Statistical Analysis of

High Dimension, Low Sample Size

Data

(Subtitle: Functional Data Analysis)

J. S. Marron

Department of Statistics
University of North Carolina

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

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

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

E.g. Corpora Collosa

Show CorpColl\CCFrawAlls3.mpg

An FDA Goal: population "structure"

I. "center"

e.g. "mean": vector $\rightarrow$ shape

Show CorpColl\CCFpcaSCs3PC1.mpg Mean Only

II. "spread"

PCA (Principal Component Analysis): "directions of max. var."

Show CorneaRobust\SimplePCAeg.ps

How to view eigenvector? $\rightarrow$ "march through shapes"

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

PCA Aside

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)

Others???? (I am collecting….)

HDLSS Statistical Analysis

Common Medical Imaging 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

1st Question: motivation for this?

Medical Imaging: YES

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?

Another FDA goal: 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 $d = 80$

Show CCFrawSs3.mpg and CCFrawCs3.mpg

Naïve Approach

PCA:

- hope: find “separated clusters”

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 HDLSSoldDisc1.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

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

Show Toy Data in SubSpProj\EgSubProj1Raw.ps

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

Show EgSubProj1.ps

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

Corpora Collosa Example:

- Finds useful directions

Show CCFospSCs3RS11o2VS.mpg and CCFospSCs3RS12o1VS.mpg

- Shaky “relabelling error rate”…