ELASTIC FUNCTIONAL AND SHAPE DATA ANALYSIS

Anuj Srivastava

Department of Statistics, Florida State University

Presented at SOCN LECTURES June 2018
First things first

- Let’s introduce ourselves.
- Grade, School, Topic of Study?
Goals for this class

- Learn about the general areas of functional data analysis and shape analysis.
- Focus on fundamental issues and recent developments, not on derivations and proofs.
- Use examples from both simulated and real data to motivate the ideas.
- As much as interactions as possible. Learn by discussion. (Your English is better than my French or Dutch!). Ask questions!
- I can provide matlab code as needed but there are no coding or homeworks planned. Unless you insist.
Background Requirement

- Assume knowledge of real analysis, linear algebra, and numerical analysis.
- This topic area of this class is multidisciplinary, not just interdisciplinary:

We will use ideas from geometry, algebra, functional analysis, and statistics to build up concepts. These are elementary ideas in their own fields but not as elementary for newcomers.
See the word file.
1 Introduction
   - What is FDA
   - What is Shape Analysis

2 Motivation for Elastic Shape Analysis

3 Discrete Versus Continuous
A term coined by Jim Ramsay and colleagues—perhaps in late 1980s or even earlier.

Data analysis where random quantities of interest are functions, i.e. elements of a function space $\mathcal{F}$. $f : D \rightarrow \mathbb{R}^k, M$

Statistical modeling and inference takes place on a function space. One typically needs a metric structure, often it is a Hilbert structure.

Several textbooks have been written with their own strengths and weaknesses.
1. Introduction
   - What is FDA
   - What is Shape Analysis

2. Motivation for Elastic Shape Analysis

3. Discrete Versus Continuous
Restrict to a certain class of objects (planar curves, 3D objects, etc).

Shape Analysis: A set of theoretical and computational tools that can provide:

- **Shape Metric**: Quantify differences in any two given shapes.
  
  How different are these shapes?

- **Shape Deformation/Geodesic**: How to optimally deform one shape into another.
Shape Analysis

- **Shape summary**: Compute sample mean, sample covariance, PCA, and principal modes of shape variability.

- **Shape model and testing**: Develop statistical models and perform hypothesis testing.

- Related tools: ANOVA, two-sample test, $k$-sample test, etc.
Much older and richer area, with ideas from many perspectives.

Generally interested in quantifying differences in shapes of objects. Kendall: *Shape is a property left after removing shape preserving transformations.*

Historically statistical shape analysis is restricted to discrete data; each object is represented by a set of points or landmarks.

Current interest lies in considering continuous objects (examples later). This includes curves and surfaces. These representations can be viewed as functions.

Functions have shapes and shapes are represented by functions. FDA and shape analysis are quite similar in challenges and solutions.
Motivation: Complex Biological Structures

- A lot of interest in studying statistical variability in structures of biological objects, ranging from simple to complex. Abundance of data!
- Working hypothesis —

**Biological Structures Equate with Functionality**

*Proteins: sequence → folding (structure) → function.*
Understanding functions requires understanding structures.

- Structure analysis — a platform of mathematical representations followed by probabilistic superstructures.
- Mathematical representations often form Riemannain manifolds, orbifolds. Probability distributions/inferences on manifolds are challenging.
Some Examples of Biological Structures

- Mitochondria contours – study of shapes.

- Simple closed curves in a plane.
- Interested in ANOVA type analysis with shape as a response variable. Decide significance of external factors.
- Need tools to quantify variability in shapes, and model that regress shapes against predictors.
Some Examples of Biological Structures

- Leaves: Classification of leaves using shapes.
- Boundary: Simple closed curves in a plane.
- Not just boundary but vein structure should also be used.
Some Examples of Biological Structures

- Nanoimaging: Supervising material properties using EM

![Frame i and Frame j](image)

Zoom-ins

- Simple closed curves in a plane.
- Goal is to compare populations of dynamic shapes, not just individual static shapes.
Proteins, RNAs – Structure Analysis

Backbones are represented by curves in $\mathbb{R}^3$. Backbones with side-chain labels are modeled as colored curves.

Protein database has thousands of structures. The goal is analyze these structures for biological variability and functionality.
Some Examples of Biological Structures

- Chromosome structure estimation: Hi-C data

  ![Single Cell](image1)
  ![Ensemble](image2)
  ![Estimation](image3)

- Single cell data is sparse and noisy. Ensemble data averages all structures and loses individuality. Structure estimation is a big problem.

- Once the structures are estimated, one is interested in comparing them and using their shapes to predict/diagnose functionality.
Human biometrics is a fascinating problem area.

Facial Surfaces: 3D face recognition for biometrics

Human bodies: applications – medical (replace BMI), textile design.

Shapes are represented by surfaces in $\mathbb{R}^3$
Examples of Biological Structures - Surfaces

- Protein surfaces, anatomical surfaces

Shapes are represented by surfaces in $\mathbb{R}^3$

- Goal is to cluster and classify proteins, and study structure versus functionality.
Examples of Biological Structures - Trees

- Neurons, axons

- Complex branching structures, different numbers and shapes of branches.

- Interested in neuron morphology for various medical reasons – cognition, genomic associations, diseases.

- More interested in neuronal networks but the starting point is individual neurons.
Human activity data using remote sensing — kinect depth maps

Each skeleton is considered either as an element of $\mathbb{R}^{60}$ or Kendall’s shape space (20 landmarks in $\mathbb{R}^3$).

An action is then a curve on that representation space.

Goal is **action classification** while being invariant to rate at which action is performed.

- Consider data that is sampled from an underlying function.

- If the time points are synchronized across observations, and the focus is only on the heights, then one can work with the vector $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$.

- If the time points are also of significance, then one needs to keep them:
  \[
  \begin{bmatrix}
  (t_1, y_1) \\
  (t_2, y_2) \\
  \vdots \\
  (t_n, y_n)
  \end{bmatrix}
  \]
- The models take the form $y_i = f(t_i) + \text{noise}$.
- What if the time points themselves are noisy?
- One can try to impose noise models on $t_i$s also, but a more natural approach is to consider the full functions.
- Working with continuous functions allows us to resample them at arbitrary points and treat $\{t_i\}$s as variables.

![Graphs showing Interpolation and Resampling](image)

- Allows for compositional noise: $y_{i,j} = a_j f_j(\gamma_j((t_i)) + \epsilon_{i,j}$