Statistical Analysis of

High Dimension, Low Sample Size

Data

J. S. Marron

Department of Statistics
University of North Carolina

Special thanks to S. Ho and G. Gerig, UNC Computer Science
& C. Perou, Lineberger Cancer Center

Functional Data Analysis

Ramsey and Silverman(1997) *Functional Data Analysis*

The “atom” of the statistical analysis

Statistical Context	Atom
1^{st} Course	Number
Multivar. Analysis	Vector
F. D. A.	Complex Object (curve, image, shape, …)

Data Representation

Object Space $\leftrightarrow$ Feature space

Curves

Images

Shapes

Vectors

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

Data Representation (cont.)

E.g. 1: curves as data (Ramsay and Silverman)

Show CurvDat\ParabsRaw.ps

Object Space $\rightarrow$ Feature Space via “digitization”

Feature Space $\rightarrow$ Object Space via “Parallel Coordinates”

E.g. 2: Shapes of Corpora Collosa

Show CorpColl\CCFrawAlls3.mpg

E.g. 3: Genetics Micro Arrays

Show GeneArray\MicroArray.jpg & GA5IntrinsicRaw.ps

Data Representation, (cont.)

1. Landmarks: Bookstein, Dryden & Mardia

   - very slippery, e.g. Corpora Collosa data

2. Fourier Boundary Representation

   - Corpora Collosa data: use 80-dim'al basis

show CorpColl\CCFappFourAlls3C2.mpg

3. Medial Representations: Pizer, Yushkevich & Co.

show Ccmrep\CCMnormRaw.cfm animate, overlay, tile

Functional Data Analysis

Common Goal 1: “understand population structure”

Toy Example: curves as data points

Again Show CurvDat\ParabsRaw.ps

Useful analysis approach: Principal Component Analysis

Show CorneaRobust\SimplePCAeg.ps

PCA in Feature Space $\rightarrow$ Visualization in Object Space

- gives “decomposition of structure

show CurvDat\ParabsCurvDat.ps

Functional Data Analysis (cont.)

PCA Toy examples (cont.)

- finds clusters

show CurvData\ParabsUpDnCurvDat.ps

Real Example: Corpora Collosa data

Show CorpColl\CCFpcaSCs3PC1.mpg & CCFpcaSCs3PC2.mpg & CCFpcaSCs3PC3.mpg

  - shows "shape components of variation"

Functional Data Analysis (cont.)

Common Goal 2:   Discrimination (classification)

Given groups – find rule for "assigning new data to groups"

e.g.   disease diagnosis, based on measurements

Corpora Collosa data: Schizophrenics vs. Controls

Show CorpDoll\CCFrawSs3.mpg & CCFrawCs3.mpg

See the difference?

FDA for Medical Images

An early reference:

Cootes, Hill, Taylor, and Haslam (1993) in *Information Processing in Medical Imaging*, (H. H. Barret and A. F. Gmitro, eds.), **Springer Lecture Notes in Computer Science 687**, 33-47.

Common Problem: $n << d$

High Dimension Low Sample Size

Corpora Callosa: $n = 71 < 80 = d$

Trend: 3-d shapes, worse in both directions

Show Stat321FDA\GreggTracton.html

HDLSS Statistical Analysis

A “land of opportunity” for:

- Statisticians
- Probabilists
- …

1st Question: motivation for this?

Medical Imaging: YES

Gene Micro Arrays: YES

HDLSS Statistical Analysis (cont.)

Further motivation: Polynomial embedding

Idea from Statistical Pattern Recognition:

Embed data in higher dimensional space
to improve performance of linear discrimination

A teaser example: “the donut”

Show PolyEmbed\PEdonFLDcombine.pdf

2nd Question: How do we think about HDLSS data?

Old Conceptual Model

Projections into 1, 2 or 3 dimensions,

Show HDLSSoldCMod1.ps

Using:

- Coordinates
- Principal Components
- …

Nature of HDLSS Gaussian Data

For d dim'al "Standard Normal" dist'n:

$$\underset{\rightarrow}{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} \sim N\left(\underset{\rightarrow}{0}, I\right)$$

Euclidean Distance to Origin:

$$\left\|\underset{\rightarrow}{Z}\right\| = \sqrt{d} + O_p(1)$$

as $d \to \infty$.

Conclusion: data lie roughly on surface of sphere of radius $\sqrt{d}$

Nature of HDLSS Gaussian Data (cont.)

Paradox:

- Origin is point of highest density

- Data lie on “outer shell”

Nature of HDLSS Gaussian Data (cont.)

Lessons:

- High dim'al space is "strange" (to our percept'l systems)
- "density" needs careful interp'n (high d space is "vast")
- Low dim'al proj'ns can mislead
- Need new conceptual models

Nature of HDLSS Gaussian Data (cont.)

High dim'al Angles:

For any (fixed or indep. random) $\underset{\rightarrow}{x}$,

$$Angle(\underset{\rightarrow}{Z}, \underset{\rightarrow}{x}) = 90^\circ + O_p\left(\frac{1}{\sqrt{d}}\right)$$

Lessons:

- High dim'al space is vast (where do they all go?)
- Low dim'al proj's "hide structure"
- Need new conceptual models

A New Conceptual Model

Data lie in “sparse, high dim’al ring”

Show HDLSSnewCMod1.mpg

What about non-Gaussian data?

Personal View: OK, to build ideas in Gaussian context, if they “work outside”

e.g. PCA

Corpora Colosa: non-Gaussian (via Parallel Coordinate Plot)

Show CorpColl\ CCFParCorAlls3.ps

So What?

- What does this “new model” bring us?

e.g. Discrimination (i.e. Classification)

Disclaimers:

  - Will develop a new (?) method (hopefully fun)
  - Please suggest other approaches

So What? (cont.)

Corpora Colosa: Separate

"Schizophrenics" from "Controls"

$n = 40$ $n = 31$

clearly HDLSS, since $d = 80$

Again show CorpColl: CCFrawSs3.mpg & CCFrawCs3.mpg

Naïve Approach

PCA:

- hope: find “separated clusters”

Again show CorpColl\: CCFpcaSCs3PC1.mpg, CCFpcaSCs3PC2.mpg & CCFpcaSCs3PC3.mpg

Result:

- Poor “separation” of subpop’ns

Classical Multivar. Analysis:

Fisher Linear Discrimination:

Idea: Look at “direction separating means”, then “adjust for covariance”.

Show HDLSSod1Raw.ps, HDLSSod1PCA.ps, HDLSSod1Mdif.ps & HDLSSod1FLD.ps

HDLSS Implementation:

Use pseudo-inverse

Fisher Linear Discrimination

Results:

- Excellent separation of subpop'ns

Show CorpColl\CCFfldSCs3.mpg

- but useless answer

Show CorpColl\CCFfldSCs3mag.mpg

Why did Fisher fail?

Reason 1: data in 71d Space, so ∃ many "80d separating hyperplanes"

(and they are "very noisy")

Bootstrap "visual stability":

Show CorpColl\ CCFfldSCs3VisStab.mpg

Reason 2: Means are "too similar"

- Need to focus on cov. structure

Show CorpColl\ CCFmeanSCs3.ps

Solution based on new model

Show HDLSSnewDisc1.mpg

Approach: “Orthogonal Subspace Proj’n”

Idea: exploit vast size of high dim’al space.

Key on “subspaces generated by data”

(note: useless idea for large data sets, or low dimensions)

Orthogonal Subspace Projection

Toy example (in 2d)

Show SubSpProj\EgSubProj1Raw.ps

Idea: Project Data in Class 2, onto subspace gen'd by Class 1

Show EgSubProj1.ps

1^{st} Discrim. Dir'n is 1^{st} Eigenvector of projected data.

Corpora Collosa Example:

- Good Discrimination – Finds useful directions

Show CorpColl\CCFospSCs3RS11o2.mpg & CCFospSCs3RS12o1.mpg

- Visually Stable

Show CorpColl\CCFospSCs3RS11o2VS.mpg & CCFospSCs3RS12o1VS.mpg

- Poor “relabelling error rate”…

Show CCFospSCs3RS1stab.ps