

The Functional Data View of Longitudinal Data

by

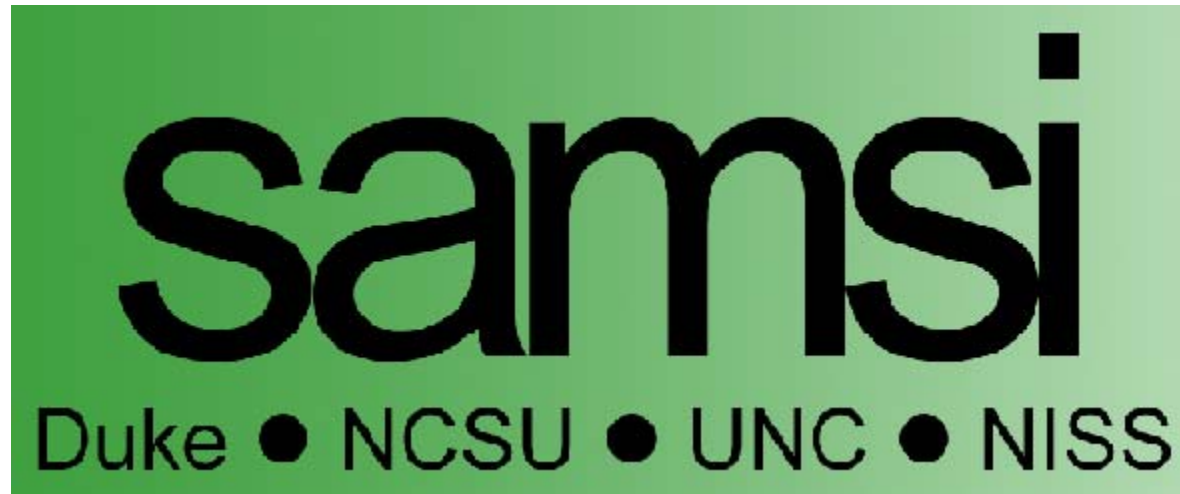
Xin Zhao
Department of Statistics
Cornell University

J. S. Marron
Department of Statistics
University of North Carolina

Special thanks to:

- Cornell University, [O. R. & I. E.](#)
- Participants in the course [OR779](#), Fall, 2001
- Martin T. Wells, Cornell University, ILR

A Brief Advertisement



[\[additional information\]](#)

Functional Data Analysis

“Parallel statistical culture” to Longitudinal Data Analysis?

- Often analyze similar data sets (e.g. this talk)
- But much different viewpoint
- FDA ideas useful for “conventional” LDA?

Functional Data Analysis (cont.)

A Personal View:

What is the “atom” of the statistical analysis?

Statistical Context

1st Course

Multivar. Analysis

F. D. A.

Atom

Number

Vector

Complex Object
(curve, image, shape, ...)

Data Representation

Object Space



Feature space

Curves

Vectors

Images

Shapes

$$\begin{pmatrix} x_{1,1} \\ \vdots \\ x_{d,1} \end{pmatrix}, \dots, \begin{pmatrix} x_{1,n} \\ \vdots \\ x_{d,n} \end{pmatrix}$$

Data Conceptualization

Feature space

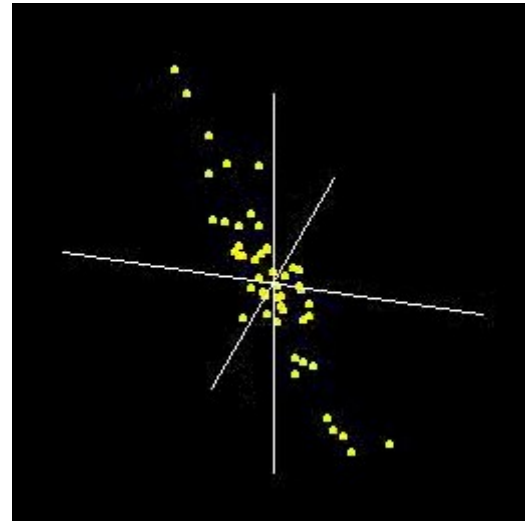


Point Clouds

Vectors

$$\begin{pmatrix} x_{1,1} \\ \vdots \\ x_{d,1} \end{pmatrix}, \dots, \begin{pmatrix} x_{1,n} \\ \vdots \\ x_{d,n} \end{pmatrix}$$

Points



Data Representation (cont.)

Toy Example: curves as data [[Raw Data Graphic](#)]

Object Space \rightarrow Feature Space

via “digitization”

Feature Space \rightarrow Object Space

via “Parallel Coordinates” (Inselberg)

Functional Data Analysis (cont.)

2 main goals

- Understanding “Population Structure”
- Discrimination

(care about this? Then ask about
“Distance Weighted Discrimination”
... an improvement of the SVM)

Understanding Population “Structure”

Common Tool: Principal Component Analysis

Idea: find “directions that maximize variability”

Point Cloud View:

- same point cloud, but “new basis” (coordinate axes)
- **PC1 direction** maximizes variation
- **PC2 direction** maximizes variation in subspace ortho to PC1
- Directions found by eigen-analysis (of covariance matrix)

Understanding Population “Structure” (cont.)

[\[PCA of Toy Example\]](#) of curves

- Separates variation into “intuitive components”
- Finds non-obvious “characteristics of population”
- ANOVA style Sums of Squares give insights
- Method should be in LDA “bag of tricks”???

Real Data Example

“Gene Expression” – Micro-array data

Data (after major preprocessing): Expression “level” of:

- thousands of genes
- but only dozens of “cases”

Interesting statistical issue:

High Dimension Low Sample Size data

Real Data Example (cont.)

Interesting Case:

“Yeast Cell Cycle Data” from

Spellman, P. T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998), “Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization”, *Molecular Biology of the Cell*, 9, 3273–3297.

Cell Cycle Data (cont.)

Lab experiment:

1. Chemically “synchronize cell cycles”, of yeast cells
2. Do cDNA micro-arrays over time
3. Used 18 time points, over “about 2 cell cycles”
4. Studied 4,489 genes (whole genome)

Time series view of data: 4,489 time series of length 18

Functional Data View: 4,489 “curves” [[graphic](#)]

Cell Cycle Data (cont.)

Central question: which genes are “periodic” over 2 cell cycles?

Naïve approach: simple PCA [[graphic](#)]

- No apparent (2 cycle) periodic structure?
- Eigenvalues suggest large amount of “variation”
- PCA finds “directions of maximal variation”
- Often, but not always, same as “interesting directions”
- Here need better approach to study periodicities

Cell Cycle Data (cont.)

To filter out “noise”, project onto subspace of “frequency 2 curves”

(a hyperplane in point cloud plot)

Found via Fourier transform:

Reduce to frequency 2 & harmonics (i.e. even frequencies)

(Just a change of basis and truncation,

i.e. rotation and reduction of coordinate axes)

Even Frequency Analysis

PCA on periodic component of data [\[Graphic\]](#)

- Hard to see periodicities in raw data
- But very clear in PC1 ($\sim \sin$) and PC2 ($\sim \cos$)
- PC1 and PC2 explain 65% of variation (see residuals)
- Recall linear combos of sin and cos capture “phase”

since: $\cos(x - \varphi) = \cos \varphi \cos x + \sin \varphi \sin x = c_1 \cos x + c_2 \sin x$

Frequency 2 analysis

Important features of data appear only at frequency 2,

Hence project data onto 2-dim space of sin and cos (freq. 2)

Useful view: [scatterplot](#)

- Angle (in polar coordinates) shows *phase*
- Colors: Spellman's cell cycle phase classification
- Black was labeled "not periodic"
- Within class phases approx'y same, but notable differences
- Next try to improve "phase classification"

Modified Phase Classification

1. Use outer 200 genes (other numbers tried, “less resolution”)
2. Study “distribution of angles”
3. Used SiZer analysis [[Graphic](#)]

(finds “significant bumps, etc., in histogram”)
4. Carefully redrew boundaries [[Resulting classification](#)]
5. Check by studying angles [[Kernel Density Estimates](#)]

Modified Phase Classification (cont.)

Study Phase Groups as subsets of curves

(Original unprojected versions)

- shows clear separation of phases
- all members strongly periodic
- should include more “boundary case” genes???

Modified Phase Classification (cont.)

Phase groups within full population [\[graphic\]](#)

- Top 200 include “most variable curves in population”
- Other have much less variation (cDNA artifact?)
- Shows periodic curves very important
- Not found by PCA, since too many a-periodic curves

Main Messages

1. FDA viewpoint is powerful and insightful

2. Useful conceptual framework is:

Object Space ↔ Feature Space ↔ Point Cloud

3. FDA approach belongs in tool box of Longitudinal Analyst?

[Good source for more info: Ramsay & Silverman (1997)]

Open Problems and Future Work

- Frequency not exactly 2???
- How serious is the problem?
- Fine tune with complex demodulation?
- Thomson's global warming methodology useful???
- Allow different frequencies for different time series?
- Threshold for labeling genes as "periodic"?