ORIE 779: Functional Data Analysis

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”.

From last meeting (cont.)

Important duality:

Object Space $\leftrightarrow$ Feature Space

Powerful method: Principal Component Analysis

Recall many names:

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)
Geo-Sciences: Empirical Orthogonal Functions (EOF)

From last meeting (cont.)

Recall 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

Goal: find "direction of greatest variability"

[Spinning point Cloud] - [Axis of greatest variability]

Visual Aside: *Motion* helps "understand" 3-d data in 2-d environment

Question: "direction" from where?

PCA, Optimization View (cont.)

Step 1: Start with Center Point:

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

Aside: "mean vector" = "vector of means" is not obvious

Notation: "under-arrow" used for vectors

PCA, Optimization View (cont.)

Step 2: Work with re-centered data:

$\underset{\rightarrow}{x}_i - \overline{\underset{\rightarrow}{x}}, \quad i = 1, \ldots, n,$ the "mean residuals"

Step 3: Consider all possible "directions"

Step 4: Project (find closest point) data onto direction vector

Step 5: Maximize "spread" (sample variance), over direction

Step 6: Project data onto orthogonal subspace, and repeat.

PCA, 2-d Illustration

Reasons:

- easy to see everything in 2-d
- build ideas that generalize to higher dimensions

Raw Toy Data: Graphic   Shifted, slanted Gaussian point cloud

Recall two views:

- "Point Cloud" (scatterplot in 2-d)
- "curves" (corresponding parallel coordinate plot)
- useful one to one correspondence

PCA, 2-d Illustration (cont.)

Steps 1 & 2: Recenter by sample mean

Graphic 1: Find the mean

- Looks like “the center”
- Mean in curve view shows “average of components”

Graphic 2: Find vectors from mean to data

- called “mean residuals”

PCA, 2-d Illustration (cont.)

Graphic 3: Subtract the mean, i.e. “shift mean to the origin”

- Now “direction” makes more sense
- Note change of axis in curve view
- Now have “mean zero” in both views

PCA, 2-d Illustration (cont.)

Interesting numerical comparison:

Quantify "how much shifting is done", using sums of squares

Terminology: Analysis of Variance (ANOVA)

- Decomposition of Sums of Squares
    - main substance of ANOVA
    - not hypothesis testing (as many think)
- Contains useful insights
- Interpret as "energy" or "signal power"

PCA, 2-d Illustration (cont.)

Graphic 4: Overlay sums of squares

- Total Sum of Squares ≈ 662
  - Squared lengths of black line segments
- Sum of Squares for Mean ≈ 606
- Squared length of green (times n)
- Fraction is ≈ 92%
- Terminology: “mean contains 92% of energy in signal”

PCA, 2-d Illustration (cont.)

ANOVA (cont.)

- Residual (from mean) Sum of Squares ≈ 55
  - Squared length of mean residuals
  - Fraction is ≈ 8%
  - "Mean Residuals contain 8% of total energy"

Aside: Nonzero means are often a large fraction of total variation. Thus conventional "R-squared Analysis" is defined with the mean subtracted everywhere.

PCA, 2-d Illustration (cont.)

Important point: this analysis "makes sense"

because of "Pythagorean Theorem":

$$SS_{total} = \sum_{i=1}^{n}\sum_{j=1}^{d}\left(x_{i,j}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{d}\left(x_{i,j} - \bar{x}_j + \bar{x}_j\right)^2 =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{d}\left(x_{i,j} - \bar{x}_j\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{d} 2\left(x_{i,j} - \bar{x}_j\right)\bar{x}_j + \sum_{i=1}^{n}\sum_{j=1}^{d}\left(\bar{x}_j\right)^2 =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{d}\left(x_{i,j} - \bar{x}_j\right)^2 + 0 + \sum_{i=1}^{n}\sum_{j=1}^{d}\left(\bar{x}_j\right)^2 = SS_{resid} + SS_{mean}$$

Power of L^2 ("Hilbert Space")

PCA, 2-d Illustration (cont.)

Pythagorean Theorem?   Where is the triangle?

- Put data in "space of concatenated vectors" $\begin{pmatrix} \underset{\rightarrow}{x}_1 \\ \vdots \\ \underset{\rightarrow}{x}_n \end{pmatrix}_{nd \times 1}$

- Mean is projection onto subspace

$$\left\{ \underset{\rightarrow}{x} : x_{1,1} = \cdots = x_{n,1}, \ldots, x_{1,d} = \cdots = x_{n,d} \right\}$$

- So mean residuals are orthogonal

- Triangle (in this big space) has vertices:
  - the origin
  - the big data vector
  - the big mean vector

PCA, 2-d Illustration (cont.)

Steps 3 & 4: Find "direction of greatest variability"

Graphic 5: Direction vector

- determines 1-d subspace
- i.e. line through origin

Graphic 6: Projections onto Direction vector

- Projection is nearest point in 1-d subspace

PCA, 2-d Illustration (cont.)

Step 5: Optimize direction for "greatest variability of project'ns"

Graphic 6: (answer was already shown above)

- major axis of "ellipse of data"
- most efficient 1-d representation of data
- minimizes length of residuals
- least squares solution to "line closest to data"
- Note "fairly close" in corresponding curve view

PCA, 2-d Illustration (cont.)

How "close"? Use ANOVA to quantify:

Graphic 7: Pieces of ANOVA

- Sum of Squares of recentered data ≈ 55
    - Sum of squared lengths of blue segments
    - Represents "energy of recentered data"

PCA, 2-d Illustration (cont.)

- Sum of Squares of Projected Data ≈ 51
  - Contains about 91% of relevant sum of squares
  - Shows this 1-d representation is a "good approx'tion"
  - This comparison is more useful than to total SS
- Sum of Squares of Resdiuals ≈ 4
  - Has only 7% of energy in recentered data
  - i.e. "little left over after 1-d approximation"

PCA, 2-d Illustration (cont.)

Alternate view: Orthogonal to direction of greatest variability

Graphic 8: Direction vector

Graphic 9: Corresponding residuals

- Current Residuals are previous Projections
- Current Projections are previous Residuals
- In 2-d this direction minimizes the variation
- Because of another Pythagorean Theorem
- Driven by orthogonality of directions

PCA, 2-d Illustration (cont.)

Effects of “poor 1-d representation” on Curve View:

- much worse approximation of data
- describes less of the structure in the data
- but contains some useful information
- “orthogonal direction’’ looks flat instead of slanted

Note ANOVA analysis uses same numbers

Graphic 10 but they “swap places” (in expected way)

PCA, 2-d Illustration (cont.)

Drawback to this type of visualization:

Useless for higher dimensions

Did this to build ideas, now extend insight to high dimensions

First revisit previous example, using only Curve View,

but summarize different views in single:

Curve View Graphic

PCA, 2-d Illustration (cont.)

Curve View Graphic: Approach to viewing PCA

Upper far left: Raw Data

- Colors allowing easy indentification across panels
- Curves are just line segments since only 2-d

Upper center left: Mean Vector

Upper center right: Mean Residuals

- this is difference of previous 2

PCA, 2-d Illustration (cont.)

Curve View Graphic (cont.)

Upper far right: Power plot

- Shows Fraction of Sum of Squares, in each direction
- Fractions shown in blue
- Cumulative Fractions shown in red
- Will make more sense for higher dimensions

Next Rows: Two directional projections

Middle Row: projection in dominant direction

PCA, 2-d Illustration (cont.)

Curve View Graphic (cont.)

Middle Row: projection in dominant direction (cont.)

Middle far left: all Projections represented as curves
- contains "a large amount of simple structure in data"
- "good one dimensional representation" (as noted before)
- Note: mean is not in this view

Middle center left: View as "mean +- extreme projections"
- Sometimes this view is more useful
- I.e. additional insight comes from including the mean
- Fraction of sum of squares appears here

PCA, 2-d Illustration (cont.)

Curve View Graphic (cont.)

Middle Row: projection in dominant direction (cont.)

Middle center right: Residuals from mean

- Family of curves above, minus far left
- orthogonal to far left
- this Fraction of SS also shown

Middle far right: projection coefficients (numbers)

- each dot is on coefficient
- color is linked to data curves
- random height (“jitter plot”) allows visual separation
- curve is “smooth histogram” (discussed more later)
- these look “quite Gaussian” (OK, since simulated that way)
- usefulness illustrated later

PCA, 2-d Illustration (cont.)

Curve View Graphic (cont.)

Bottom Row: projection in orthogonal direction (cont.)

Bottom far left: projections

- same as residuals (since only in 2-d)

Bottom center left: View as "mean +- extreme projections"

- Different impression from this orthogonal direction

Bottom center right: Remaining residuals

- Have subtracted only from residuals above
- Nothing left, since in 2-d
- Sum of Squares is 0, since "nothing left"

Bottom far right: projection coefficients (numbers)

- Again Gaussian (as expected)

PCA, 2-d Illustration (cont.)

Now try variations, to study differences:

Similar simulation, but with mean 0:

Raw Data   Curve View Graphic

- Raw data look like mean recentered from before
- Mean (upper center left) looks visually negligible
- Confirmed by very small SS (non-zero, since sim'd data)
- Directions (and ANOVA) all very similar to before

PCA, 2-d Illustration (cont.)

Simulated from spherical Gaussian

Raw Data   Curve View Graphic

- Neither Raw Data view shows much structure
- Directions not informative
- Just driven by luck of the draw
- Note Sums of Squares much more evenly split
- But somewhat different: again luck of the draw
- Power Plot (upper far right) shows this nicely