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

Aside on Smoothing

Interesting Research Directions?

"Old Question": How should we estimate, and how good is that?

"Modern Q": Which features in a smooth are "really there"?

{SiZer, SSS (Signif. In Scale Space)}

Aside on Smoothing (Cont.)

Two Provocative Statements:

1. Bandwidth selection is not as important as I once thought
2. Confidence bands are the wrong way to measure "variability" in curve estimators.

{SiZer, SSS (Signif. In Scale Space)}

Functional Data Analysis

Ramsey and Silverman(1997) *Functional Data Analysis*

The "atom" of the statistical analysis

<u>Statistical Context</u>	<u>Atom</u>
1st 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

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 & Co.

show Stat321FDA\PaulYMrepRaw2.png & PaulYMrepFine2.png

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

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 $d \rightarrow \infty$.

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 $d = 80$

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

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

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:

- Good Discrimination

Show top 2 rows of SubSpProj\ccf25d3sp1p1.ps

- Visually Stable

Show CCFospSCs31o2.mpg and CCFospSCs32o1.mpg

- Finds useful directions

Show CCFospSCs3RS11o2VS.mpg and CCFospSCs3RS12o1VS.mpg

- Poor “relabelling error rate”…

Show CCFospSCs3RS1stab.ps

Future work:

Atoms (of the FDA) are “trees”

Again show Paul Y trees and Tracton’s rep’,n