ORIE 779: Functional Data Analysis

From last meeting

Important duality:

Object Space $\leftrightarrow$ Feature Space

Principal Component Analysis – for curves

Gave "decomposition of variation":

Toy E.g. PCA for Parabolas (cont.): Curve View PCA

PCA for Images

Real Data Example: Cornea Data

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

Visualization (generally true for images):

- more challenging than curves (since can't overlay)
- instead view sequence of images
- harder to see "population structure" (then for curves)
- so PCA type decomposition of variation is more important

PCA for Images (cont.) Cornea Raw Data

Recall: nature of images (on disk)

- Color is “curvature”
- Along radii of circle (direction with most effect on vision)
- Hotter (red, yellow) for “more curvature”
- Cooler (blue, green) for “less curvature”
- Feature vector is coefficients of Zernike expansion
- Zernike basis: related to Fourier basis, on disk
- Conveniently represented in polar coordinates

PCA for Images (cont.) Cornea Raw Data

Recall: PCA can find (often insightful) dir'n of greatest variability

Feature Space Viewpoint: [Simple Illustrative Example]

Main problem: display of result (no overlays for images)

Solution: show movie of "marching along the direction vector"

[Cornea Data PC1 movie]

PCA for Images (cont.) Cornea Raw Data

PC1: [movie version]

Mean (starting image): mild vertical astigmatism

- known population structure called “with the rule”

Main direction: “more curved” & “less curved”

- corresponds to first optometric measure
  (89% of variation, in usual Mean Resid. SS sense)

Also: “stronger astigmatism” & “no astigmatism”

Note: found correlation between astigmatism and curvature

Scores (blue lines): Apparent Gaussian (Normal) dist’n

PCA for Images (cont.) Cornea Raw Data

PC2: [Movie version]

Mean: same as above

- common centerpoint of point cloud
- Are studying “directions from mean”

Images along direction vector:

- Looks terrible???
- Why?

PCA for Images (cont.) Cornea Raw Data

PC2 (cont.): [Movie version]

Reason made clear in Scores Plot (blue lines):

- Single outlying data object drives PC direction
- A known problem with PCA
- Recall finds direction with “max variation”
- In sense of variance
- Which is easily dominated by single large observation

Toy example graphic

PCA for Images (cont.) Cornea Raw Data

PC2 (cont.): [Movie version]

How bad is this problem?

View 1: Statistician: Arrrgghh!!!!

- Outliers are very *dangerous*
- Can give arbitrary and meaningless directions
- What does 4% of MR SS mean???

PCA for Images (cont.) Cornea Raw Data

PC2 (cont.): [Movie version]

How bad is this problem?

View 2: Ophthalmologist: No Problem

- Driven by "edge effects" (note many such in raw data)
- Artifact of "light reflection" data gathering
  ("eyelid blocking", and drying effects)
- Routinely "visually ignore" those anyway
- Found interesting (and well known, see data) direction of:
  steeper superior vs steeper inferior

PCA for Images (cont.) Cornea Raw Data

For the moment continue with opthalmologists view

PC3: [Movie version]

Edge Effect Outlier is present

But focusing on “central region”

- shows changing direction of astigmatism (3% of MR SS)
- “with the rule” (vertical) vs. “against the rule” (horizontal)
- most astigmatism is “with the rule”
- most of rest is “against the rule” (known folklore)

PCA for Images (cont.) [Cornea Raw Data]

For the moment continue with ophthalmologists view

PC4: [Movie version]

- Other direction of astigmatism???
- Location (often called “registration”) effect???
- Harder to interpret
- OK, since only 1.7% of MR SS
- Substantially less than for PC2 & PC3

PCA for Images (cont.) Cornea Raw Data

Ophthalmologists View (cont.)

Overall Impressions / Conclusions:

- Useful decomposition of population variation
- Useful insight into population structure

PCA for Images (cont.) Cornea Raw Data

Now return to Statistician's View:

How can we handle these outliers?

Even though not fatal here, can be for other examples:

Recall Simple Toy Example (in 2d)

Enhancement of Parabolas Toy Example: Raw Data

Outliers in PCA (cont.)

Parabolas + Outlier Toy Example: Raw Data

- Why is it an outlier?
- never leaves range of other data
- but Euclidean distance to others very large
- relative to other distances
- also major intuitive difference in terms of "shape"
- and even "smoothness"
- Important lesson: $\exists$ *many directions* in $\Re^d$

Outliers in PCA (cont.)

Parabolas + Outlier Toy Example: PCA

- At first glance, mean and PC1 look similar to Parabs PCA
- PC2 clearly driven completely by outlier
- Score plot on right gives clear outlier diagnostic
- Outlier does not appear in other directions
- Previous PC2, now appears as PC3
- Total Power (upper right plot) now “spread farther”

Outliers in PCA (cont.)

Parabolas + Outlier Toy Example (cont): PCA

Closer look:

Mean “influenced” a little, by the outlier [toy illustration]

- appearance of “corners” at every other coordinate

PC1 substantially “influenced” by the outlier

- Clear “wiggles”

Outliers in PCA (cont.)

What can (should?) be done about outliers?

Context 1: outliers are important aspects of the population

- they need to be highlighted in the analysis
- although could separate into subpopulations

Context 2: outliers are “bad data”, of no interest

- recording errors? Other mistakes?
- Then very important to avoid poor view by PCA

Outliers in PCA (cont.)

Common approaches to dealing with outliers:

I. Outlier deletion: Kick out “bad data”

II. Robust Statistical methods:

Work with full data set, but “downweight” bad data

“Reduce influence”, instead of “deleting”

Outliers in PCA (cont.)

Outlier Deletion:

Useful Diagnostic: Score plot (seen above)

Example Cornea Data:

- Can find PC2 outlier (by looking through data (careful!))
- Problem: after removal, another point dominates PC2
- Could delete that, but then another appears
- After 4th step have eliminated 10% of data ($n = 43$)

Outliers in PCA (cont.)

Example (cont.) Cornea Data:

Motivates alternate approach: Robust Statistical Methods

Recall main idea: downweight (instead of delete) outliers

$\exists$ a large literature. Good intro's (from different viewpoints) are:

Huber (1981) *Robust Statistics*, Wiley, New York.

Hampel, Ronchetti, Rousseeuw and Stahel (1986) *Robust statistics: the approach based on influence functions*, Wiley, New York.

Staudte, R. G. and Sheather, S. J. (1990) *Robust estimation and testing*, Wiley, New York

Robust Statistics

A simple robustness concept: “breakdown point”

- how much of data “moved to ∞” will “destroy estimate”?
- Usual mean has breakdown 0 [toy example]
- Median has breakdown ½ (best possible)
- Conclude median much more robust than mean
- Median uses all data
- Median gets good breakdown from “equal vote”

Robust Statistics (cont.)

Controversy: is median's "equal vote" scheme good or bad?

Huber: Outliers contain some information,

- so should only control "influence" (e.g. median)

Hampel, et. al.: Outliers contain no useful information

- should be assigned weight 0 (not done by median)
- using "proper robust method" (not simply deleted)

Robust Statistics (cont.)

Robustness Controversy (cont.):

- *both* are "right" (depending on context)
- Source of major (unfortunately bitter) debate!

Application to Cornea data:

- Huber's model more sensible
- Already know $\exists$ some useful info in each data point
- Thus "median type" methods are sensible