INTRODUCTION TO FUNCTIONAL DATA ANALYSIS, PART 2: FUNCTIONAL PCA AND FUNCTIONAL REGRESSION

MATTHEW MALLOURE

RStudio Community Meetup

August 16th, 2022

Matthew Malloure Biography

- Education:
  - Grand Valley State University
    - ✓ 2010 – BS in Statistics
    - ✓ 2012 – MS in Biostatistics
  - Texas A&M University
    - ✓ 2017 – PhD in Statistics
    - ✓ Research: Bayesian Nonparametric Goodness-of-Fit Testing Using Cross-Validation Bayes Factors (Advisor Jeffrey D. Hart)
    - ✓ Hart, J.D. and Malloure, M. (2019) “Prior-free Bayes Factors Based on Data Splitting”. *International Statistical Review. 87 (2) 419-442*
- Dow Experience:
  - 2017-2022: Statistician in Core R&D
  - 2022-Present: Data Scientist in Packaging & Specialty Plastics R&D
  - Specialty Areas Include:
    - ✓ Functional Data Analysis
    - ✓ Nonparametric Statistics
    - ✓ Statistical Computing/Simulation
  - Primary Project Areas:
    - ✓ Polymer Resin Design
    - ✓ Development of Sustainable Products

What Is Functional Data Analysis (FDA)?

Functional Data Analysis (FDA)

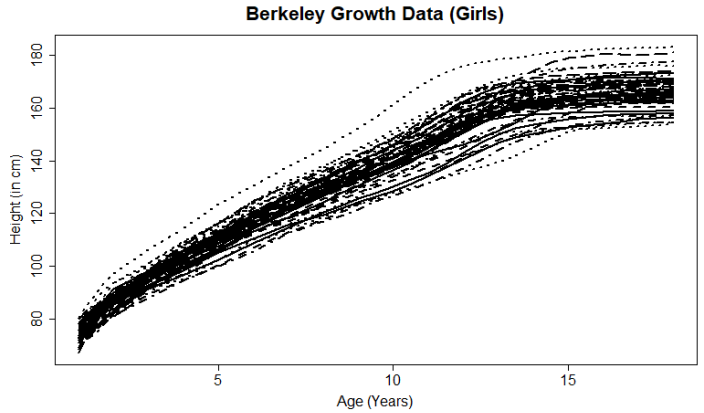
Variable Types:

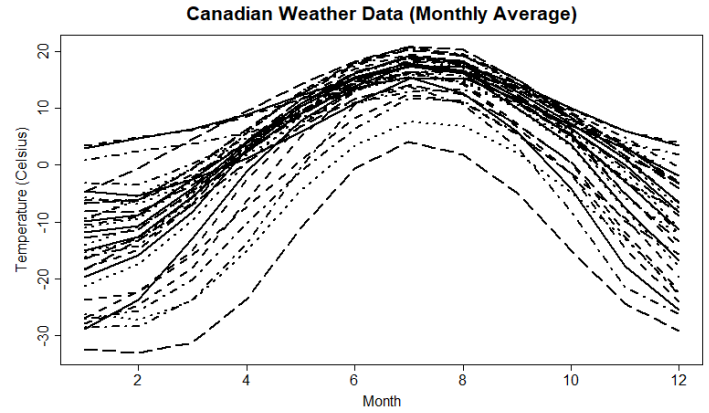
- **Continuous Variable** → single scalar value
- **Categorical Variable** → values consist of strings representing labels
- **Functional Variable** → an entire process, distribution, or curve.

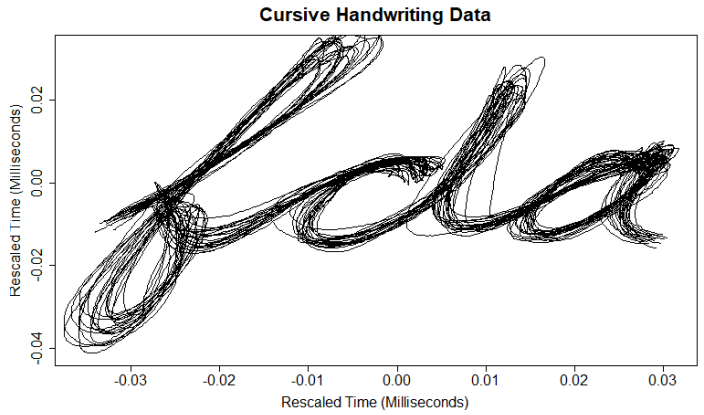
Functional Data Analysis is a class of methods that allow us to incorporate entire functional variables in an analysis!

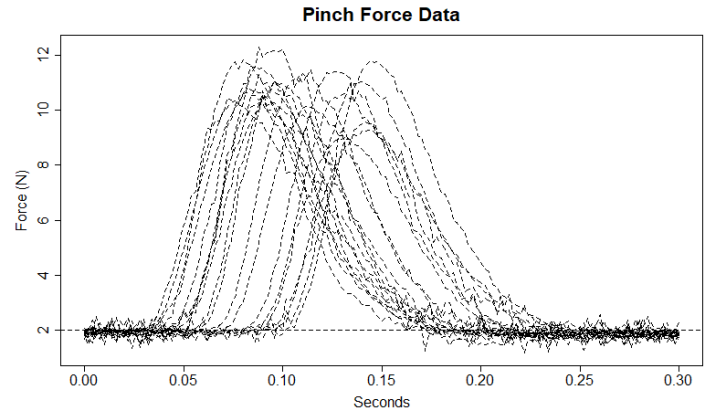
For a sample of functional data, we can perform functional analogues to standard statistical methods:

- Curve Estimation (Smoothing)
- Exploratory Data Analysis – Mean & Variance
- Principal Components Analysis
- (Generalized) Linear Models
- Functional Experimental Design
- Analysis of Derivative Functions







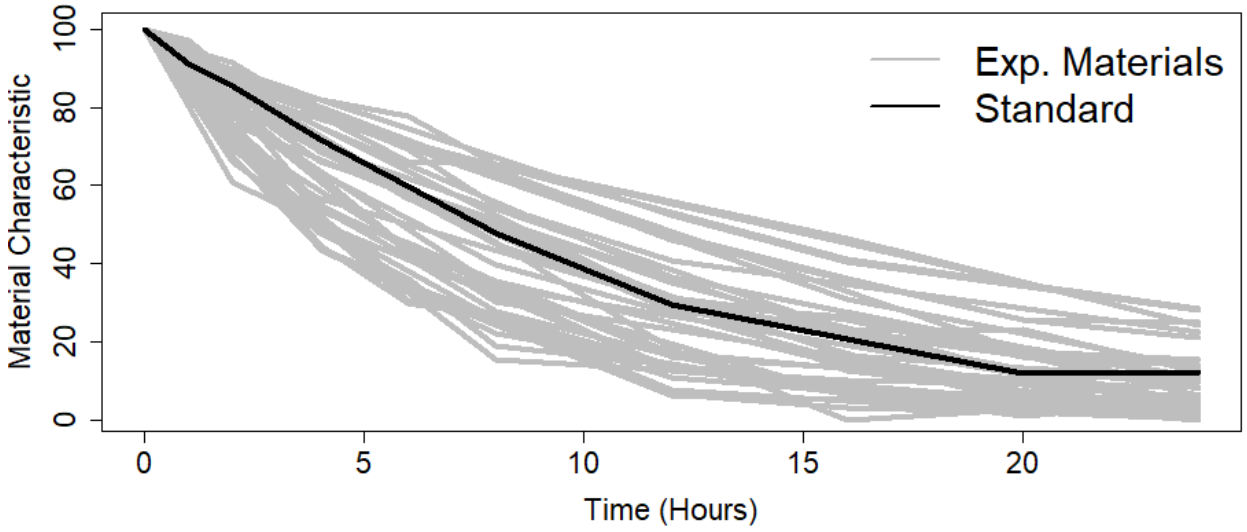


Data available in 'fda' package and referenced in *Ramsay, James O., and Silverman, Bernard W. (2006), Functional Data Analysis, 2nd ed., Springer, New York*

Why Do We Need FDA?

Case Study Data: A material characteristic is measured over time for 41 different additives

- Automatic data capture at scheduled time points
  - t=1, 2, 4, 6, 8, 12, 16, 20, and 24 hours
- Due to the rigorous data processing steps at each measurement, missing data is quite common.
  - Curves can have between 2 and 5 missing points out of 10
- Data simulated for sharing along with working code.



Questions: Using the additive data,

1. Can we identify a subset of 3-5 experimental materials with most similar characteristic profiles to the standard?
2. Can we predict fabricated article performance from the material characteristic profiles?

What traditional methods might we use to answer these questions?

- Time Series / Repeated Measures / Longitudinal Data Analysis
- Moment-Based Regression Models
- Principal Components Analysis / Partial Least Squares

Each of these methods comes with some limitation(s):

- Measurements may not be equally spaced
- Differing number of measurements in each function
- Information loss when summarizing functions using moments
- Analysis of data matrix ignores functional relationship between columns

FDA allows us to answer the posed questions without these limitations!

Initial Step in FDA Applications: Smoothing

Converting Measured Functional Data to Smooth Functions

Smoothing Functional Observations

For a single observation of a functional variable, we measure the true function/process $x(t)$ at a finite set of domain points $t_1, t_2, \ldots, t_N \in T$ often with some measurement error ϵ_j

$$y_j = x(t_j) + \epsilon_j$$

The initial smoothing step converts the measured, discrete data vector to a smooth, continuous function across the entire domain

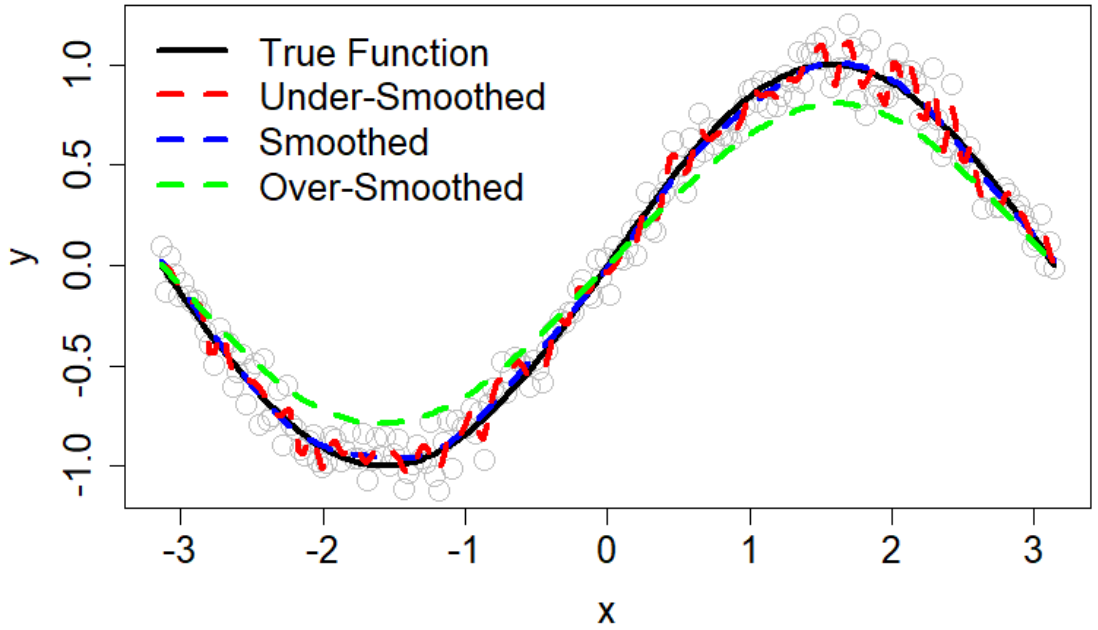
The most common method utilizes penalized splines:

- Fourier series for periodic data
- Cubic B-splines for non-periodic data
- Fine tune the smoothing parameter λ to balance bias/variance
- Smoothing can be done for functional observations individually or all at once

Note we can use smoothing to convert discrete data with unequally spaced and/or unequal number of measurements to a suitable matrix for times series, repeated measures, etc.

True Function: $y_j = \sin(t_j) + N(0,0.1), t_j \in [-3, 3]$

Estimated Functions: 100 Fourier Basis Functions with $\lambda = 1e^{-12}, 1e^{-3}, 1e^{-1}$



Once the data are smoothed, we can implement FDA methods!

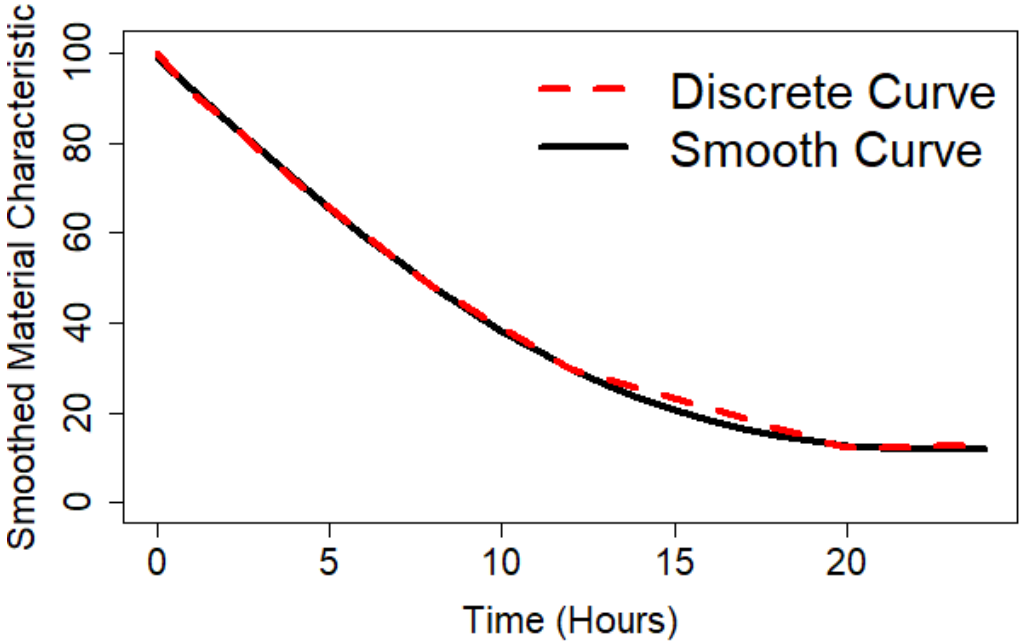
Additive Screening Case Study: Data Smoothing

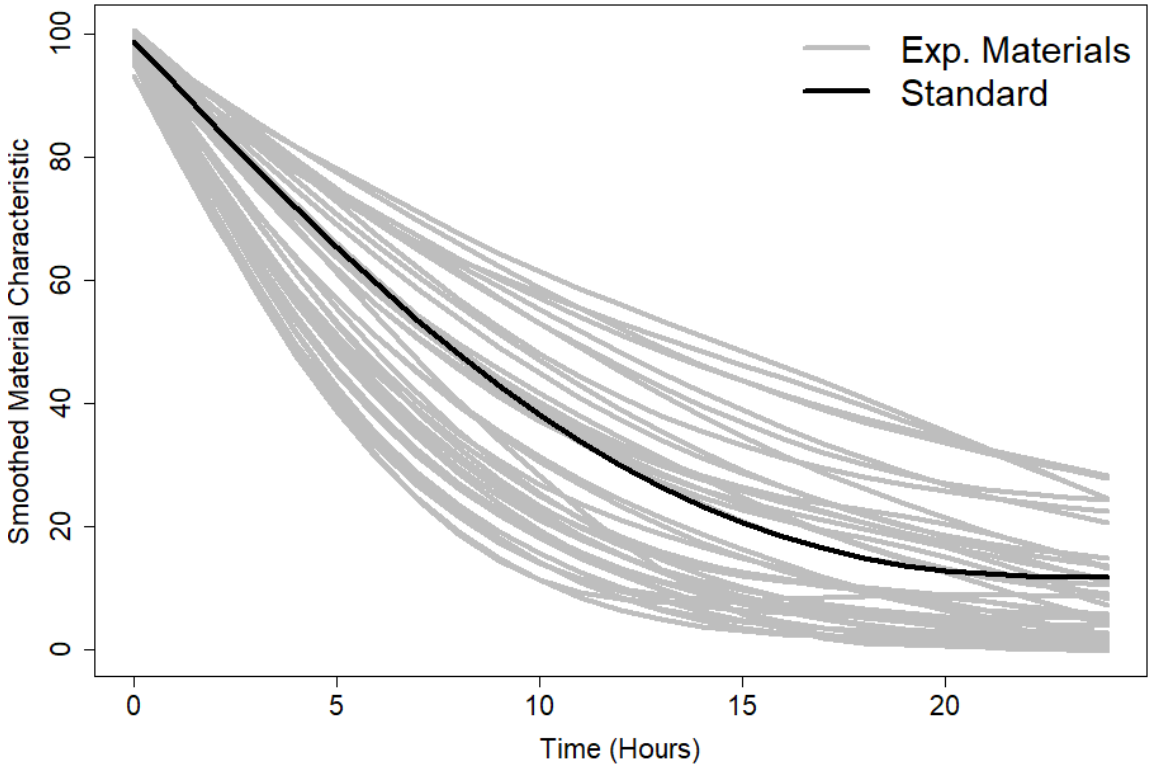
Step 1: Smooth profiles individually using:
- 20 cubic B-splines
- Smoothing parameter $\lambda = 10$

Step 2: Evaluate the estimated functional observation at $t = 0,1,2,\ldots,23,24$ hours

Step 3: Stack all 41 smoothed & evaluated observations creating a 41 x 25 matrix

Step 4: Create a functional data object in R.





Now that we have converted the measured discrete data to estimated functional observations, how do we answer the two key questions?

Screening Additives Compared to Control

Functional Principal Components Analysis (FPCA)

FPCA: Introduction and Results

Traditional Scalar PCA:

- Data Transformation: Construct uncorrelated linear combinations from correlated data vectors
- Dimension Reduction: Select primary modes of variation that capture majority of the total variability
- Mathematics: Compute Eigenvalues/Eigenvectors of covariance (or correlation) matrix

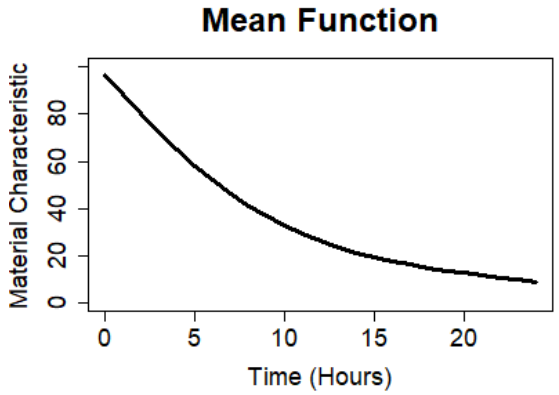
FPCA is the functional analog for PCA:

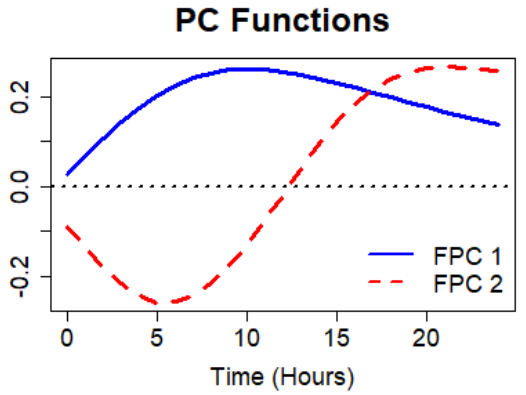
- Perform eigen analysis on functions
- Capture primary modes of variation in orthonormal functions
- We can use the eigenfunctions as empirical basis functions

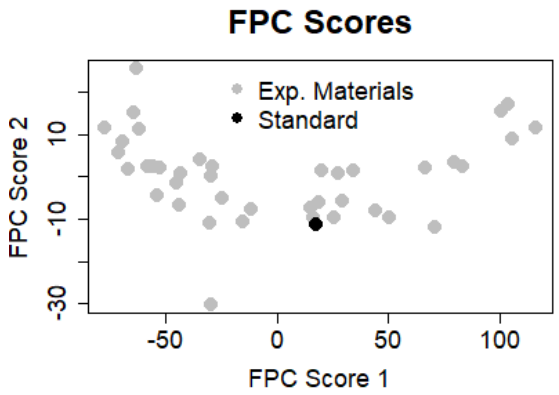
Karhunen-Loeve Expansion:

$$y(t) = \bar{y}(t) + \sum_{j=1}^{J} c_j \xi_j(t)$$

FPCA Results for Performance Decay Curves

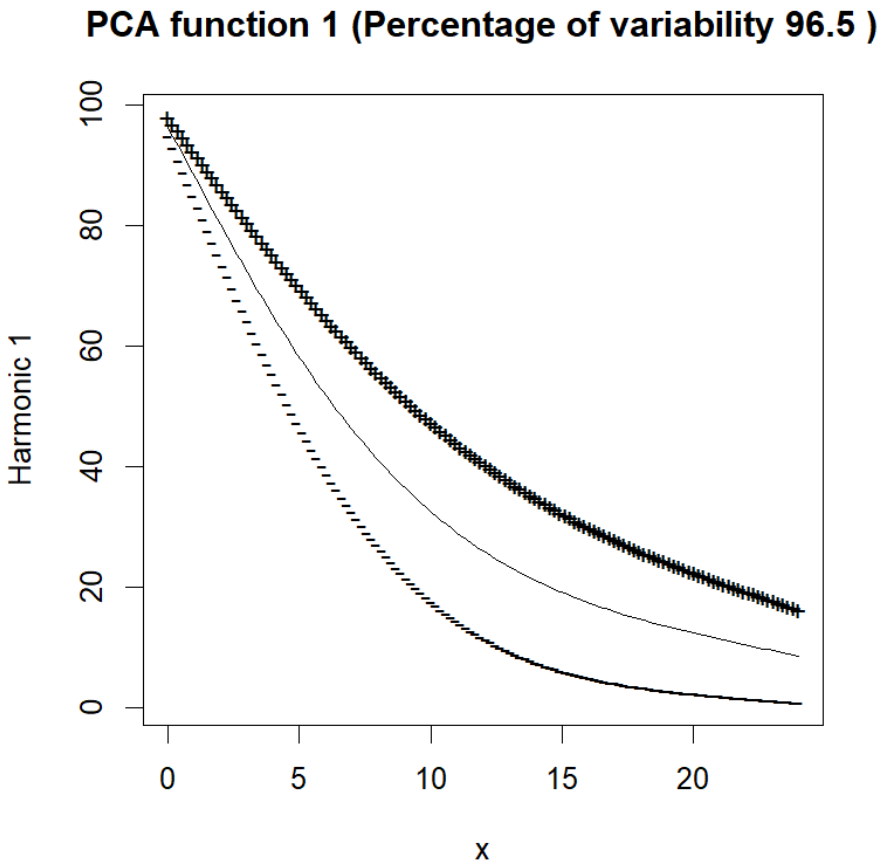


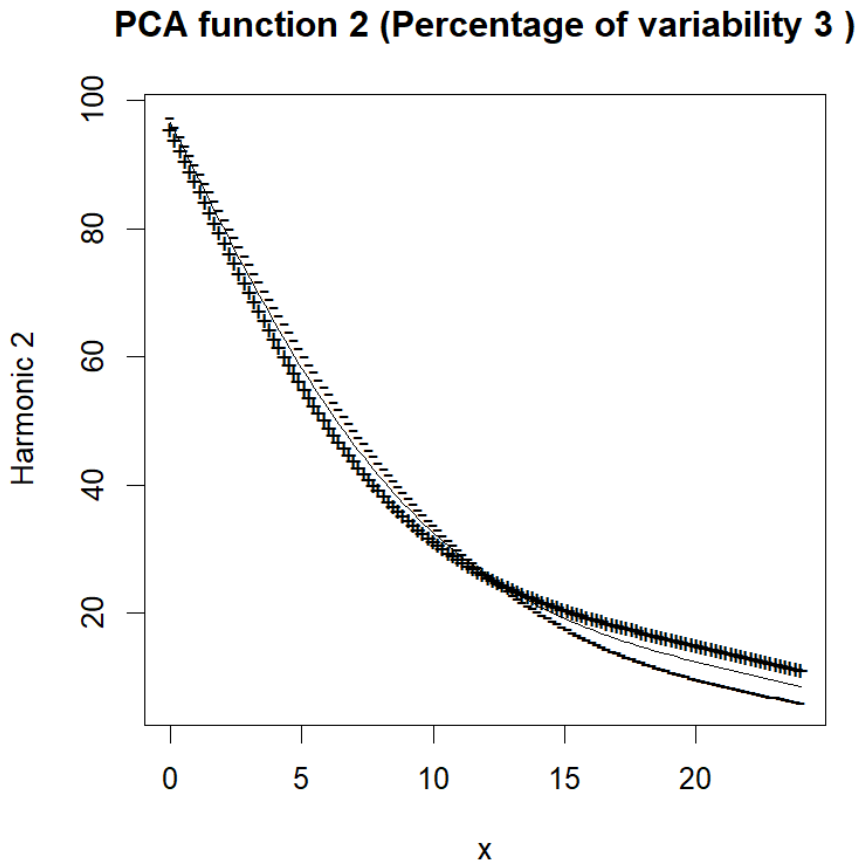




Proportion Variance Explained:
FPC 1: 96.5%
FPC 2: 3.0%

FPCA: Interpretation





PC Functions can be difficult to interpret.

Default plot output in R provides nice visual summary

For each FPC:

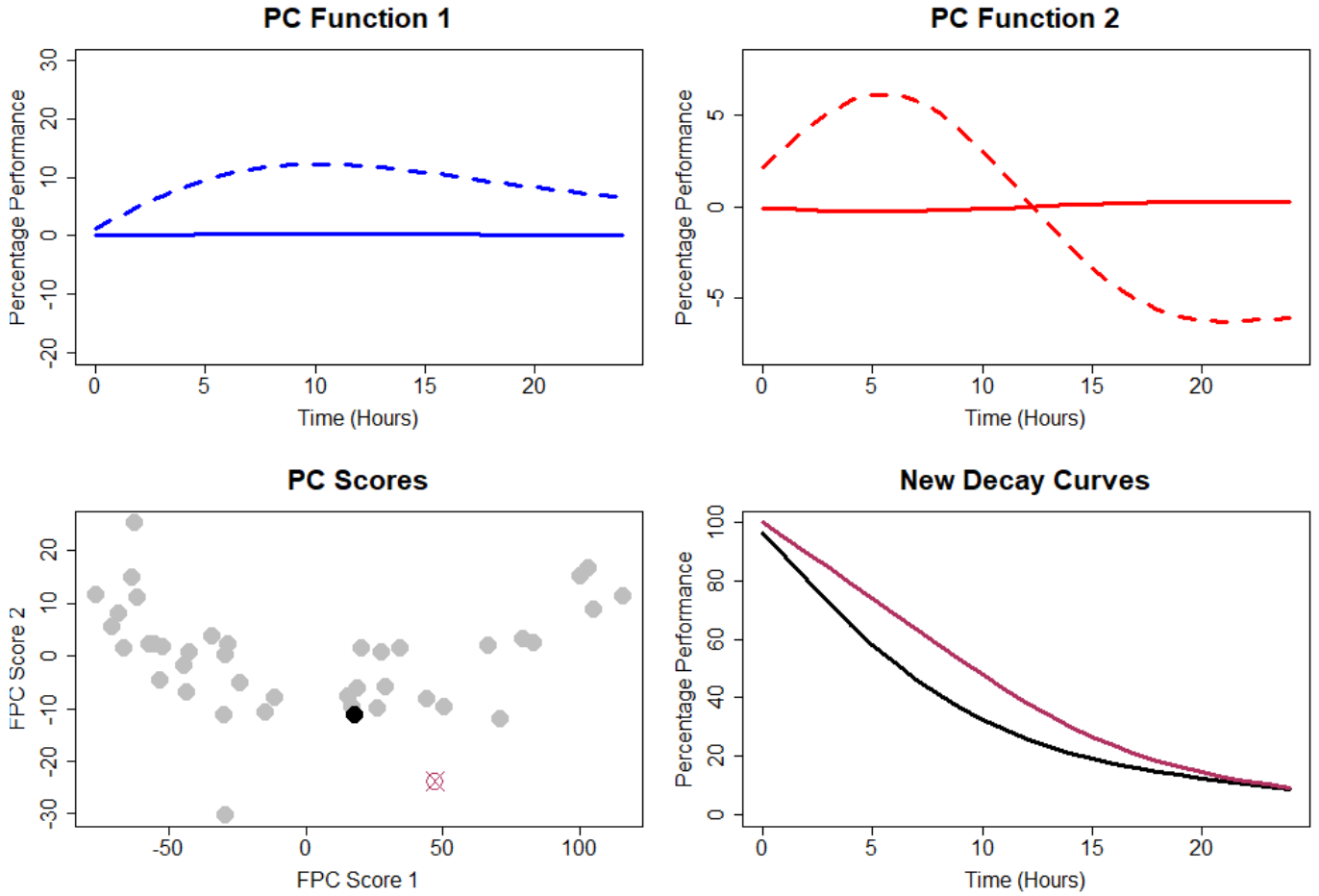
- Percentage of variability provided in main title
- Mean function plotted (solid line)
- Perturbations of equal weight c from the mean represented by '+' and '-' lines.
- '+' line corresponds to $\bar{y}(t) + c\xi_1(t)$
- '-' line corresponds to $\bar{y}(t) - c\xi_1(t)$

In our case study:

- FPC1 controls the 'steepness' of the decay during the first 12 hours
- FPC2 controls decay during the last 12 hours

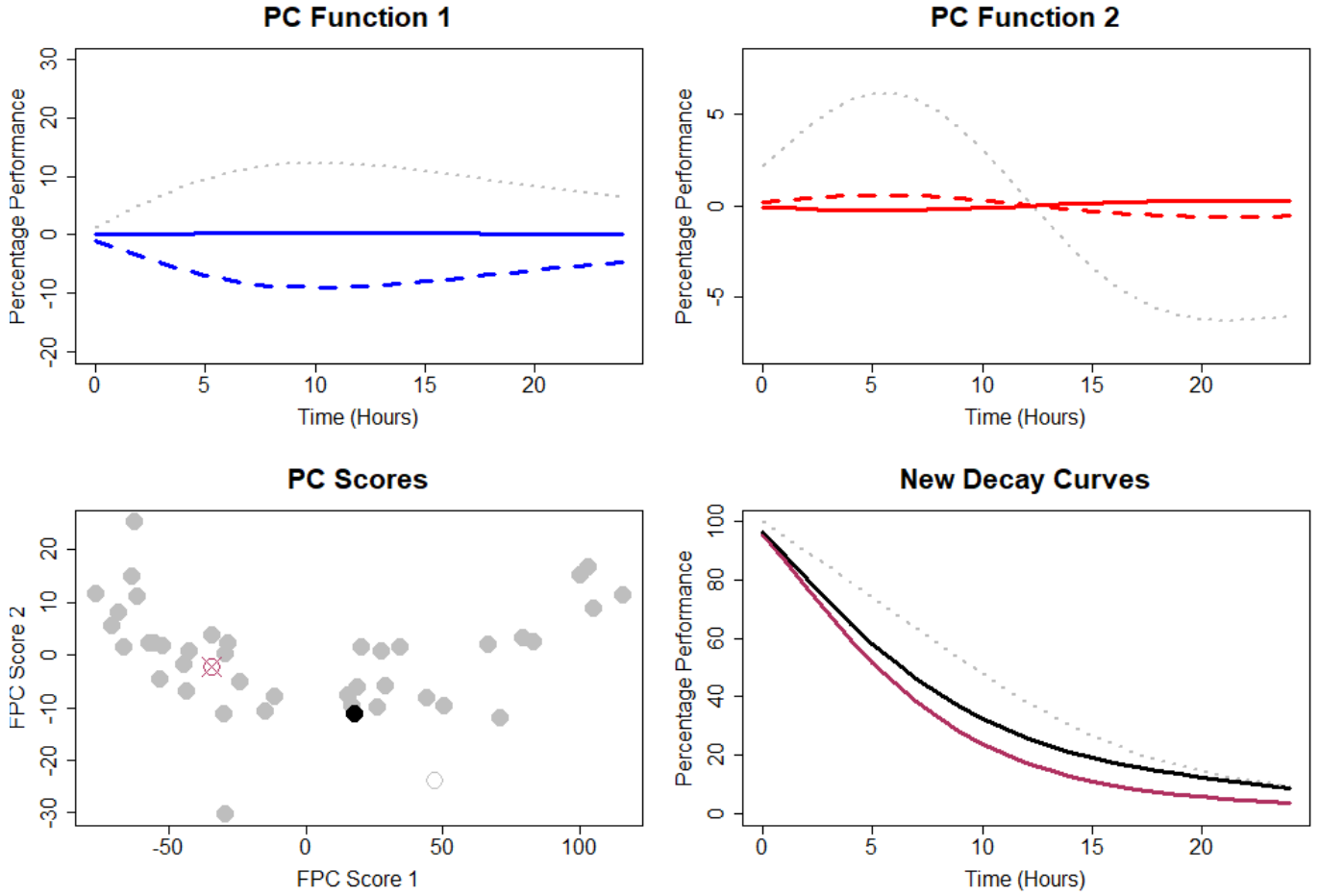
FPCA: Dimension Reduction

Treating the Mean and PC Functions as fixed basis functions (learned from the data), the profile's shape is controlled by 2 scalar coefficients!



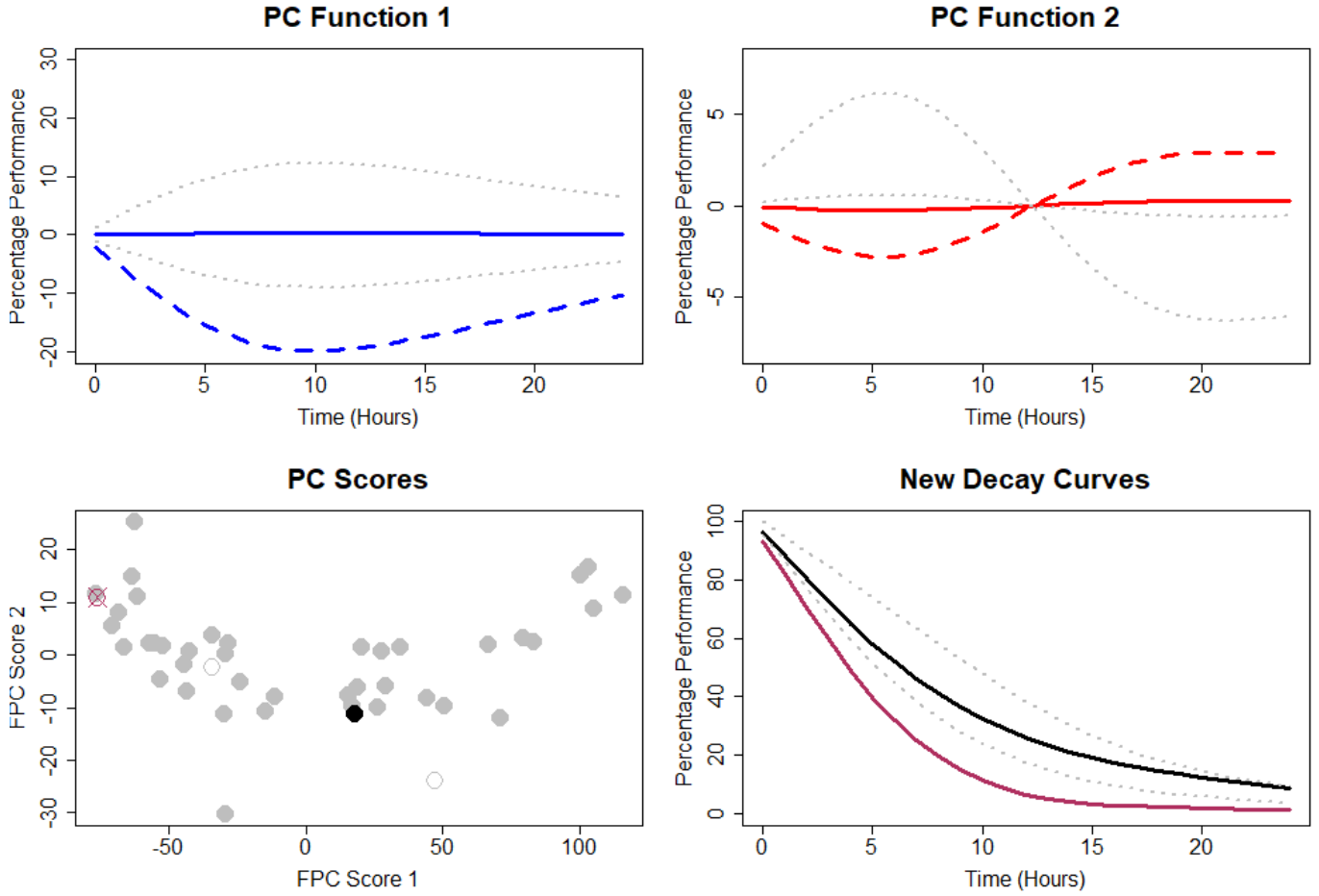
FPCA: Dimension Reduction

Treating the Mean and PC Functions as fixed basis functions (learned from the data), the profile's shape is controlled by 2 scalar coefficients!



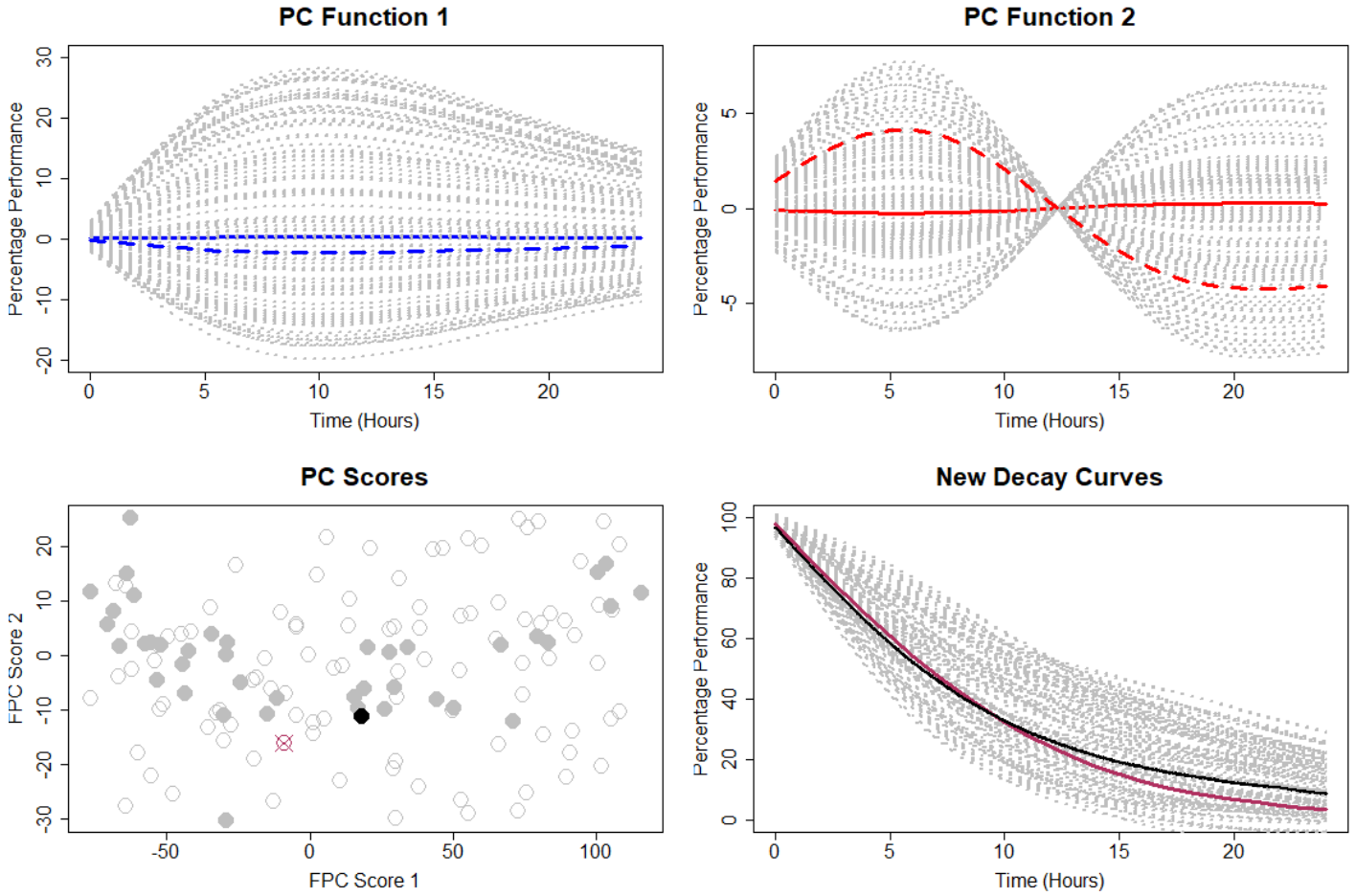
FPCA: Dimension Reduction

Treating the Mean and PC Functions as fixed basis functions (learned from the data), the profile's shape is controlled by 2 scalar coefficients!

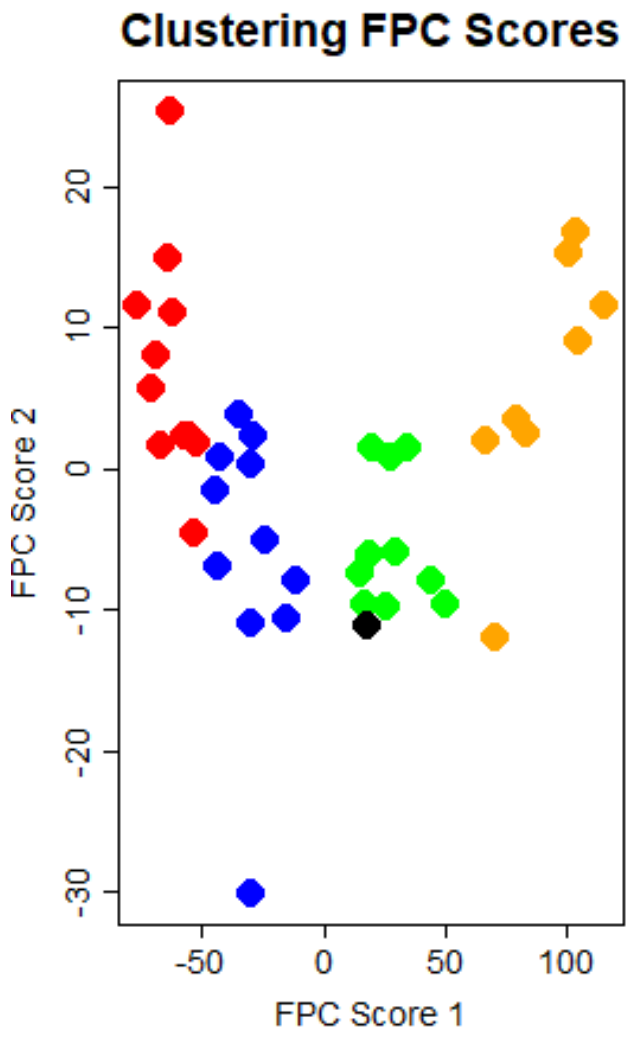


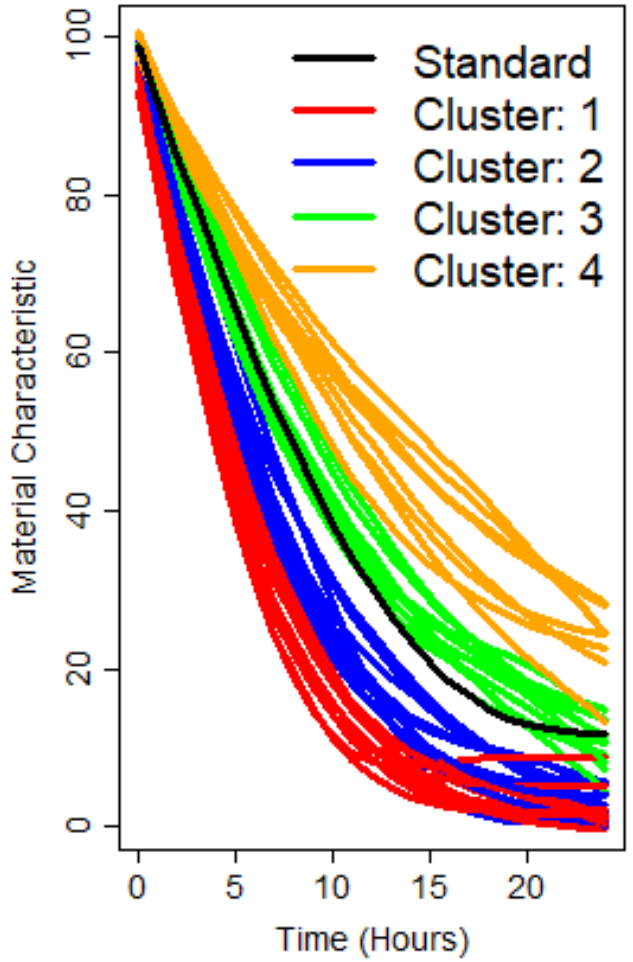
FPCA: Dimension Reduction

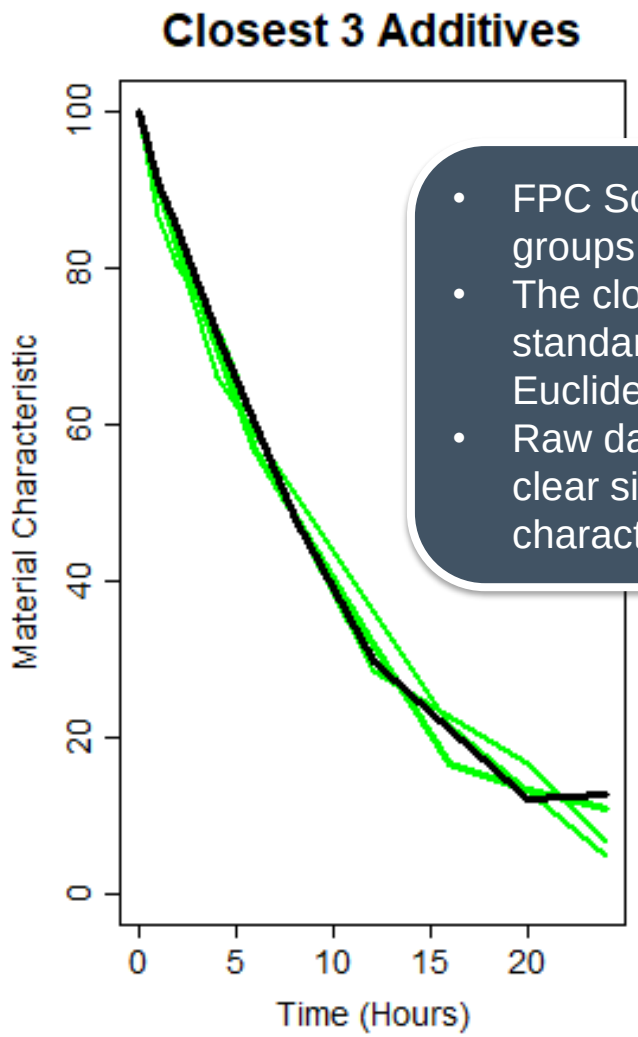
Treating the Mean and PC Functions as fixed basis functions (learned from the data), the profile's shape is controlled by 2 scalar coefficients!



k-Means Clustering of FPC Scores







- FPC Scores clustered into 4 groups
- The closest 3 additives to the standard selected using Euclidean distance.
- Raw data overlay illustrates clear similarity in material characteristic!

PREDICTING FABRICATED ARTICLE PROPERTIES USING MATERIAL CHARACTERISTICS

FUNCTIONAL REGRESSION

Functional Regression: Introduction

Scalar-on-Function: Functional ANOVA (FANOVA)

One Factor Model:

$$y_{ij}(t) = \mu(t) + \alpha_j(t) + \epsilon_{ij}(t)$$

- $y_{ij}(t)$: i-th response function in j-th group
- $\mu(t)$: mean response function
- $\alpha_j(t)$: effect function of j-th group

Canadian Weather Example: Predicting Annual Temperature or Precipitation Profile Based on Climate Zone (Atlantic, Pacific, Continental, Arctic)

Function-on-Function

$$y_i(t) = \alpha(t) + \int x_i(s)\beta(s,t)ds + \epsilon_i(t)$$

- y_i: i-th functional response
- $\alpha(t)$: intercept function
- $x_i(s)$: i-th covariate function
- $\beta(s,t)$: regression surface (approximated with basis expansion)

Canadian Weather Example: Predicting Precipitation Profile from the Temperature Profile Across Weather Stations

Function-on-Scalar: Functional Linear Model

$$y_i = \alpha + \int x_i(s)\beta(s)ds + \epsilon_i$$

- y_i: i-th scalar response
- α: scalar intercept term
- $x_i(s)$: i-th covariate function
- $\beta(s)$: slope function (approximated with basis expansion)

Canadian Weather Example: Predicting Total Precipitation from the Temperature Profile Across Weather Stations

Functional Principal Components Regression

Scalar Response:

$$y_i = C\beta + \epsilon_i$$

Functional Response:

$$y_i(t) = C\beta(t) + \epsilon_i(t)$$

- C: a design matrix containing FPC scores for one or more functional covariates.
- Functional analogue to Principal Components Regression
- For JMP users, FPCR is the method implemented in Functional Data Explorer and to analyze Functional DOE Data

Functional Regression: Simulating Responses

The application performance response values are simulated as follows:

- True Slope Function:

$$\beta(t) = \frac{1}{500}\sqrt{t},$$

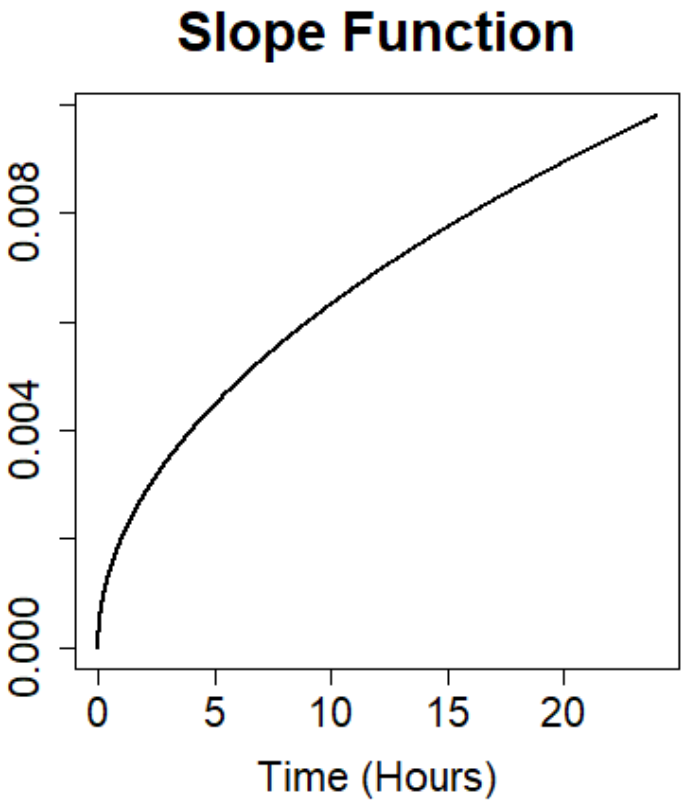
- Random Error:

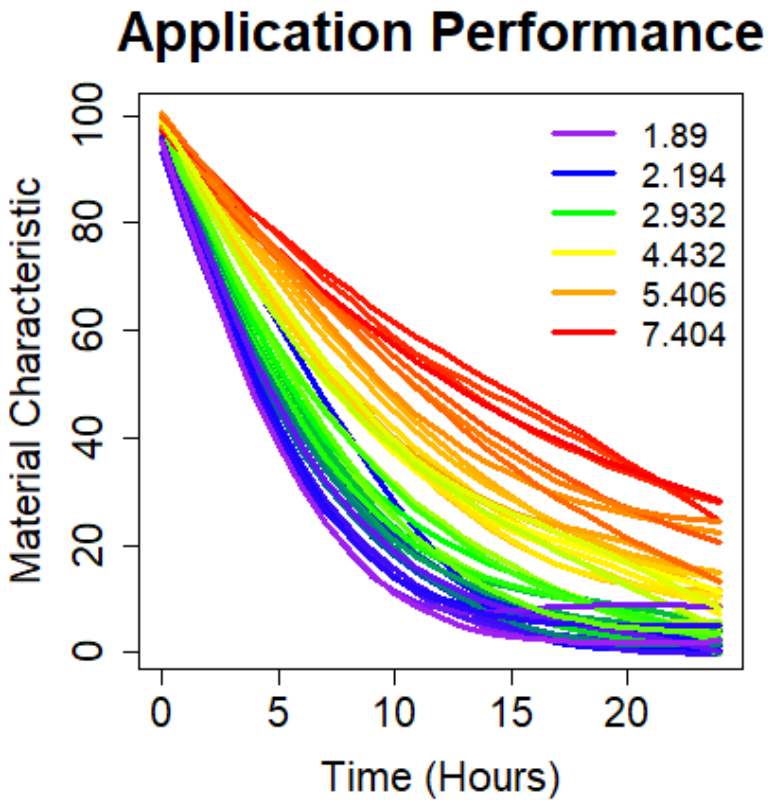
$$\epsilon \sim N(0, \sigma = 0.5)$$

- Response Value Computation:

$$y_i = \int_0^{24} \beta(t) x_i(t) dt + \epsilon_i$$

- For simplicity, no intercept term is added
- Notice both the true slope and true exponential decay functions are used to generate the response





Intuitively, the final performance of a given material is related to slower decay in material characteristic.

Functional Regression: Results

Both the FPCR and FLM models perform equivalently well due to the 2 FPC components capturing 99.5% of the variability in the functional observations

FPCR Model:

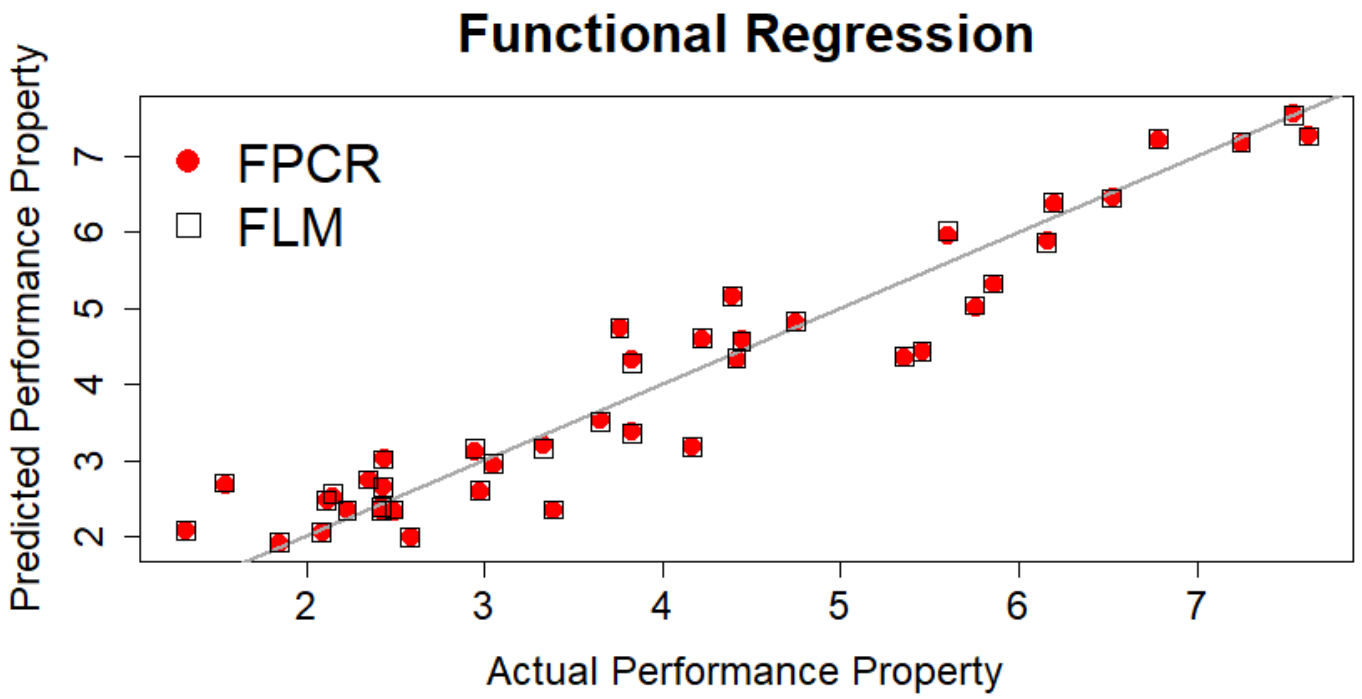
$$y_i \sim 3.99 + 0.029FPC_1 + 0.015FPC_2$$

$$MAE = 0.398$$

FLM Model:

$$y_i \sim 0.64 + \int_0^{24} \hat{x}_i(t)\hat{\beta}(t)dt$$

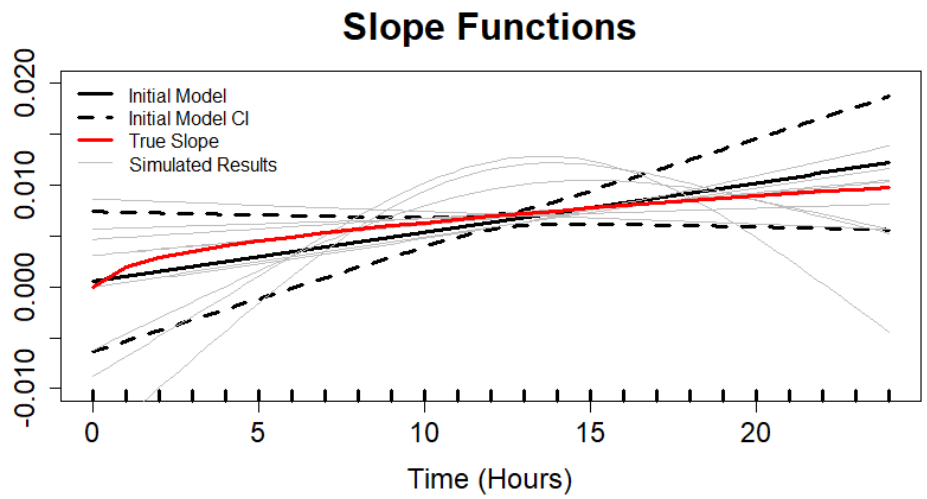
$$MAE = 0.400$$



Functional Regression: FLM vs FPCR Comparison

Functional Linear Model

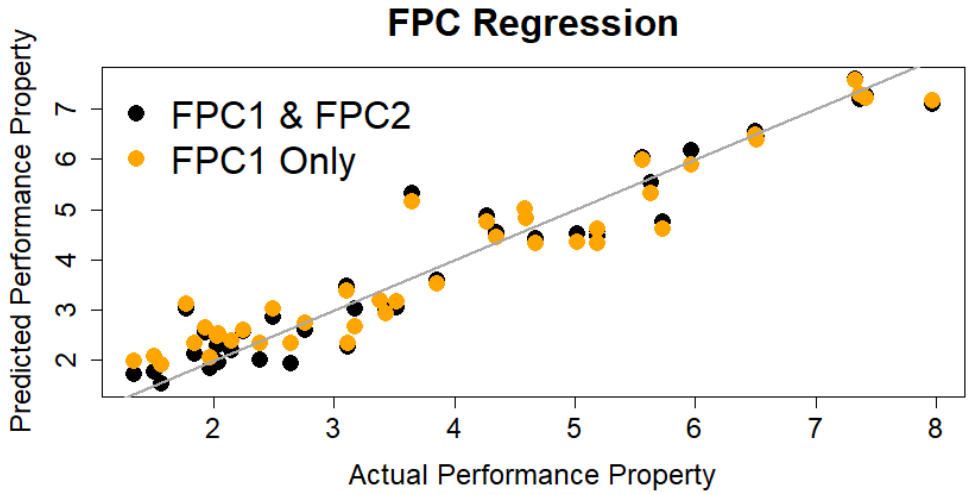
- Pros
  - Produces an estimate of the slope function
  - Utilizes entire functional observation
  - Pre-smoothing data not required when using ***pfr*** function
- Cons
  - Requires more degrees of freedom (basis expansion coefficients)
  - Estimated slope function may not resemble the truth
  - Model may be unstable → large standard errors/CI bands
  - Model inversion/optimization more challenging



- 10 different random seeds generated
- New response values simulated
- Slopes have different shapes
- MAEs range from 0.386 to 0.512

Functional PC Regression

- Pros
  - Simple regression model based on small number of uncorrelated FPC scores
  - Easier model inversion/optimization
  - FPC scores can be used to define functional DOE
- Cons
  - Performance is directly related to proportion of explained variation



- FPC1 Explains 96.5% of the variation
- Losing 3.5% after removing FPC2 increases the MAE from 0.400 to 0.439

Comparing FPCR and FLM to Alternative Modeling Approaches

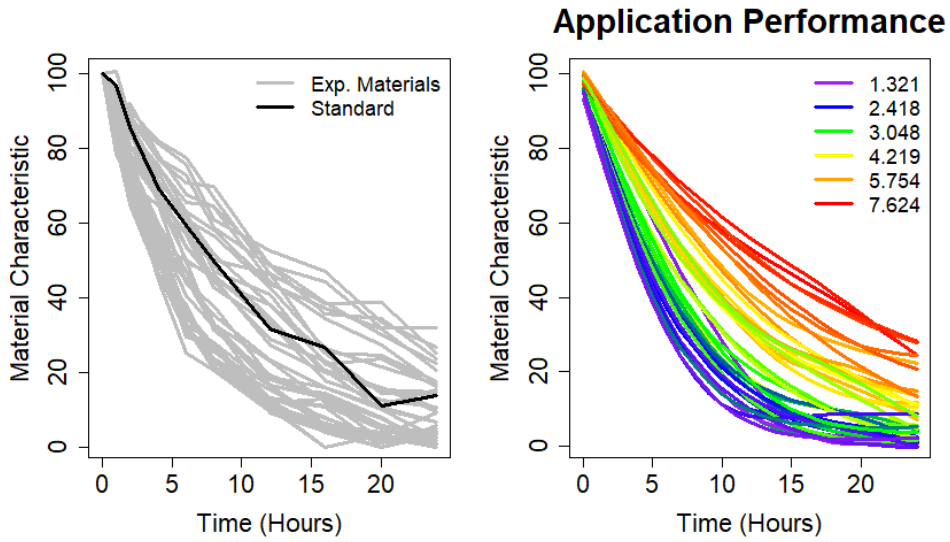
Alternative Modeling Approaches (Original Data)

How do FPCR and FLM compare to alternative modeling approaches:

- Moment Based Linear Models:
  - Numerical Derivatives (Rise/Run)
    - 1 measurement across entire domain [0,24]
    - 3 binned measurements [0,6], [6,16], [16,24]
  - Numerical Integration (AUC)
- Multivariate Based Method
  - Principal Components Regression
- True Data Generating Basis Function
  - Model response using estimated exponential decay parameter

Note: Original 41 decay profiles with non-missing data used for this comparison

Methods	MAE	RMSE
One Derivative	0.596	0.717
Binned Derivatives	0.404	0.507
AUCs	0.384	0.483
PCR	0.34	0.433
Exp. Decay	0.379	0.484
FPCR	0.356	0.449
FLM	0.355	0.447



- In general, moment based models underperform due to information loss when summarizing the profiles
  - This even includes when summarizing the curves using the true basis function model
- Multivariate models can often perform similarly well or better than functional methods for prediction, but:
  - Complete data is required (imputation or smoothing could be a first step)
  - Functional relationship between vectors still lost
  - Interpretation of model and model inversion more difficult

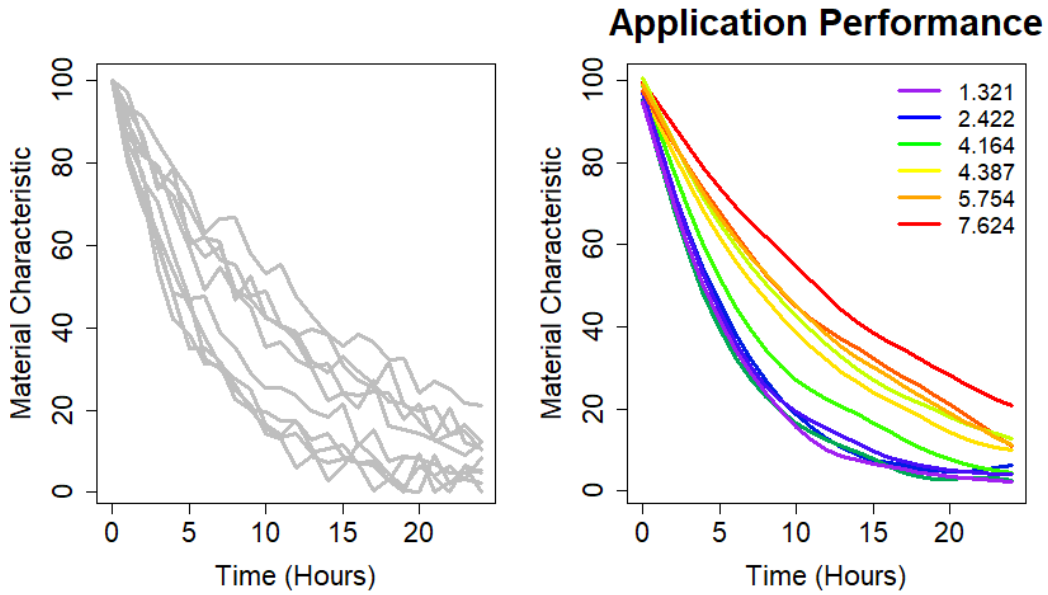
Alternative Modeling Approaches (Extreme Data)

Here we compare the same methods, but in a more extreme scenario

Suppose we have:
- **Only 10 materials characterized**
- **Improved the measurement process and can now measure performance every hour**

This wide data matrix scenario is where functional methods shine

Methods	MAE	RMSE
One Derivative	0.89	0.977
Binned Derivatives	0.463	0.542
AUCs	0.416	0.536
PCR	0.437	0.522
Exp. Decay	0.44	0.566
FPCR	0.385	0.513
FLM	0.368	0.486



Functional PCR and Functional Regression allow us to:

- Incorporate the functional relationship between measurements in a regression model
- Minimize or eliminate information loss
- Interpret the relationship between functional observations and predictors
- More easily invert the regression model to identify profiles which optimize target response values.

Takeaways and Further Thoughts

- Thanks to R packages like ***fda*** and ***refund,*** FPCA, FPCR, and Functional Regression are very approachable and powerful methods to analyzing functional data
  - Those interested in a Python offering, check out ***scikit-fda***
- Depending on the problem at hand Smoothing and FPCA can be useful utilities
  - Consider smoothing as a form of imputation to correct for missing values and/or unequally spaced measurements
  - When no obvious parametric model exists to model distributional or functional data, consider using FPCA to 'learn' a new set of empirical basis functions
  - You could also leverage existing data to guide future experimental designs for functional data using FPCA results.
- In the example shown, we only had one functional predictor, but functional regression methods extend to:
  - Include any combination of scalar (categorical or continuous) and functional input variables.
  - Model either a functional or scalar response (including non-normal responses)
- Though not covered here, for growth/decay data, analyzing the derivative functions is a common approach and no scalar/multivariate analogue exists
- I have shared the code used to produce all data, figures, and analyses in this presentation for you to explore and/or leverage for your own analyses

References

- Textbooks:
  - Ramsay, J. and Silverman, B.W. (2005). Functional Data Analysis.
  - Ramsay, J., Hooker, G., and Graves, S. (2009). Functional Data Analysis with R and MATLAB.
  - Ramsay, J. and Silverman B. W. (2002). Applied Functional Data Analysis: Methos and Case Studies.
  - Kokoszka, P. and Reimherr, M. (2021). Introduction to Functional Data Analysis.
  - Srivastava, A. and Klassen, E. (2016). Functional and Shape Data Analysis.
  - Shi, J. Q. and Choi, T. (2011). Gaussian Process Regression Analysis for Functional Data.
  - Zhang, J. (2013). Analysis of Variance for Functional Data.
- Papers:
  - Morris, J. S. (2015). Functional Regression. *Annual Review of Statistics and Its Application.* Volume 2, pages 321-359
- YouTube:
  - Functional Data Analysis Lecture Series (https://www.youtube.com/c/FunDataScience/videos)

Matthew Malloure

Email: MRMalloure@dow.com

LinkedIn: in/matthew-Malloure

GitHub: @MatthewMalloure