ORIE 779: Functional Data Analysis

What is it?

A personal view: what is the "atom" of the statistical analysis?

1st course in statistics: "atoms" are numbers

Statistical multivariate analysis: "atoms" are vectors

Functional Data: "atoms" are more complex objects

Functional Data Analysis (cont.)

FDA: “atoms” are more complex objects, e. g.

- curves (focus of Ramsay and Silverman) [example]
- images, e.g. Cornea data (Cohen, Tripoli) [example]
- shapes, e.g. Corpus Callosum Data (Ho, Gerig) [example]
- 3-d shapes, e.g. Vertebra (Gregg Tracton) [example]

Functional Data Analysis (cont.)

Viewpoint: "analyzing" populations of complex objects

2 common major goals:

I. Understanding "population structure".
   - "visualization"
   - "intuition"

II. Statistical Classification, i.e. Discrimination
   - put into "known groups", based on "training data"
   - e.g. disease diagnosis

Visualization of Population Structure

More effort and thought needed
(by statisticians and scientists in general)

Central Problem:

- Human perceptual system excellent in 1, 2 and 3 dimensions
- But very weak beyond that (time, temperature)
- We “can’t visualize” higher dimensional structures
- Thus challenging to build “intuitive ideas” about populations

Data Representation

Object Space $\leftrightarrow$ Feature space

Curves

Vectors

Images

Shapes

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

One to one mapping couples visualization in Object Space, with statistical analysis in Feature Space

Data Representation (cont.)

Terminology: “feature vector” from field of

Statistical Pattern Recognition

Famous reference (∃ many):

Devijver, P. A. and Kittler, J. (1982) *Pattern Recognition: A Statistical Approach*, Prentice Hall, London.

Caution:

- “features” in that field are entries of vectors

- For me, “features” are “aspects of populations”

Data Representation (cont.)

Interesting early reference (for this type of data analysis):

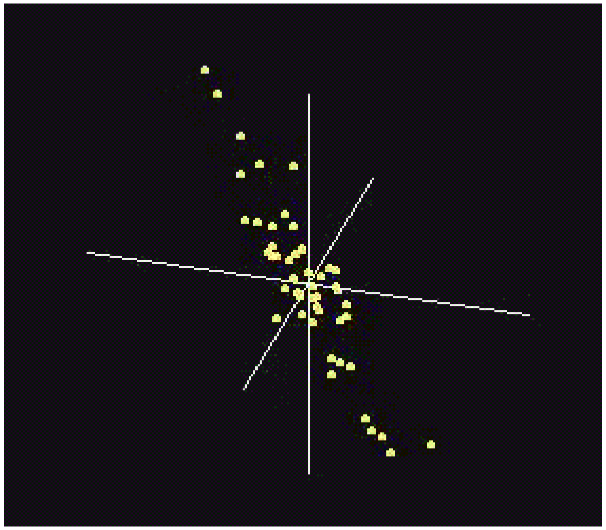
Cootes, T. F., Hill, A., Taylor,C. J. and Haslam, J. (1993) The use of active shape models for locating structures in medical images, *Information Processing in Medical Imaging*, H. H. Barret and A. F. Gmitro, eds., Lecture Notes in Computer Science 687, 33-47, Springer Verlag, Berlin.

High Dim'al Data Conceptualization

Feature space $\leftrightarrow$ Point Clouds

Vectors

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



[Spinning Point Cloud Graphic] .

E.g. 1: Curves [example]

Data Objects: $f_1(x),\ldots,f_n(x)$ (conceptual model)

Digital version: $\begin{pmatrix} f_1(x_1) \\ \vdots \\ f_1(x_d) \end{pmatrix},\ldots,\begin{pmatrix} f_n(x_1) \\ \vdots \\ f_n(x_d) \end{pmatrix}$, for a “grid” $x_1,\ldots,x_d$

Object Space View: Overlay plots of curves

Feature space: $\left\{ \begin{pmatrix} f_i(x_1) \\ \vdots \\ f_i(x_d) \end{pmatrix} : i = 1,\ldots,n \right\}$, e.g. dimension $d = 10$

Aside

In statistical visualization community, this is called the "parallel coordinates" or "parallel axes" plot

Proposed by:

Inselberg, A. (1985) The plane with parallel coordinates, *The Visual Computer*, 1, 69-91.

as a method to visualize high dim'al data (in general)

In particular: visualize each data vector $\underset{\rightarrow}{x_i} = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{id} \end{pmatrix}$,

by "connecting the dots" through $(1, x_{i1}), (2, x_{i2}), \ldots, (d, x_{id})$

E.g. 2: Images, Corneas [example]

Special thanks to K. L. Cohen and N. Tripoli,
UNC Ophthalmology

Reference:

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. and Cohen, K. L. (1999) Robust Principal Component Analysis for Functional Data, *Test*, 8, 1-73.

Data Objects: color map of "temperature scale radial curvature"

- "hot" = more curvature

- "cool" = less curvature

E.g. 2: Images, Corneas [example] (cont.)

Feature vectors: Digitized version is “large and wasteful”

Instead use coefficients of Zernike Basis repres’n, $d = 66$

Schwiegerling, J., Greivenkamp, J. E., and Miller, J. M. (1995) Representation of videokeratoscopic height data with Zernike polynomials, *Journal of the Optical Society of America, Series A*, 12, 2105-2113.

Born, M. and Wolf, E. (1980) *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Pergamon Press, New York.

E.g. 2: Images, Corneas [example] (cont.)

Object Space view: can’t overlay images

Instead show images sequentially

Hard to see “population structure”

E.g. 3: shapes, Corpora Callosa [example]

Special thanks to G. Gerig and S. Ho, UNC Computer Science

Reference:

Kelemen, A., Szekely, G. and Gerig, G. (1997) Three dimensional model-based segmentation, TR-178 Technical Report Image Science Lab, ETH Zurich.

Data Objects: boundaries of “segmented” corpora callosa

E.g. 3: shapes, Corpora Callosa [example]

Feature vectors: use coefficients of Fourier boundary representation, $d = 80$

Object Space view: can either overlay, or show sequentially

In either case: hard to see "population structure"

Finding and visualizing structure in populations

Powerful method: Principal Component Analysis

Our approach:

1. Focus on visualization, with only heuristic mathematics

2. Later carefully revisit mathematics & computation
   (easier with intuition and motivation in mind)

Principal Component Analysis (PCA)

There are many names (lots of reinvention?):

Statistics: Principal Component Analysis (PCA)

Social Sciences: Factor Analysis (PCA is a subset)

Probability / Electrical Eng: Karhunen – Loeve expansion

Applied Mathematics: Proper Orthog'l Decomposition (POD)

Geo-Sciences: Empirical Orthogonal Functions (EOF)

PCA, (cont.)

There are many applications / viewpoints:

- dimension reduction (statistics / data mining)
- change of basis (linear algebra)
- transformation (statistics)
- data compression (electrical engineering)
- signal denoising (acoustics / image processing)
- optimization (operations research)