ORIE 779: Functional Data Analysis

From last meeting

Fisher Linear Discrimination

- Mahalanobis distance view
- Likelihood view
- Generalized to Gaussian Likelihood ratio
- Generalized to “uneven weights”
- Generalized to multiple classes
- I.e. Principal Discriminant Analysis
- Corpora Callosa data (failed because of…)

High Dimension Low Sample Size Statistical Analysis

Last Time: Fisher Linear Discrimination

Corpora Callosa application:

Recall data: Schizophrenics Controls

Movie display of FLD direction vector and projections

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

Last Time: 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?

Last Time: Corpora Callosa Fisher Linear Discrimination (cont.)

Zoom in on FLD direction:

- Only pixel sampling artifacts
- Expect big changes with new data
- Direction neither useful nor insightful

Last Time: Big Picture View

This motivate new area of statistical analysis:

High Dimension - Low Sample Size (HDLSS)

Idea: face common Problem: $n << d$

Last Time: Standard Approach to HDLSS

Dimensionality Reduction

Example: Medial Representation of Corpora Callosa data

No longer had HDLSS, since $d = 20 < n = 31,40$

But still FLD gave similar poor performance

Maybe not "far from HDLSS"?

Rethink Big Picture Views of FLD

Classical View (assumes $n >> d$):

- have "good estimates" of $\underrightarrow{\mu}$ and Σ
- Thus "instability of estimation" is negligible
- FLD works when Mean Difference does [toy example]
- But Mean Diff. can fail when FLD works [toy example]
- So FLD is *always recommended* (no loss, potential gain)
- This idea is *pervasive* in statistical (and beyond) folklore

Rethink Big Picture Views of FLD (cont.)

HDLSS view:

- Gap in above argument is unstable estimation
- FLD very unstable for $n < d$
- And appears unstable for $n \geq d$, but $n \approx d$
- Thus FLD *might* lose out to Mean Difference

Interesting Research Questions:

"Boundaries" between HDLSS and classical analyses???

Possible to develop diagnostics?

General Trends in FDA

Try to draw "big picture trends" from:

Some personal examples of HDLSS contexts

Cornea Data: $n = 42 < 66 = d$

Corpora Callosa (Fourier B'dry Rep'n): $n = 71 < 80 = d$

Genetic Micro-arrays: $n = 78 < 459 = d$

General Trends in FDA (cont.)

Towards Higher Dimensions:

- Research tending towards more complex “data objects”
- Appetite grows with capability (and understanding)

Towards Lower Sample Sizes:

- More complex data objects more costly too acquire
- Price comes down, but not as fast as above growth

General Trends in FDA (cont.)

Personal Conclusions:

- Neither trend will end soon
- Foolish to insist on "dimension reduction"
- Critical to learn to analyze HDLSS data
- HDLSS is a research "Land of Opportunity"
- Reinvention of most of multivariate analysis is needed

Will now give one example of this….

Old Conceptual Model for HDLSS data

Projections into 1, 2 or 3 dimensions [toy graphic]

(where our perceptual systems work),

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:

$$\left\| \underset{\rightarrow}{Z} \right\| = \left(\sum_{j=1}^{d} Z_j^2 \right)^{1/2} \sim \left(\chi_d^2\right)^{1/2}$$

$$\left\| \underset{\rightarrow}{Z} \right\| = \left(d + \sqrt{2d} \cdot O_p(1)\right)^{1/2}$$

(recall: $E\chi_d^2 = d$ and $\mathrm{var}\left(\chi_d^2\right) = 2d$)

Nature of HDLSS Gaussian Data (cont.)

So (for $\underset{\rightarrow}{Z} \sim N\left(\underset{\rightarrow}{0}, I\right)$), as $d \rightarrow \infty$,

$$\left\|\underset{\rightarrow}{Z}\right\| = \left(d\left(1 + d^{-1/2} O_p(1)\right)\right)^{1/2} = \sqrt{d}\left(1 + d^{-1/2} O_p(1)\right)^{1/2}$$

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

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

Nature of HDLSS Gaussian Data (cont.)

Paradox:

- Origin, $\underset{\rightarrow}{0}$, 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 (hi dim'al space is "vast")

  (mass of "solid ball" is "concentrated near boundary")

- *Nobody* is anywhere near "average in all respects"   ?!?
- 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 independent random) $\underset{\rightarrow}{x}$,

$$Angle(\underset{\rightarrow}{Z}, \underset{\rightarrow}{x}) = \cos^{-1}\left(\frac{\langle \underset{\rightarrow}{Z}, \underset{\rightarrow}{x} \rangle}{\|\underset{\rightarrow}{Z}\| \cdot \|\underset{\rightarrow}{x}\|} \right) = \cos^{-1}\left(\frac{\sum_{i=1}^{d} Z_i x_i}{\|\underset{\rightarrow}{Z}\| \cdot \|\underset{\rightarrow}{x}\|} \right)$$

$$Angle(\underset{\rightarrow}{Z}, \underset{\rightarrow}{x}) = \cos^{-1}\left(O_p\left(d^{-1/2} \right) \right)$$

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

Nature of HDLSS Gaussian Data (cont.)

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” [toy graphic]

What about non-spherical data?

- suitably stretch axes?
- Still makes sense to think of:

  “data on surface of $d-1$ dim’l ellipse”???

A New Conceptual Model (cont.)

What about non-Gaussian data?

Personal View:

OK to build ideas in Gaussian context, if they "work outside"

e.g. PCA

Corpora Collosa: non-Gaussian (via Parallel Coord. Plot)

Yet PCA, "shows population structure" [PC1]

So What?

- What does this "new model" bring us?

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

Corpora Colosa: try to separate

Schizophrenics [graphics] from Controls [graphics]

$n = 40$ $n = 31$

clearly HDLSS, since $d = 80$

Recall Background:

PCA failed: data not in "separated clusters" PC1 PC2 PC3

Fisher Linear Discrimination Failed:

- means too close [graphic]
- singular covariance found useless directions

Problem 1: based on old conceptual model [graphic]

Problem 2: Must use "covariance structure", not means

Solution Based on New Conceptual Model

Idea: Want to separate “two sparse rings of data” [toy graphic]

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)

Subspace Projection

Toy Example:

Idea: Project Data in Class 2, onto subspace orthogonal to subspace generated by Class 1 [graphic]

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

Corpora Collosa Example:

Best visual result: [OSP 1 on 2] [OSP 2 on 1]

- Directions show "shape"?

Comparison? Try "X view":

- Separate: directions look "similar" [1 on 2 X] [2 on 1 X]
- Combined: really found anything useful here???

Subspace Projection (cont.)

Important Questions:

- Is this effect really there?

- I.e. Is it stable with respect to new data?

- Is it useful?

(some answers coming later)

An Aside on High Dimensions

Deep questions in probability:

- Are there general limiting results as $d \to \infty$?

- In particular, for non-Gaussian dist'ns (indep. only?)

- Distance to Origin $\sim \sqrt{d}$? Angles $\sim 90^\circ$

- Do data always "cluster along $d-1$ dim'al manifold"?

High Dimensional Space Is Strange

Example from Ed George:

1. Start with "unit cube" $\{\underset{\rightarrow}{x} : -1 < x_i < 1, i = 1,\ldots,d\}$

2. Inscribe spheres in "quadrants"

$$\{\underset{\rightarrow}{x} : 0 < x_i < v_i, i = 1,\ldots,d\} \quad \text{indexed by} \quad \underset{\rightarrow}{v} = \begin{pmatrix} \pm 1 \\ \vdots \\ \pm 1 \end{pmatrix}$$

3. Consider sphere centered at $\underset{\rightarrow}{0}$, tangent to others

4. How "big" is that sphere? [graphic in 2-d]

High Dimensional Space Is Strange (cont.)

Strange Properties of Unit Cube in d dimensions:

- Volume = 2^d
- Number of "faces" = $2d$
- Distance from $\underset{\rightarrow}{0}$ to face = 1
- Number of "vertices" = 2^d (vertices are the $\underset{\rightarrow}{v}$ above)
- Distance from $\underset{\rightarrow}{0}$ to vertex = $\sqrt{d}$
- Where is the "mass"?

High Dimensional Space Is Strange (cont.)

"Mass" of the Unit Cube in d dimensions:

- Consider uniform distribution on unit cube
- I.e. $\underset{\rightarrow}{U}$, where U_i are independent Uniform $(-1,1)$
- Marginal 2nd Moment: $EU_i^2 = \int_{-1}^{1} \frac{1}{2} u^2 du = \frac{1}{2} \frac{u^3}{3} |_{-1}^{1} = \frac{1}{3}$
- By C.L.T.: $\frac{1}{d} \sum_{i=1}^{d} U_i^2 = EU_i^2 + O_p\left(\frac{1}{\sqrt{d}}\right) = \frac{1}{3} + O_p\left(\frac{1}{\sqrt{d}}\right)$
- Euclidean distance to $\underset{\rightarrow}{0}$:

$$\left\|\underset{\rightarrow}{U}\right\| = \left(\sum_{i=1}^{d} U_i^2\right)^{1/2} = \left(d\left(\frac{1}{3} + O_p\left(d^{-1/2}\right)\right)\right)^{1/2} = \sqrt{\frac{d}{3}} + O_p(1)$$

High Dimensional Space Is Strange (cont.)

"Mass" of the Unit Cube in d dimensions (cont.):

- So "most of the mass" is $\sqrt{d/3} \approx 0.58\sqrt{d}$ away from $\underset{\rightarrow}{0}$
- Recall *farthest point* from $\underset{\rightarrow}{0}$ has distance $\sqrt{d}$
- And faces have distance 1 to $\underset{\rightarrow}{0}$
- Conclude "mass is mostly near vertices"???
- Careful: only $2d$, but 2^d vertices
- Suggests very strong potential for ICA as d grows

High Dimensional Space Is Strange (cont.)

Size of Inscribed Sphere:

- Centers of Quadrant Spheres: $\frac{1}{2}\underset{\rightarrow}{v}$
- Distance from center to $\underset{\rightarrow}{0}$: $\sqrt{d}/2$
- Radius of Quadrant Spheres: $1/2$
- Radius of Inscribed Sphere: $\left(\sqrt{d}/2\right)-1/2$
- Inscribed Sphere "pops out of face", for $d \geq 9$?!?!
- Quadrant Spheres "move out towards vertices" ?!?!
- Makes "mass of Unit Cube" effect seem plausible?