ORIE 779: Functional Data Analysis

From last meeting

Dual eigen-problem
- Allows fast computation in HDLSS settings

Statistics of PCA
- Gaussian Likelihood view
- Dimension reduction view
- Data Compression view

PCA for shape
- Corpus Callosum data
- Fourier Boundary representation

PCA for shapes (cont.)

Raw Data   Modes of shape variation?

PC1:

- Direction is "overall bending"
- Colors explained later (sub populations)
- An outlier?
- Find it in the data?   [numbered data]
- Case 2:   could delete

PCA for shapes (cont.)

Corpus Callosum data (cont.)

PC2:

- Rotation of right end
- "Sharpening" of left end
- "Location" of left end
- These are correlated with each other

PC3:

- "thin" vs. "thick"

PCA for shapes (cont.)

Alternate summarization of Corpus Callosum data:

Medial Representation: “M-Reps”

Idea: discrete “skeleton” of shape

Summarization: features are “location and angle parameters”

Special thanks to Paul Yushkevich, UNC Computer Science

PCA for shapes (cont.)

Raw data

- from same data as above Fourier boundary rep'n
- but they look different
- since different type of fitting was done
- also, worst outlier was deleted

modes of variation?

PCA for shapes (cont.)

PC1:

- “Overall bending”
- Same as PC1 for Fourier boundary analysis, above
- Correlated with “right end fattening”

PC2:

- “Rotation of ends”
- similar to PC1 for Fourier boundary analysis, above

PCA for shapes (cont.)

PC3:

- systematic “distortion of curvature”
- this time *different* from above Fourier boundary PC3
- Lesson: different rep’ns focus on different aspects of data
- I.e. not just differences in fitting
- But instead on features that are emphasized
- Thus choice of “features” is *very important*

PCA for shapes (cont.)

PC4:

- more like fattening and thinning
- i.e. similar to Fourier boundary PC3
- but “more local” in nature
- an important property of M-reps

Variation on PCA

Replace covariance matrix with correlation matrix

I.e. do eigen analysis of $$R = \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,d} \\ \rho_{1,2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{d-1,d} \\ \rho_{1,d} & \cdots & \rho_{d-1,d} & 1 \end{pmatrix}$$

Where $$\rho_{i,j} = \frac{\operatorname{cov}(X_i, X_j)}{\sqrt{\operatorname{var}(X_i)\operatorname{var}(X_j)}}$$

Variation on PCA (cont.)

Why use correlation matrix?

Reason 1: makes features "unit free"

e.g. M-reps:

- mix "lengths" with "angles" (degrees? radians?)
- are "directions in point cloud" meaningful or useful?
- Will unimportant directions dominate?

Variation on PCA (cont.)

Alternate view of correlation PCA:

Ordinary PCA on standardized (whitened) data

I.e. PCA on data matrix $\tilde{\tilde{X}} = \begin{pmatrix} \frac{X_{1,1} - \overline{X}_1}{s_1} & \cdots & \frac{X_{1,n} - \overline{X}_1}{s_1} \\ \vdots & \ddots & \vdots \\ \frac{X_{d,1} - \overline{X}_d}{s_d} & \cdots & \frac{X_{d,n} - \overline{X}_d}{s_d} \end{pmatrix}$

Distorts “point cloud” along coordinate directions

Variation on PCA (cont.)

Reason 2 for correlation PCA:

Sometimes “whitening” is a useful operation

(e.g. M-rep data)

Caution: sometimes this is *not* helpful

- can lose important structure this way

E.g. 1: Cornea data - elliptical vs. spherical PCA

Variation on PCA (cont.)

E.g. 2: Corpus Callosum Data

Correlation PC1, PC2, PC3

- Not useful directions
- No insights about population
- Driven by “high frequency” artifacts
- Reason: “whitening” has damped the important structure
- By magnifying high frequency noise
- Parallel coordinates show what happened

Variation on PCA (cont.)

Summary on correlation PCA:

- Can be useful (especially with “noncommensurate units”)
- But not always, can also hide important structure in data
- To make choice, decide whether “whitening” is useful
- My personal use of correlation PCA is rare
- Other people use it “most of the time”

PCA and clusters

Recall Toy Example of “2 clusters of parabolas”

Recall PCA:

- Dominant direction finds very distinct clusters
- “skewer through meatballs” (in point cloud space)
- shows up clearly in scores plot
- An important use of scores plot is finding such structure

PCA and Clusters (cont.)

A deeper example: the Mass Flux Data

Data from Enrica Bellone,

National Center for Atmospheric Research

- “Mass Flux” for quantifying “cloud types”
- How does “mass change” when “moving into” a cloud
- Tried Standard PCA

Mass Flux PCA (cont.)

Mean: Captures “general mountain shape”

PC1: Generally “overall height of peak”

- shows up nicely in mean +- plot (2nd column)
- 3 apparent clusters in scores plot
- Are those “really there”?
- If so, could lead to interesting discovery
- If not, could waste effort in investigation

Mass Flux PCA (cont.)

PC2: Location of peak

- again mean +- plot very useful here

PC3: Width adjustment

- again see this most clearly in mean +- plot

Mass Flux PCA (cont.)

Investigation of PC1 Clusters:

Main Question: “Important structure” or “sampling variability”?

Approach: SiZer (SIgnificance of ZERo crossings of deriv.)

Idea: at a “bump” $\hat{f}$ goes up then down, so highlight as

Blue when deriv. significantly > 0

Purple when deriv. not significant

Red when deriv. significantly < 0

Mass Flux PCA (cont.)

Will discuss SiZer next time, in the meantime can look at:

http://www.stat.unc.edu/faculty/marron/DataAnalyses/SiZer_Intro.html

SiZer conclusion: find 3 significant clusters!

- Correspond to 3 known “cloud types”
- Worth deeper investigation

Mass Flux PCA (cont.)

Improved view of mass flux PCA, color the clusters

[Colored PCA (parts)]

- Use minima of smooth histogram to draw boundaries
- Clusters well separated in full data
- Although not clear a priori
- Same for residuals
- Can see “gaps” in PC1

Mass Flux PCA (cont.)

Another useful view: [2-d scatterplots of scores]

Terminology: these linked scatterplots are called

“Draftsman’s Plots”

- Clear systematic patterns
- But not well separated by these directions
- PCA optimizes “variation”, not “separation of clusters”
- Can find “better directions”?

Mass Flux PCA (cont.)

An attempt at "better directions" for PC3 and PC4

Idea: "rotate" subspace gen'd by PC3 and PC4

To better "visually separate" colors

Manually selected axes

Resulting Draftsman's plot

- better color separation in many plots

Really useful direction???? Resulting curves

To do later (???):

1. SiZer intro
2. PCA time series – chemometrics data
3. Independent Component Analysis
4. In vector space, orthogonal basis introduction
5. Fourier basis
6. Legendre basis
7. Tensor product Fourier Legendre basis
8. Zernike basis
9. Revisit cornea data? (compare “raw image” with “fit images”, fiddle with Cornean power map? (do this at home?), use Figure from LMTZ paper, see directories D:\DellInspiron7000\SW30\Docs\Steve and D:\DellInspiron7000\SW30\Pictures)
10. Elliptical Fourier bases
11. Complex plane representation (no simple real valued basis)
12. Corpora Collosa Approximation
13. Discrimination – Corpus Collosum Data

14. Fisher Linear Discrimination
15. High dimensional geometry?
16. Support Vector Machines
17. Polynomial Embedding
18. Micro-Array Data analysis
19. Normal KerCli discrimination (in Cornean/demo)