ORIE 779: Functional Data Analysis

From last meeting

Corpora Callosa data

- Fourier Boundary Representation & PCA
- Medial Representation & PCA
- Correlation PCA

PCA & clusters

- Mass Flux Data

From last meeting

Mass Flux Data: PCA

- Scores plot suggested clusters
- Are clusters “really there”?
- Saw interesting structure in 2d scores plots
- Organized as “draft’smans plots”
- Deep question: when are clusters “really there”?
- Approach: SiZer

SiZer Background

- settings: 2-d scatterplot smoothing and1-d histograms
- Fossils Data
- Incomes Data
- Central Question:

  Which features are “really there”?

- Solution, Part 1: Scale space
- Solution, Part 2: SiZer

SiZer Background (cont.)

Smoothing Setting 1: 2-d Scatterplots

E.g. Fossil Data

- from T. Bralower, Dept. Geological Sciences, UNC
- Strontium Ratio in fossil shells
- reflects global sea level
- surrogate for climate
- over millions of years

SiZer Background (cont.)

Smooths of Fossil Data (details given later)

- dotted line: undersmoothed (feels sampling variability)
- dashed line: oversmoothed (important features missed?)
- solid line: smoothed about right?

Central question: Which features are “really there”?

SiZer Background (cont.)

My scatterplot smoothing method (others disagree):

local linear smoothing

Main idea: (illustrated by toy example)

use kernel window to “determine neighborhood”

then “fit a line within the window”

then “slide window along”

Window Width, h, is critical

SiZer Background (cont.)

Smoothing Setting 2: Histograms

Family Income Data: British Family Expenditure Survey, 1975

- Distribution of Family Incomes
- ~ 7000 families

Histogram Analysis:

- Again under- and over- smoothing issues
- Perhaps 2 modes in data?
- Histogram Problem 1: Binwidth (well known)

Central question: Which features are “really there”?

- e.g. 2 modes?
- Same problem as existence of “clusters” in PCA

SiZer Background (cont.)

Why not use (conventional) histograms?

Histogram Problem 2: Bin shift (less well known)

- For same binwidth
- get much different impression
- by only “shifting grid location"

Solution to binshift problem: average over all shifts

- 1st peak all in one bin: bimodal
- 1st peak split between bins: unimodal

Smooth histogram provides understanding,
so should use for data analysis

Another name: Kernel Density Estimate

SiZer Background (cont.)

Kernel density estimation

Recommended Reference (∃ many books):

Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*,
Chapman and Hall: London.

View 1: Smooth histogram

View 2: Distribute probability mass, according to data

E.g. Chondrite data (from how many sources?)

SiZer Background (cont.)

Kernel density estimation (cont.)

Central Issue: width of window, i.e. “bandwidth”, h

E.g. Incomes data: controls critical amount of smoothing

Old Approach: data based bandwidth selection

Recommended reference:

Jones M. C., Marron, J. S. and Sheather, S. J. (1996) A brief survey of bandwidth selection for density estimation, *Journal of the American Statistical Association*, 91, 401-407.

New Approach: "scale space" (look at all of them)

SiZer Background (cont.)

"Scale Space" – idea from Computer Vision

Conceptual basis:

- Oversmoothing = "view from afar" (macroscopic)
- Undersmoothing = "zoomed in view" (microscopic)

Main idea: all smooths contain useful information,
so study "full spectrum" (i. e. all smoothing levels)

Recommended reference:

Lindeberg, T. (1994) *Scale space theory in computer vision*,
Kluwer, Boston.

SiZer Background (cont.)

Fun views:

- Spectrum Movie
- Spectrum Overlay
- Spectrum Suface

Note: the scale space viewpoint makes

"data based bandwidth selection"

much less important (than I once thought….)

SiZer Background (cont.)

SiZer:

Significance of Zero crossings, of the derivative, in scale space

Combines:

- needed statistical inference
- novel visualization

To get: a powerful exploratory data analysis method

Main reference:

Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structure in curves, *Journal of the American Statistical Association*, 94, 807-823.

SiZer Background (cont.)

Basic idea: a “bump” is characterized by:

an increase, followed by a decrease

Generalization: many “features of interest” captured by

sign of the slope of the smooth

Foundation of SiZer:

Statistical inference on slopes, *over scale space*

SiZer Background (cont.)

SiZer Visual presentation:

Color map over scale space:

- Blue: slope significantly upwards (deriv . CI above 0)
- Red: slope significantly downwards (der. CI below 0)
- Purple: slope insignificant (deriv. CI contains 0)

SiZer Background (cont.)

SiZer analysis of Fossils data:

Upper Left: Scatterplot, family of smooths, 1 highlighted

Upper Right: Scale space rep'n of family, with SiZer colors

Lower Left: SiZer map, more easy to view

Lower Right: SiCon map – replace "slope" by "curvature"

Slider (in movie viewer) highlights *different smoothing levels*

SiZer Background (cont.)

SiZer analysis of Fossils data (cont.)

Oversmoothed (top of SiZer map):

- Decreases at left, not on right

Medium smoothed (middle of SiZer map):

- Main valley significant, and left most increase
- smaller valley not statistically significant

Undersmoothed (bottom of SiZer map):

- “noise wiggles” not significant

Additional SiZer color: gray - not enough data for inference

SiZer Background (cont.)

SiZer analysis of Fossils data (cont.)

Common Question: which is “right”?

- decreases on left, then flat (top of SiZer map)
- up, then down, then up again (middle of SiZer map)
- no significant features (bottom of SiZer map)

Answer: *All* are “right”, just different “scales of view”,

i.e. “levels of resolution of data”

SiZer Background (cont.)

SiZer analysis of Incomes data:

Oversmoothed: Only one mode

Medium smoothed: Two modes statistically significant

Confirmed by PhD dissertation of H. P. Schmitz (U. Bonn):

Schmitz, H. P. and Marron, J. S. (1992) Simultaneous estimation of several size distributions of income, *Econometric Theory*, 8, 476-488.

Undersmoothed: many “noise wiggles”, not significant

Again: all are “correct”, just different “scales”

SiZer Background (cont.)

Simulated example 1: Marron - Wand Trimodal, #9

n=100: only one mode "significant"

n=1000: two modes now "appear from background noise"

n=10,000: finally all 3 modes are "really there"

Simulated example 2: Marron - Wand Discrete Comb, #15

- similar lessons to above
- someday: "draw" local bandwidth on SiZer map