From last meetings

Goal 1: Understanding Population Structure

PCA: illustrated with Cornea Data

Goal 2: Discrimination (classification)

Corpora Callosa data

F. L. D. failed

Now derive "Orthogonal Subspace Projection"

Corpora Callosa Data

Show CorpColl\CCFrawAlls3.mpg, CorpColl\CCFrawSs3.mpg and CorpColl\CCFrawCs3.mpg

PCA: poor discrimination

Show CorpColl\CCFpcaSCs3PC1.mpg

Fisher Linear Discrimination: found useless, noise driven, dir'n

Show CorpColl\CCFfldSCs3.mpg and CorpColl\CCFfldSCs3mag.mpg

Key observation: means are very close

Show CorpColl\CCFmeanSCs3.ps

So to discriminate must use "covariance structure", not means

Background (for motivation)

New area of statistical analysis:

High Dimension - Low Sample Size (HDLSS)

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

Old Conceptual Model

Projections into 1, 2 or 3 dimensions
(where our perceptual systems work),

Show HDLSSoldCMod1.ps

Using:

- Coordinates
- Principal Components
- …

Nature of HDLSS Gaussian Data, I

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, II

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

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

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

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

Nature of HDLSS Gaussian Data, III

Paradox:

- Origin, $\underset{\rightarrow}{0}$, is point of highest density

- Data lie on “outer shell”

Nature of HDLSS Gaussian Data, IV

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")
- Low dim'al proj'ns can mislead
- Need new conceptual models

Nature of HDLSS Gaussian Data, V

High dim'al Angles:

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

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

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

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

Nature of HDLSS Gaussian Data, VI

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-spherical data?

- suitably stretch axes?
- Still makes sense to think of:
  “data on surface of 2-d manifold (ellipse)”???

A New Conceptual Model, II

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)

Again show CorpColl\CCFParCorAlls3.ps

Yet PCA, "shows population structure"

Show CorpColl\CCFpcaSCs3PC1.mpg

An aside

Deep questions in probability:

- Do data always "cluster along 2-d manifold"?
- Are there general limiting results as $d \to \infty$?
- Distance to Origin $\sim \sqrt{d}$? Angles $\sim 90^\circ$

So What?

- What does this “new model” bring us?

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

Corpora Colosa: try to separate

Schizophrenics from Controls

$n = 40$ $n = 31$

clearly HDLSS, since $d = 80$

Again show CorpColl\CCFrawSs3.mpg and CorpColl\CCFrawCs3.mpg

Recall Background:

PCA failed: data not in “separated clusters”

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

Fisher Linear Discrimination Failed:

- means too close
- singular covariance found useless directions

Again show CorpColl\CCFmeanSCs3.ps

Old conceptual model

Show HDLSSoldDisc1.ps

Solution based on new conceptual model

Idea: Want to separate “two sparse rings of data”

Show HDLSS\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)

Subspace Projection

Toy Example:

Show Toy Data in SubSpProj\EgSubProj1Raw.ps

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

Show SubSpProj\EgSubProj1.ps

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

Corpora Collosa Example:

Best visual result:

Show CorpColl/CCFospSCs3RS11o2.mpg and CorpColl/CCFospSCs3RS12o1.mpg

- Directions show “shape”

Comparison? Try “X view”:

- Separate: directions look “similar”

Show CorpColl/CCFospSCs3RS11o2X.mpg and CorpColl/CCFospSCs3RS12o1X.mpg

- Combined: really found anything useful here???

Show CorpColl\CCFospSCs3RS1bothX.mpg

Short Term Future Plans

a. Mathematics and Numerics behind PCA

b. Independent Component Analysis?

c. Goodness of Approximation?

d. Mathematics for Fisher Linear Discrimination

e. Validation of Discrimination

f. Polynomial Embedding and Support Vector Machines