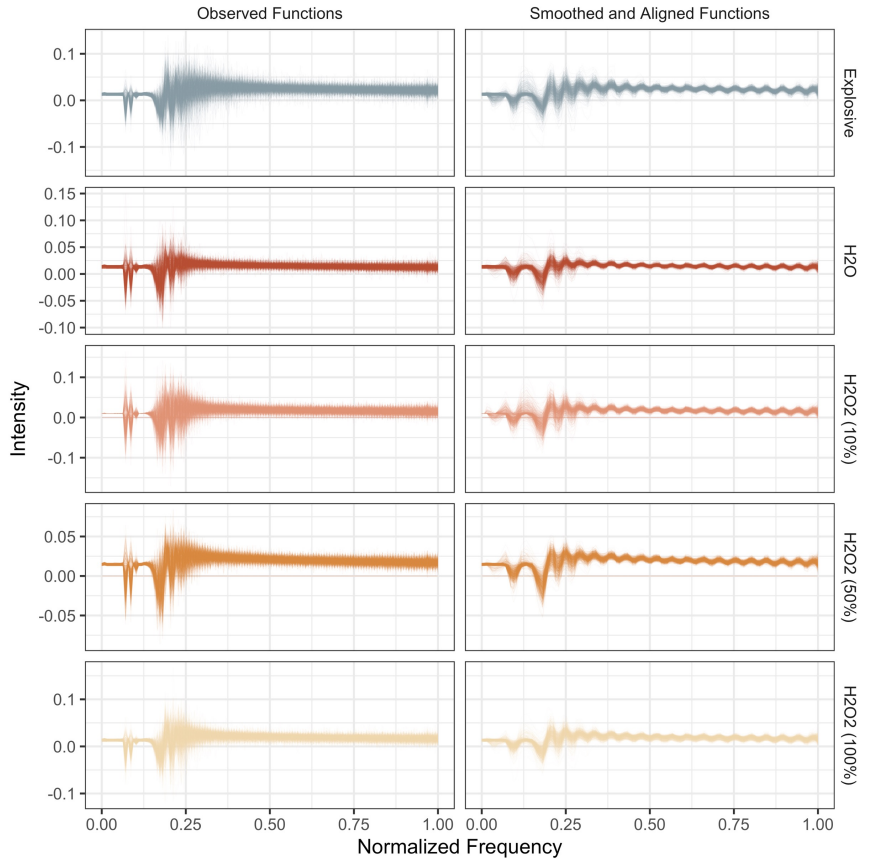
- **V**ariable importance **E**xplainable **E**lastic **S**hape **A**nalysis Pipeline
  - Accounts for **functional nature** of data
  - Incorporates explainability through **variable importance** (without bias from correlation)
  - **Model-agnostic**

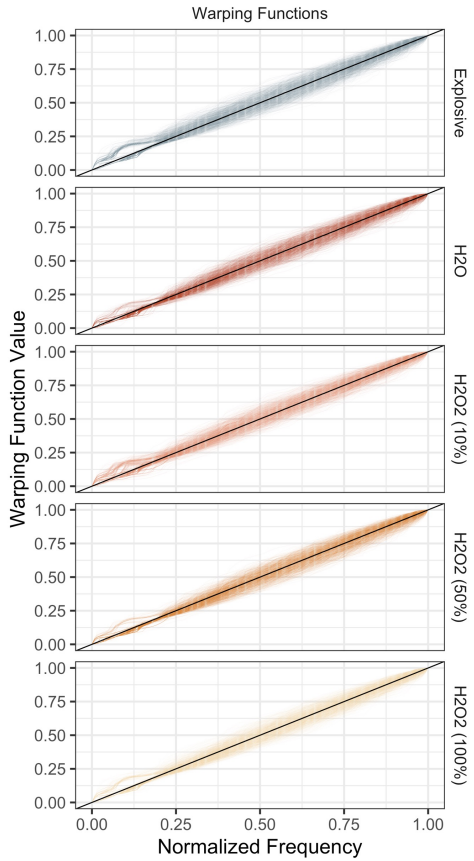
- Makes use of two previously developed techniques:
  - **Elastic shape analysis (ESA)** framework for functional data
  - **Permutation feature importance (PFI)**

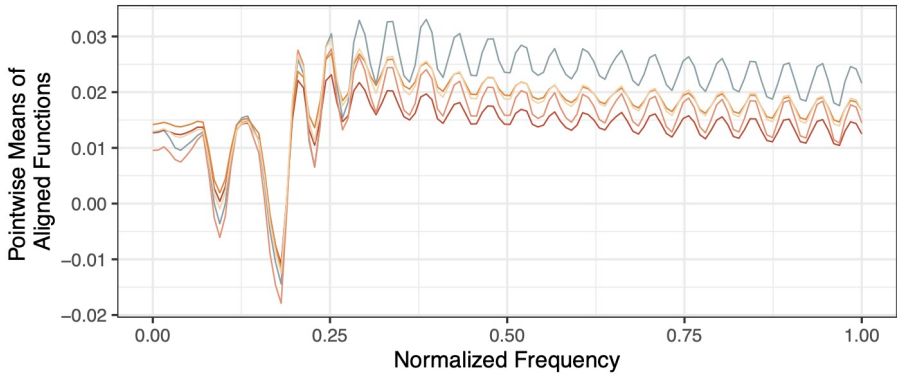


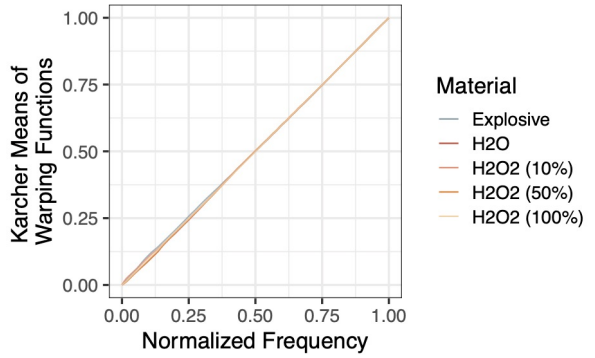
VEESA Pipeline

- **Pre-processing** (elastic shape analysis techniques)
  - Separate functional variability into vertical and horizontal parts
  - Obtain joint functional principal components (jfPCs)
- **Modeling**
  - Use jfPCs as features in machine learning model



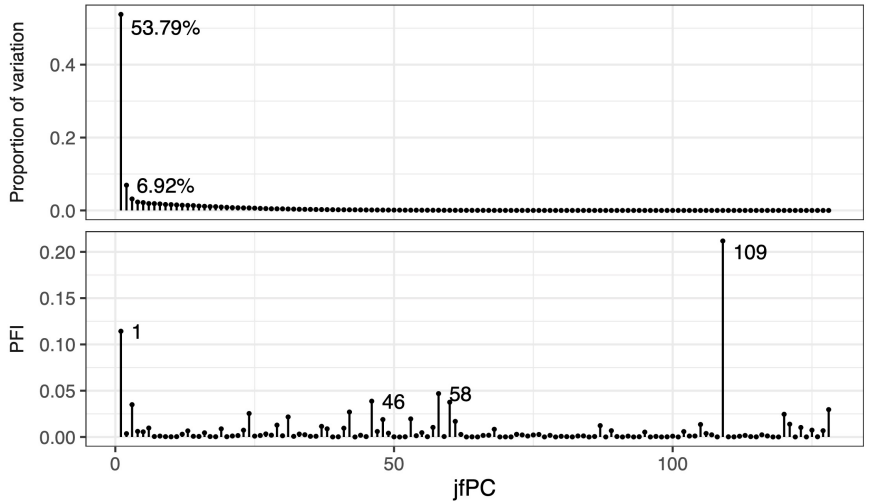


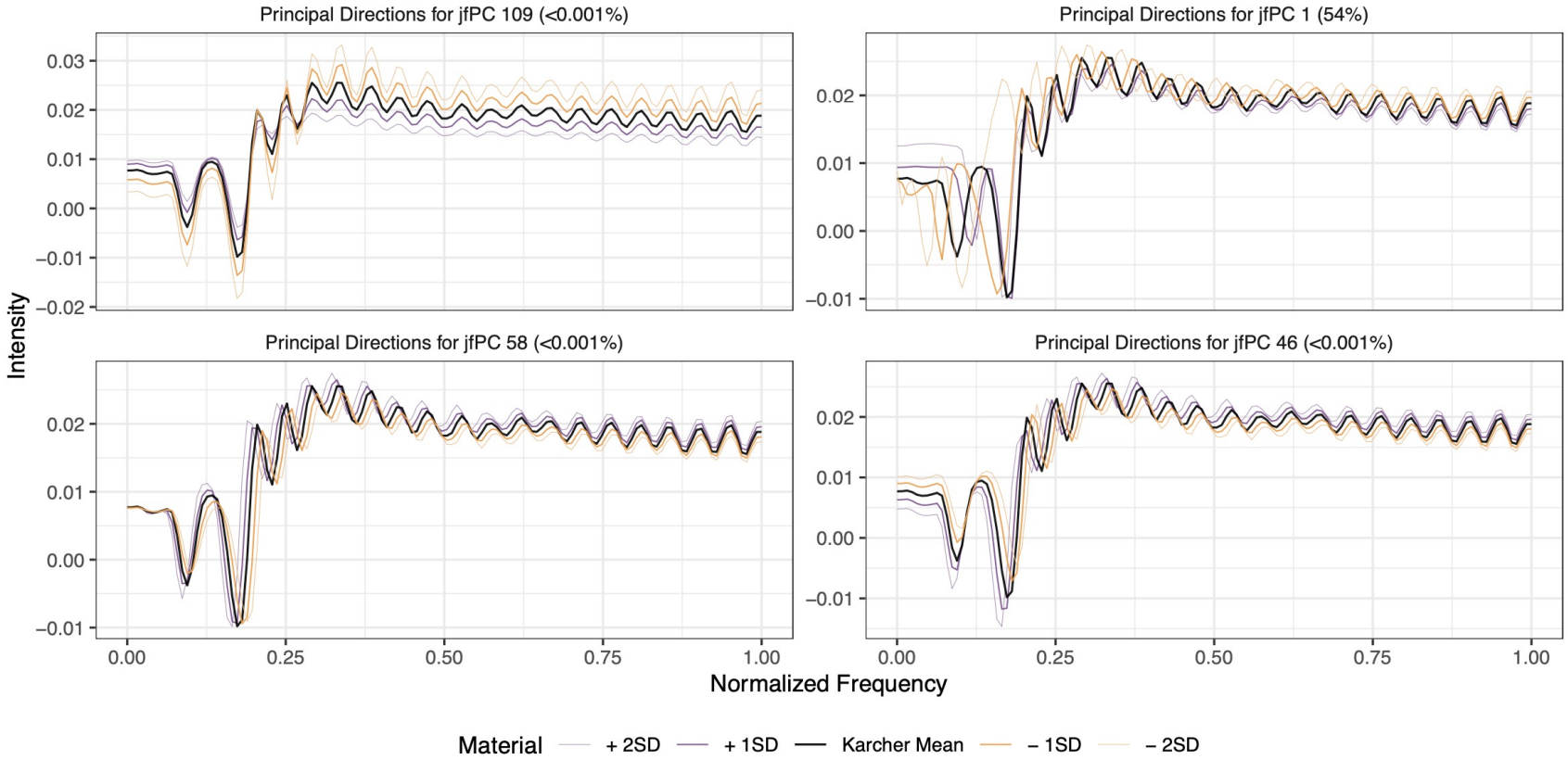




VEESA Pipeline (continued)

- **Post-processing**
  - Apply PFI and interpret important jfPCs using visualization





Going Forward

Challenges and Limitations

- How to best **assess explanations**?
  - Models are black-boxes, so we don't know the "true" explanations
  - Possible route: Develop index for high/low explainability based on number of PCs needed to capture model "mechanism"
- How to implement **PFI with large datasets**?
- How to best perform **feature selection based on PFI**?
- How to best **interpret jfPCs when many are "important"**?

Ideas for Future Work

- Check for important **interactions** (not only main-effects)
- Try incorporating **other explainability methods** such as:
  - Shapley values
  - Partial dependence plots
- Explore **other avenues** for modeling functional data with machine learning models in ways that allow for explainability

Questions?

Contact Information:

Katherine Goode
kjgoode@sandia.gov
505-844-1998