Introduction to Functional Data Analysis
A CSCU Workshop

Giles Hooker
Biological Statistics and Computational Biology
gjh27@cornell.edu

www.bscb.cornell.edu/~hooker/FDAWorkshop
Agenda

- What is Functional Data Analysis? What is Functional Data?
- Exploring functional data
- Modeling with functional data
- Making use of derivatives
- Registration
- Mechanics and resources
What is Functional Data Analysis?

Functional Data Analysis (FDA) is the study of the distribution of functional data, relationships between functional random variables and relationships between functional random variables and other quantities.
What is Functional Data?

Measures of position of nib of a pen when writing "fda". 20 replications, measurements taken at 200 hertz.
Characteristics

- Data are measurements of smooth processes over time
- We usually do not want to make parametric assumptions about those processes.
- Often have multiple measurements of the same process
- We are interested in describing the variation of processes.
- Frequently, collected data have high resolution and low noise.
- Can be applied to any estimate of a smooth process.
Data may be measured more noisily

We need to find the *smooth* process under the data.
Data may be measured more sparsely

- Data are low noise but low-resolution
- Measured at unequal intervals
- We know that the curves must always increase
Characteristics of Functional Data

- Represent a collection of smooth processes over time (or space, or wavelength, or ...)
- Are comparable by time-point; it makes sense to compare time 2.34 in curve 1 to time 2.34 in curve 2.
- Cannot be modeled by simple formulae.
- May be measured noisily and infrequently, but have enough data to estimate a smooth process underlying the data.
- Need not be all measured at the same time points.

There are methods when you only get a few points for each functional observation, but these are beyond our scope.
Exploring Functional Data

How do we understand

1. mean
2. variance
3. covariance
4. important parts of covariance

in functional data?
Taking a Mean

Functional $\bar{x}(t) = \frac{1}{n} \sum x_i(t)$. 
Calculating the Variance

Functional $s^2(t) = \frac{1}{n-1} \sum (x_i(t) - \bar{x}(t))^2$
The Covariance Surface

The covariance between $x(s)$ and $x(t)$ is a \textit{surface}

$$C(s, t) = \frac{1}{n - 1} \sum (x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t))$$
A Closer Look

\[ C(s, t) = \frac{1}{n-1} \sum (x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t)) \]
The Correlation Surface

Normalize the covariance by the variance

\[ R(s, t) = \frac{C(s, t)}{\sqrt{s^2(s)s^2(t)}} \]
Decomposing Covariance

Functional Principal Components Analysis: which function represents *most* of the variance.

Plot mean + pc1, mean - pc1 to assess nature of variation.
More Complicated Covariances

For both $x$ and $y$ co-ordinates:

Principal Component 1

Principal Component 2
Assessing Difference: The Functional t-Test

$$t(t) = \frac{\bar{x}_1(t) - \bar{x}_2(t)}{s(t)}$$

Growth for boys and girls:

*But: don’t look up a threshold in a traditional table!*
Linear Models: Canadian Weather Data

Average daily temperatures for Canadian cities:

![Graph showing mean temperature over the year for Canadian cities.](image)

Farmers want to know how weather relates to *total* precipitation in the year.
Functions as Covariates

In standard linear regression:

\[ y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \]

In FDA, \( x(t) \) is a covariate for each \( t \):

\[ y_i = \beta_0 + \int \beta(t)x_i(t)dt + \epsilon_i \]
Coefficient Functions

*Didn’t my stats prof say that we can’t have more covariates than observations?*

Yes, but also know that $\beta(t)$ is smooth.

- $\beta(t) < 0$ in first part of the year ⇒ high spring temperatures imply lower rainfall.
- $\beta(t) > 0$ in latter part of the year ⇒ high fall temperatures imply high rainfall.
Derivatives: Growth Data

Sometimes, it is the rate of change that is more interesting or informative.
Acceleration

The second derivative is acceleration; it allows us to see when a system is "speeding up":

![Graphs showing velocity and height acceleration over age]
Dynamics

We might also want to look at relationships between derivatives: this gets us closer to the physical laws describing a system.
Registration

Variation is not always vertical

About 25% of variation is explained by changing the *timing* of growth curves.
Basis Expansions

How do I go from data to $x_i(t)$?
We represent it by a basis expansion:

$$x_i(t) = \sum_{j=1}^{K} c_{ij} \phi_j(t)$$
Resources

- [www.functionaldata.org](http://www.functionaldata.org)
- fda packages for both R and Matlab.
- In R package: code available to produce examples from all three books.