ORIE 779: Functional Data Analysis

From last meeting

Functional Data Analysis: what is the “atom”?

Important duality:

Object Space $\leftrightarrow$ Feature Space

Powerful method: Principal Component Analysis

Built Ideas in 2d (where can see everything)

PCA, Point Cloud View

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

- "directions of greatest variability"
- "natural coordinate axes"
- "maximal 1-d descriptions of data"

Red is 1^{st} PC (dominant direction)

Yellow is 2^{nd} PC (dominant direction in subspace ortho'l to 1^{st})

Cyan is 3^{rd} PC (dominant direction in ortho'l to 1^{st}two)

PCA, Curve View

Corresponding to above data: graphic

Top Row: Mean shift (as before)

2nd Row: Decomposition in 1st PC direction

3rd Row: Decomposition in 2nd PC direction

4th Row: Decomposition in 3rd PC direction

PCA for curves, 3d

E.g. 1: "Dog Legs" (simulated example) [curve view]

Guess "structure of population"?

- Mean like "v"?
- x_1 correlated with x_3?
- Intuitive content of dominant direction?

Since $d = 3$, try spinning point cloud view

- Can see "one direction will explain a lot of the data"?
- But "meaning in curve space"??? (x_1 correlated with x_3?)

PCA for curves, 3-d

PCA for Dog Legs: Curve View PCA

Mean: "somewhat tilted V" (~40% of SS)

PC1: "multiples of symmetric V" (~92% of MRSS)

- shows "x_1 correlated with x_3" is a very important aspect

PCA for curves, 3-d (cont.)

PCA for Dog Legs (cont.): Curve View PCA

PC2: "change *only* in x_2 direction" (~7% of MRSS)

PC3: "slants" (note: ortho to PC1 direction) (1% of MRSS)

Remaining Residuals: nothing, since in only 3-d

Note: overall intuitively & useful "decomposition of variation"

PCA for curves, 3-d (cont.)

A different 3-d example: Fans   curve data graphic

Again guess "population structure"?

- Mean is slanted line?
- x_3 has most variation?
- x_2 is correlated with x_3?

Again, since $d = 3$,   try spinning point cloud view

- data lie near "slab" (vs. "line" in Dog Legs e.g. above)

PCA for curves, 3-d (cont.)

PCA for Fans: Curve View PCA

Mean: Slanted Line (65% of SS)

PC1: Driven by x_3 variation, with x_2 correlated (86% of MRSS)

PC2: Part of x_2, that is independent of x_3 (13% of MRSS)

PC3: all x_1 variation, much smaller (1% of MRSS)

Verify in Spinning Point Cloud View and PC axes view

Note: "data lie in slab" reflected by large PC2 (than for dog legs)

PCA for curves

Now try higher dimension

- no more spinning clouds
- can only use curve view (but now know main ideas)

Toy Example "Random Parabolas": Raw data graphic

$n = 50$ curves in $d = 10$ dimensions, guess structure?

PCA for curves (cont.)

PCA for Parabolas: Curve View PCA

Mean: Captures *all* of the parabolic structure (90% of SS)

- dominant shape is *not* part of variation

PC1: Vertical shift (88% of MRSS)

- Can see that in raw data
- How about structure in PC1 residuals?

PC2: Tilt (10% of MRSS)

- *can't* see this in raw data

PCA for curves (cont.)

PCA for Parabolas (cont.): Curve View PCA

Remaining PCs:

- very small fraction of MRSS (see upper right Power plot)
- random directions?
- were simulated as I.I.D. Gaussians

Overall: Intuitive decomposition of “population structure”

- shows features invisible in full data set.

PCA for curves (cont.)

Interesting question: what is PCA for I.I.D. Gaussians?

Initial idea: $N(0, I)$ random vectors have

"spherically symmetric distribution"

So expect:

- random directions
- SS's evenly separated

PCA for curves (cont.)

Actual answers:

1. Directions are random

2. But SS's depend on sample size

Case 1: Small n: d = 10, n = 10 PCA Curve Graphic

- SS's are not constant, instead "fall off linearly"
- Clearly visible in Power Plot (upper right)
- Because data naturally "extend more in some directions"

PCA for curves (cont.)

Case 2: Large n: $d = 10$, $n = 200$ PCA Curve Graphic

- now SS's look much more constant
- but still some small decrease
- reason is more data $\Rightarrow$ more "large directions"

There is some mathematical theory for this:

Johnstone (2001) On the distribution of the largest principal component, *Annals of Statistics*, 29, 291-327, internet available at: http://www-stat.stanford.edu/~imj/Reports/2000/largepc.ps

PCA for curves (cont.)

One more toy data set: “2 Clusters” Raw Data Graphic

Goal: illustrate use of “smoothed histograms” (on right)

Form of data: 2 “clusters”

- widely separated subpopulations

Guess PCA?

- “Maximal variability” along “skewer between 2 meatballs”?

PCA for curves (cont.)

PCA for 2 Clusters Data: Graphic

Mean: negligible (only 2% of SS)

PC1: Clearly captures 2 clusters (93% of MRSS)

- visible in projection plot (far left)
- and also in jitter plot & smooth histo. (bimodal pop'n)

PCA for curves (cont.)

PCA for 2 Clusters Data (cont.): Graphic

PC2: part of vertical shift, *but not all* (4% of MRSS)

- since vertical shift *not quite* orthogonal to PC1 direction
- no guarantee that PCA finds “right” directions
- only “orthogonal directions of greatest variability”
- recall vertical shift was PC1 above (less “important” now)

PCA for curves (cont.)

PCA for 2 Clusters Data (cont.): Graphic

PC3: Tilt (2% of MRSS)

- this was PC2 before
- feature of population that is not visually apparent

Remaining PCs: negligible, just Gaussian noise

PCA for curves (cont.)

Potential Problem:

PCA directions different from "interesting directions"

Generally: a very challenging problem for future work

A first simple solution: VARIMAX from Section 6.3.3 of Ramsey and Silverman (1997)

(a good topic for student presentation)