From last meeting

Class Web Page:
http://www.stat.unc.edu/faculty/marron/321FDAhome.html

Functional Data Analysis: what is the “atom”?

Goal I: Understanding “population structure”.

Important duality:

Object Space $\leftrightarrow$ Feature Space

Powerful method: Principal Component Analysis

Principal Component Analysis (PCA)

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)

PCA, II

There are many applications / viewpoints:

- dimension reduction (statistics / data mining)
- change of basis (linear algebra)
- transformation (statistics)
- data compression (electrical engineering)
- signal denoising (acoustics / image processing)
- optimization (operations research)

PCA, Optimization View

Find “direction of greatest variability”

Show HierArch\HierArchEG1d0p2.mpg and HierArch\HierArchEG1d0p4.mpg

1. Center Point: Sample Mean: $\underset{\rightarrow}{\bar{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_d \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{id} \end{pmatrix}$,

Aside: “mean vector” = “vector of means” is not obvious

2. Work with re-centered data: $\underset{\rightarrow}{x}_i - \underset{\rightarrow}{\bar{x}}, \quad i = 1, \ldots, n$
“mean residuals”

PCA, Optimization View, II

3. Consider all possible “directions”

4. Project (find closest point) data onto direction vector

5. Maximize “spread” (sample variance), by choice of direction

Show CorneaRobust/SimplePCAeg.ps

6. Project data onto orthogonal subspace, and repeat.

PCA, Optimization View, III

Results:

Again show HierArch\HierArchEG1d0p2.mpg and HierArch\HierArchEG1d0p4.mpg

- “directions of greatest variability”

- “natural coordinate axes”

- “maximal 1-d descriptions of data”

PCA for curves

E.g. 1: "Dog Legs" (simulated example)

Show CurvDat\DogLegsRaw.ps

Note: since $d = 3$, have direct "point cloud" visualization

Show CurvDat\DogLegs3d.ps

PCA:

Show CurvDat\DogLegsCurvDat.ps

- Plot 1,1: Raw data
- Plot 1,2: Center point, i.e. mean vector, i.e. average curve
- Plot 1,3: Mean Residuals, i.e. re-centered point cloud

PCA for curves,   E.g. 1: "Dog Legs"

- Plot 1,4: discussed later

- Plot 2,1: Projections (centered) data onto PC1
  (recall object $\leftrightarrow$ feature duality)
  shows "dominant component of variability"

- Plot 2,2: "Extremes view", on original (not re-centered) scale

- Plot 2,3: Residuals, i.e. data – projection
  i.e. projection onto orthog'l subspace

Again show CorneaRobust/SimplePCAeg.ps

- Plot 2,4: kernel density estimate (smooth histogram) of
  projections (say more later)

PCA for curves, E.g. 1: "Dog Legs" (cont.)

- Plots 3,1-4: Same for 2nd PC
  orthogonal to first
  captures different mode of variability
  less visual variability

- Plots 4,1-4: Same for 3rd PC
  yet another mode
  even less visual variability
  residuals are 0 (since $d = 3$)

Overall: Decomposition of "complex variability" into several simple (thus interpretable) pieces.

PCA for curves, E.g. 1: “Dog Legs” (cont.)

Sum of squares analysis

Idea: quantify “decreasing visual variability”

Statistics: ANOVA (ANalysis Of VAriance)

Signal Processing: “energy”

PCA for curves, E.g. 1: Sum of squares

Total Sum of Squares (energy): $\sum_{i=1}^{n}\sum_{j=1}^{d} x_{ij}^{\;2}$

Mean Sum of Squares: $\sum_{i=1}^{n}\sum_{j=1}^{d} \bar{x}_i^{\;2}$ (= 62% of total)

Mean Resid'l Sum of Sq's: $\sum_{i=1}^{n}\sum_{j=1}^{d}\left(x_{ij}-\bar{x}_i\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{d} x_{ij}^{\;2} - \sum_{i=1}^{n}\sum_{j=1}^{d} \bar{x}_i^{\;2}$

(Pythagorean theorem, = 38% of total)

Again show CorneaRobust/SimplePCAeg.ps

PCA for curves, E.g. 1: Sum of squares (cont.)

Decomposition of Mean Residual sum of squares:

Sum PC1 + Sum PC2 + Sum PC3
(Parseval's identity)
(Distribution of "energy")

Quantification of visual impression:

SS, PC1 = 92% of MR,	SS Resid = 8% of MR
SS, PC2 = 7% of MR,	SS Resid = 1% of MR
SS, PC3 = 1% of MR,	SS Resid = 0% of MR

Visual comparison: upper right

PCA for curves (cont.)

E.g. 2: “Parabolas” (simulated data set)

Show CurvDat\ParabsRaw.ps and CurvDat\ParabsCurvDat.ps

Similar display, main lessons:

i. Mean: where “parabolic part” appears (90% of Total SS)

ii. Mean Residuals: “random curves”????

iii. PC1: variability of “vertical shift” type (86% of MR SS)
(not obvious from mean residuals?)

iv. PC1 residuals: much less (only 14% of MR SS)
(recall projection of orthogonal subspace)

PCA, E.g. 2: “Parabolas”

v. PC2: Variability of “tilt” type (10% of MR SS)
(really can’t “see this in data”!)

vi. PC2 residuals: even less (only 3% of MR SS)

vii. PC3: “random noise” (only 0.7% of MR)

viii. PC3 residuals: contains “most of energy” of above

ix. PC4: similar to PC3, no more interesting structure

Overall: Intuitive decomposition of “population structure”, shows features invisible in full data set.

PCA for curves (cont.)

E.g. 2: "Up and Down Parabolas" (simulated data set)

Show CurvDat\ParabsUpDnRaw.ps

Idea: why are smoothed histograms of projections useful?

Form of data: 2 "clusters"

PCA:

Show CurvDat\ParabsUpDnCurvDat.ps

PC1: finds "clusters (93% of variability, see smooth histo's)

PC2: "vertical shift" (note some of that also in PC1)

(no guarantee that "right" features are found)

PC3: "tilt" (almost all variability explained now)

PCA for Images:

E.g. 3: Cornea Data

Again show CorneaRobust\NORMLWR.MPG

PCA: can find direction of greatest variability

Again show CorneaRobust/SimplePCAeg.ps

Main problem: display of result (no overlays for images)

Solution: show movie of “marching along the direction vector”

Show CorneaRobust\NORM100.MPG

PCA for Images, E.g. 3: Cornea Data

PC1:

Mean: mild vertical astigmatism
(known population structure called "with the rule")

Main direction: "more curved" & "less curved"
(corresponds to first optometric measure)

Also: "stronger astigmatism" & "no astigmatism"

Note: found correlation between astigmatism and curvature

Projections (blue lines): Looks like Gaussian (Normal) dist'n

PCA for Images, E.g. 3: Cornea Data

PC2:

Show CorneaRobust\NORM200.MPG