From last meetings

Class Web Page:
http://www.stat.unc.edu/faculty/marron/321FDAhome.html

Important duality:

Object Space $\leftrightarrow$ Feature Space

Goal I: Understanding "population structure".

PCA for curves (simulated parabolas)

Show CurvDat\ParabsCurvDat.ps

PCA for Images:

E.g. 3: Cornea Data

Again show CorneaRobust\NORMLWR.MPG

PCA: can find direction of greatest variability

Again show CorneaRobust/SimplePCAeg.ps

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

Solution: show movie of “marching along the direction vector”

Show CorneaRobust\NORM100.MPG

PCA for Images, E.g. 3: Cornea Data

PC1:

Mean: mild vertical astigmatism
(known population structure called “with the rule”)

Main direction: “more curved” & “less curved”
(corresponds to first optometric measure)

Also: “stronger astigmatism” & “no astigmatism”

Note: found correlation between astigmatism and curvature

Projections (blue lines): Looks like Gaussian (Normal) dist’n

PCA for Images, E.g. 3: Cornea Data

PC2:

Show CorneaRobust\NORM200.MPG

Mean: same as above (common centerpoint)

Projections: edge effects $\Rightarrow$ "outliers"

$\Rightarrow$ "pulls off PC direction"????

Show CorneaRobust\OutliersPCA.ps

Ophthalmologists: no problem, always "ignore edge effects",
This direction is known: "steep at the top & bottom"

PCA for Images, E.g. 3: Cornea Data

Me: Arrggghh!!!! Outliers are very dangerous

Approach described later: Robust PCA

Results:

Robust PC1: captures same structure

Show CorneaRobust\NORM122.MPG

Robust PC2: Same structure
unaffected by outlier
Gaussian projection distribution

Show CorneaRobust\NORM222.MPG

PCA for Images, E.g. 3: Cornea Data

PC3:

Regular: Edge effect outlier is present,
Astigmatism “with the rule” and “against the rule”

Show CorneaRobust\NORM300.MPG

Robust: Eliminate outliers
But main effect diminished

Show CorneaRobust\NORM322.MPG

Overall: insightful “views of population structure”

PCA for Shapes:

E.g. 4: Corpora Callosa Data

Again show CorpColl\CCFrawAlls3.mpg

PCA, part 1: shapes

PC1: major bending (note outlier)

Show CorpColl\CCFpcaSCs3PC1.mpg

PC2: shape of ends

Show CorpColl\CCFpcaSCs3PC2.mpg

PC3: fat & thin

Show CorpColl\CCFpcaSCs3PC3.mpg

PCA for Shapes: E.g. 4: Corpora Callosa Data

PCA part 2: projections

New goal: discrimination (classification)

Projected data now shown as dots (not lines), colored as:
Schizophrenics Controls

Hope: two well separated clusters

Reality: didn't happen (but 80-d space is very large!)

E.g. 4: Corpora Callosa Data

Alternate view: Parallel coordinates

Show CorpColl\CCFParCorAlls3.ps

- Top: lots of common structure (mean is large component)

- Middle: large “dynamic range”, expected from Fourier decomp. of smooth signal.

- Bottom: non-Gaussian in direction of kurtosis

E.g. 4: Corpora Callosa Data

Discrimination by parallel coordinates?

Show CorpColl\ CCFParCorSCs3.ps

- not helpful

- red looks dominant: overplot problem

- conclude: parallel coordinates not a very useful view

Fisher Linear Discrimination

Idea: separate subpop'ns by "diff'nce between sample means"

Improvement: take covariance structure into account

show: HDLSS\ HDLSSoldDisc1.ps

Corpora Callosa application:

Show: CorpColl\ CCFfldSCs3.mpg

- Great separation of subpopulations?!?
- Image doesn't change when marching along vector?

Corpora Callosa Fisher Linear Discrimination

Major problem: $n = 71 < 80 = d$:

- gives “directions of perfect separation” (~8 dim subspace!)
- $\exists$ a very small change in this direction (watch pixels)
- numerics: use pseudo-inverse of covariance matrix
- is FLD direction interesting or useful?

Corpora Callosa Fisher Linear Discrimination (cont.)

Zoom in on FLD direction:

Show: CorpColl\CCFfldSCs3.mpg

- Only pixel sampling artifacts

- Expect big changes with new data

- Direction neither useful nor insightful

- A source of difficulty is means very close

Show CorpColl\CCFmeanSCs3.ps

Corpora Callosa Discrimination

Alternate approach: “Orthogonal Subspace Projection”

Will develop method later, for now see results

Show: CorpColl\CCFospSCs3RS11o2.mpg and CorpColl\CCFospSCs3RS12o1.mpg

- seems to find “real shape difference”
- is this effect really there?
- I.e. Is it stable with respect to new data?
- Is it useful?

Big Picture

I. Data examples (curves, images, shapes)

II. PCA for Visualization

III. FLD for discrimination

Now look more carefully (but still heuristically) at:

a. Robust PCA for cornea data

b. Orthogonal Subspace Projection for Corpora Callosa data

Cornea Data

Show CorneaRobust\NORMLWR.MPG

PCA gave good insights

Show CorneaRobust\NORM100.MPG, CorneaRobust\NORM200.MPG, CorneaRobust\NORM300.MPG

But e.g. PC2 may have been affected by outliers

Naïve approach: “outlier deletion”

Problem: >4 outliers (> 10% of data)

Robust Statistics

Major dichotomy:

View 1: Outliers are “bad data”, delete them

View 2: Outliers have problems, but also “contain useful info”,
So control their “influence”

E.g. the mean “feels outliers strongly” (“breakdown pt.” = 0)
The median allows outliers to only vote
(“breakdown pt.” = 50%)

Source of major (unfortunately bitter) debate!

Robust PCA

Approaches in literature:

1. Projection pursuit: idea replace “variance” in PCA optimization problem by “robust measure of spread”

   Problem: non-quadratic optimization $\Rightarrow$ slow to compute for high d (> 4 or 6, but we have 66)

2. Robust covariance matrix estimation

   Problem: existing methods assume “affine invariance”, which requires $n > d$

Robust PCA for $d > n$

First problem (previously ignored):

Sample mean $\bar{\underset{\rightarrow}{x}}$ can be seriously affected by outliers

Show CorneaRobust\OutliersMean.ps

Fix by using "median"?

What is "multivariate median"?

Multivariate Medians

(generated from different characterizations of univariate median)

i. Coordinate-wise median: often worst
(can lie on convex hull of data)

ii. Simplicial depth: slow to compute
(idea: measure “paint thickeness” of $d+1$ dim “simplices” with corners at data)

iii. Huber’s L^1 M-estimate:
(idea: project data on sphere, move sphere to make avg. of projected data at center)

Show CorneaRobust\L1Center.ps