From Last Meeting

Finished ICA

Analysis of Mass Flux data:

- Insights from “clustering”
- Explored “rotation of PCA directions”

Goodness of Approximation

I.e. how many basis elements to use?

E.g. Corpora Callosa data

Recall "shape representations" are based on
$d = 80$ dimensional "feature vectors"

Show CCFrawAlls3.mpg

How big does d need to be?

A personal working assumption:

"shape is complicated, so need d large"

Major sticking point

For medical image shapes, usually have "few data points",

$n < d$

Personal approach:

- that complicates matters
- but "shape" is "complex" and requires complex rep'n
- hence need to develop new statistical methods:

High Dimension Low Sample Size

Classical Approach

- Statistical Multivariate Analysis is based on "standardizing"

- Multiply by $\hat{\Sigma}^{-1/2}$ (for covariance matrix)

- Requires $n > d$ (else matrix inverse doesn't exist)

- For $n \le d$, do "dimension reduction"

- For example, keep only the "1st few Principal Components"

Questions:

Is dimension reduction (e.g. PCA based) “good enough”?

Or is it important to develop HDLSS methods?

Aside: how well do ANOVA sums of squares “capture shape”?

Study in context of corpus collosum data

Fourier Approximation Background:

Represent:

$$Shape = \sum_{j=1}^{d} c_j BE_j$$

where the c_j are the “Fourier Coefficients”

and where the BE_j are “basis element” shapes

Fourier Approximation Background (Cont.):

Problem: BE_j have “parametric representation”,

so hard to view individually

Solution: Interesting web site:

http://www.cs.unc.edu/~seanho/miggg/fourdem.html

show CorpColl\BdryFourDemo\fourdem.html

Some examples of generated shapes:

Show CorpColl\CCFbasis.ps

Approximation 1: Raw Fourier Coefficients

View “goodness of approximation” of

$$k - approx.\,Shape = \sum_{j=1}^{k} c_j BE_j$$

for $k = 0,1,2,\ldots,d$

show CCFappFourAlls3C4.mpg

- $k = 0$ single point: the “zero function”
- $k = 1$ just a line
- $k = 2,3$ still a line (due to “shape normalization”)

Approximation 1: Raw Fourier Coefficients (cont.)

- $k = 4$ ellipse

- $k > 4$ more complicated shapes

- larger k get convergence towards full shape

- $k = 80 = d$ blue completely covers white

Approximation 1: Raw Fourier Coefficients (cont.)

ANOVA style Sums of Squares:

$$Signal\ Power(k-Approx.) = \sum_{j=1}^{k} c_j^2$$

Measures “goodness of fit”, on scale of “energy”

Energy decomposition: c_j^2 is “power in signal in direction BE_j”

Show upper left of CCFappFourAlls3.ps

Approximation 1: Raw Fourier Coefficients (cont.)

Useful scales:

- log scales

Show bottom row of CCFappFourAlls3.ps

- relative scale: $c_j^2 / \sum_{j'=1}^{d} c_{j'}^2$

Show center of CCFappFourAlls3.ps

- cumulative relative scale: $\sum_{k=1}^{k} c_k^2 \Big/ \sum_{j'=1}^{d} c_{j'}^2$

Show right of CCFappFourAlls3.ps

Approximation 1: Raw Fourier Coefficients (cont.)

What does “cumulative relative signal power” really measure?

Again show CCFappFourAlls3C4.mpg

- $k = 2$ line alone is 93%
- $k = 6$ nearly elliptical is 95%???
- $k = 12$ 99%, but still “misses lots of shape”
- $k = 25$ 99.9%, still don’t have all of this “shape”?

Have looked at some others: similar lessons

Approximation 2: Centered Fourier Coefficients

Main idea: subtract out the mean first

- standard in ANOVA (often huge part of Sums of Squares)
- results in much different interpretation (of <u>relative</u> SS)

When is "90% of SS explained"?

- Case 29: 31 terms: all of shape

show CCFappCFourAlls3C3.mpg

- Case 2: 11 terms: missed a lot of shape

show CCFappCFourAlls3C1.mpg

Approximation 2: Centered Fourier Coefficients (Cont.)

Paradox of cumulatives ("data compression" plots):

- Case 2 has "great compression" (high curve), yet needs ~50 terms (99.8% explained) for "good shape rep'n"

- Case 29 has "poor compression" (low curve), yet needs only ~32 terms (92.53% explained) for "good shape rep'n"

Personal conclusion:

"shape" manifestations of Sum of Square Analysis is "slippery"

Approximation 3: Principal Component Analysis

Recall Ideas:

- Find "directions of greatest variability"
- Will "maximize signal compression"
- Works in an "average sense", not individually
- Use 1st k for "dimensionality reduction"

Approximation 3: Principal Component Analysis (cont.)

Overlay of cumulatives:

Show CCFappPCAAlls3.ps

- cumulative eigenvalues ("average") shown in yellow

- much better signal compression than centered Fourier

flip back to CCFappCFourAlls3.ps

- colored cases are extremes of signal compression:

- Case 2 is "great", Case 13 and Case 29 are "poor"

- Case 35 is "closest to average"

Approximation 3: Principal Component Analysis (cont.)

How well does “90%” capture “shape”?

- Case 2: poor (happens at $k = 1$)

Show CCFappPCAAlls3C1.mpg

- Case 13 and Case 29 good (happens at $k = 16$ and $k = 17$)

Show CCFappPCAAlls3C2.mpg and CCFappPCAAlls3C3.mpg

- Case 35 not quite (happens at $k = 6$)

Show CCFappPCAAlls3C4.mpg

- k is more useful than “% variability”?

Approximation 3: Principal Component Analysis (cont.)

How many terms are needed to capture shape?

- Case 2: $k = 17$?

Show CCFappPCAAlls3C1.mpg

- Case 13 $k = 15$?

Show CCFappPCAAlls3C2.mpg

- Case 29 $k = 16$?

Show CCFappPCAAlls3C3.mpg

- Case 35 $k = 15$?

Show CCFappPCAAlls3C4.mpg

Personal conclusions

- “Sums of Squares” are very crude surrogate for “shape”
- Not enough to “just work with 1st k PCs”
- Not enough to “just work with PCs with top 95% of signal”
- Careful about “average fit” (as in PCA), vs. “individuals”
- 15 – 20 PCs “captures shape for Corpus Callosum data”
- Expect more needed for higher dim’nal objects

Show GreggTracton.html

Still worth developing HDLSS