ORIE 779: Functional Data Analysis

Note: time to schedule remaining Student Presentations

From last meeting

Independent Component Analysis

Idea: Find "directions that maximize independence"

Studied:

- Toy signal processing examples
- Toy FDA examples
- Local minima & non-linearities

Last Time: Careful look at Kurtosis

Recall for standardized (mean 0, var 1) data: $Z_1, \ldots, Z_n$,

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n} Z_i^4 - 3$$

- for $Z_i \sim N(0,1)$, Kurtosis = 0
- Kurtosis "large" for high peak, low flanks, heavy tails?
- Kurtosis "small" for low peak, high flanks, light tails?
- Can show Kurtosis $\geq$ -2 (point masses at +-1)
- Thus very "asymmetric"? (see above examples)

Last Time: Careful look at Kurtosis (cont.)

E.g. three point distribution, with probability mass function:

$$f_w(x) = \begin{cases} \frac{1-w}{2} & x = \frac{-1}{\sqrt{1-w}} \\ w & x = 0 \\ \frac{1-w}{2} & x = \frac{1}{\sqrt{1-w}} \end{cases}, \quad \text{for} \quad w \in [0,1]$$

Some simple Calculations:

- $EX = 0, \quad \mathrm{var}(X) = 1, \quad EX^4 = \frac{1}{1-w}$

Last Time: Careful look at Kurtosis (cont.)

Special Cases: [graphic]

- $w = 0$ (no weight in middle), Kurtosis = -2 (minimum)
- $w = 1/3$ (uniform), Kurtosis = -1.5
- $w = 2/3$ Kurtosis = 0, (closest to Gaussian?)
- $w > 2/3$ (heavy tails), Kurtosis > 0, (finally positive)
- $w \approx 1$ (2 outliers), Kurtosis very large

Note strong asymmetry in Kurtosis

Last Time: Careful look at Kurtosis (cont.)

Aapo Hyvärinen comments:

Solve asymmetry problem with "different nonlinearities",

i.e. replace absolute kurtosis = $\left| E(\underrightarrow{w}^t \underrightarrow{Z})^4 - 3 \right|$ with:

1. "tanh": $\left(E\left|\underrightarrow{w}^t \underrightarrow{Z}\right| - \sqrt{\frac{2}{\pi}} \right)^2$ (since $E\left|N(0,1)\right| = \sqrt{\frac{2}{\pi}}$)

2. "gaus": $\left(E\varphi\left(\underrightarrow{w}^t \underrightarrow{Z}\right) - \frac{1}{2\sqrt{\pi}} \right)^2$ (since $E\varphi\left(N(0,1)\right) = \frac{1}{2\sqrt{\pi}}$)

Last Time: Careful look at Kurtosis (cont.)

Comparison via 3 point example: [graphic]

- upper left: noncomparable scales
- upper right: max rescaling is better
  - tanh and gaus "less asymmetric" than A. Kurt.
- lower left: still shows all are asymmetric
- lower right: "best scale"
  - A. Kurt. has pole at left, but "best for small w"
  - tanh and gaus have different zeros than A. Kurt.

ICA, Toy Examples Revisited (cont.)

E.g. Parabs Up and Down   (two distant clusters)

Tanh: [graphic]

- Only IC2 finds an outlier
- IC1 and IC3 have kurt. < 0
- IC3 finds most of 2 clusters
- but not so well as PC1

ICA, Toy Examples Revisited (cont.)

Gaus: [graphic]

- IC1 is classical "heavy tail kurtosis"

- IC2 nicely finds clusters

- IC3 is another bimodal direction (no insights about data)

Conclusion: tanh and gaus work as expected, and are useful

Big Picture View of Course Material

Recall 2 vital concepts:

I. Data Representation & Conceptualization

II. Understanding “Population Structure”

Big Picture View of Course, Data Representation

Object Space $\leftrightarrow$ Feature space

Curves — Vectors

Images

Shapes

$$\begin{pmatrix} x_{1,1} \\ \vdots \\ x_{d,1} \end{pmatrix}, \cdots, \begin{pmatrix} x_{1,n} \\ \vdots \\ x_{d,n} \end{pmatrix}$$

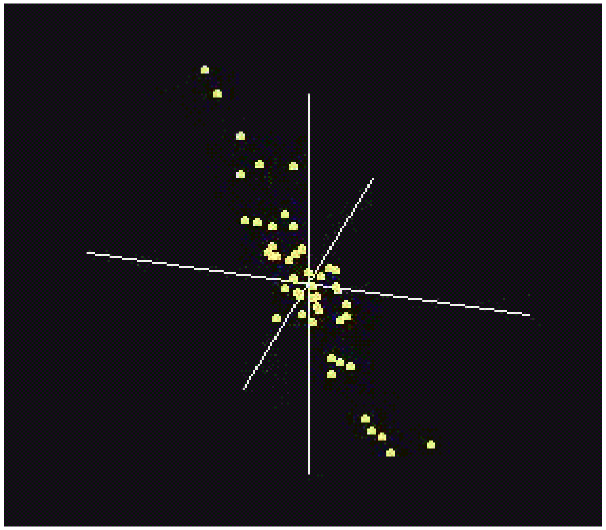
One to one mapping couples visualization in Object Space, with statistical analysis in Feature Space

Big Picture View of Course, Data Conceptualization

Feature space $\leftrightarrow$ Point Clouds

Vectors

$$\begin{pmatrix} x_{1,1} \\ \vdots \\ x_{d,1} \end{pmatrix}, \cdots, \begin{pmatrix} x_{1,n} \\ \vdots \\ x_{d,n} \end{pmatrix}$$



[Spinning Point Cloud Graphic] .

Big Picture View of Course, Population "Structure"

Main Idea: "analyzing" populations of complex objects

2 common major goals:

I. Understanding "population structure".
   - "visualization"
   - "intuition"

II. Statistical Classification, i.e. Discrimination
   - put into "known groups", based on "training data"
   - e.g. disease diagnosis

Statistical Classification, i.e. Discrimination

Interesting Example:

Corpora Callosa data, Recall from Lecture 01-21.02

Special thanks to G. Gerig and S. Ho, UNC Computer Science

Reference:

Kelemen, A., Szekely, G. and Gerig, G. (1997) Three dimensional model-based segmentation, TR-178 Technical Report Image Science Lab, ETH Zurich.

Data Objects: boundaries of "segmented" corpora callosa

Recall Corpora Callosa Data

Data Curves [example]

Feature vectors: use coefficients of Fourier boundary representation, $d = 80$

Object Space view: can either overlay, or show sequentially

In either case: hard to see “population structure”

Recall Corpora Callosa PCA

Raw Data   Modes of shape variation?

PC1:

- “overall bending”

PC2:

- Rotation of right end, “Sharpening” of left end

PC3:

- “thin” vs. “thick”

Discrimination for Corpora Callosa Data

Have 2 sub-populations:

Schizophrenics, $n = 40$ [sub-population of curves]

Controls, $n = 31$ [sub-population of curves]

Goal 1: See difference between populations? (???)

Goal 2: Given new shape: assign to a group

"automatic diagnosis (of schizophrenia)"

Discrimination for Corpora Callosa Data (cont.)

Very simple approach:

- Colored Parallel Coordinate view of data [graphic]
- Look for diff'nce between Schizophrenics and Controls
- Major "overplotting" problems (Schizos last, so "on top")
- No useful separation, since view is "too simple"
- Only looks in limited "coordinate directions"
- Perhaps "better separation" in other directions
- Caution: bottom show non-Gaussian

Discrimination for Corpora Callosa Data (cont.)

Another simple approach:

- for “widely separated data” [toy example]
- find “skewer through meatballs”
- using difference vector between means [toy example]
- Projection “separates sub-populations”

Alternate view:

- discrimination boundary is “orthogonal hyperplane”

Discrimination for Corpora Callosa Data (cont.)

Problem for Corpora Callosa Data:

- Subpopulation means nearly same
- Square of Difference, as Fraction of Total < 0.1%
- Thus effective discrimination must account for “spread”
- Perhaps can exploit covariance structure?

Discrimination for Corpora Callosa Data (cont.)

Another simple approach: PCA

- Again hope for "skewer between meatballs"
- This time focusing on covariance, not mean [toy example]
- Doesn't work for Corpora Callosa Data

Recall: PC1 PC2 PC3

- Recall PCA only feels "maximal variation"
- Different from "separating subsamples"
- PCA doesn't even use "class label information"

Discrimination for Corpora Callosa Data (cont.)

Another view of PCA problem: [toy data set]

- "maxim'l variation" can be different from "good separation"
- so PCA fails [PCA]
- mean difference better, not adequate [mean diff.]
- really want to work in "covariance structure"

Alternate Approach:

- modify mean difference, using "covariance structure"
- called Fisher Linear Discrimination

Fisher Linear Discrimination

Careful development:

Mathematical Notation (vectors with dimension d):

Class 1: $\underrightarrow{X}_1^{(1)},\ldots,\underrightarrow{X}_{n_1}^{(1)}$

Class 2: $\underrightarrow{X}_1^{(2)},\ldots,\underrightarrow{X}_{n_2}^{(2)}$

Class Centerpoints: $\overline{\underrightarrow{X}}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1}\underrightarrow{X}_i^{(1)}$ and $\overline{\underrightarrow{X}}^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2}\underrightarrow{X}_i^{(2)}$

Fisher Linear Discrimination (cont.)

Covariances: $\hat{\Sigma}^{(j)} = \widetilde{X}^{(j)} \widetilde{X}^{(j)^t}$, for $j = 1,2$ (outer products)

Based on "normalized, centered data matrices":

$$\widetilde{X}^{(j)} = \frac{1}{\sqrt{n_j}} \left(\underline{X}_1^{(j)} - \underline{\overline{X}}^{(j)}, \ldots, \underline{X}_{n_j}^{(j)} - \underline{\overline{X}}^{(j)} \right)$$

note: Use "MLE" version of normalization, for simpler notation

Terminology (useful later): $\hat{\Sigma}^{(j)}$ are "within class covariances"

Fisher Linear Discrimination (cont.)

Major assumption: Class covariances are same (or "similar")

Good estimate of "common within class covariance"?

(recall [toy example])

Pooled (weighted average) within class covariance:

$$\hat{\Sigma}^{w} = \frac{n_1 \hat{\Sigma}^{(1)} + n_2 \hat{\Sigma}^{(2)}}{n_1 + n_2} = \widetilde{X}\widetilde{X}^{t}$$

for the "full data matrix":

$$\widetilde{X} = \frac{1}{\sqrt{n}} \left(\sqrt{n_1}\,\widetilde{X}^{(1)} \; \sqrt{n_2}\,\widetilde{X}^{(2)} \right)$$

Fisher Linear Discrimination (cont.)

Note: $\hat{\Sigma}^{w}$ is similar to $\hat{\Sigma}$ from before

- i.e. "covariance matrix ignoring class labels"

- important difference is "class by class centering"

(recall [toy example])

Fisher Linear Discrimination (cont.)

Simple way to find "correct covariance adjustment":

Individ'ly transform subpop'ns so "spherical" about their means

$$\underset{\rightarrow}{Y}_i^{(j)} = \left(\hat{\Sigma}^w\right)^{-1/2} \underset{\rightarrow}{X}_i^{(j)}$$

(upper right in [toy example])

then:

"best separating hyperplane"

is

"perpendicular bisector of line between means"

Fisher Linear Discrimination (cont.)

So in transformed space, the separating hyperlane has:

Transformed normal vector:

$$\underset{\rightarrow}{n}_{TFLD} = \left(\hat{\Sigma}^{w}\right)^{-1/2} \underset{\rightarrow}{\overline{X}}^{(1)} - \left(\hat{\Sigma}^{w}\right)^{-1/2} \underset{\rightarrow}{\overline{X}}^{(2)} = \left(\hat{\Sigma}^{w}\right)^{-1/2} \left(\underset{\rightarrow}{\overline{X}}^{(1)} - \underset{\rightarrow}{\overline{X}}^{(2)}\right)$$

Transformed intercept:

$$\underset{\rightarrow}{\mu}_{TFLD} = \frac{1}{2}\left(\hat{\Sigma}^{w}\right)^{-1/2} \underset{\rightarrow}{\overline{X}}^{(1)} + \frac{1}{2}\left(\hat{\Sigma}^{w}\right)^{-1/2} \underset{\rightarrow}{\overline{X}}^{(2)} = \left(\hat{\Sigma}^{w}\right)^{-1/2} \left(\frac{1}{2}\underset{\rightarrow}{\overline{X}}^{(1)} + \frac{1}{2}\underset{\rightarrow}{\overline{X}}^{(2)}\right)$$

Equation:

$$\left\{\underset{\rightarrow}{y} : \left\langle \underset{\rightarrow}{y}, \underset{\rightarrow}{n}_{TFLD} \right\rangle = \left\langle \underset{\rightarrow}{\mu}_{TFLD}, \underset{\rightarrow}{n}_{TFLD} \right\rangle\right\}$$

(lower right in [toy example])

Fisher Linear Discrimination (cont.)

Thus discrimination rule is:

Given a new data vector $\underset{\rightarrow}{X}^0$, Choose Class 1 when:

$$\left\langle \left(\hat{\Sigma}^w\right)^{-1/2} \underset{\rightarrow}{X}^0, \underset{\rightarrow}{n}_{TFLD} \right\rangle \geq \left\langle \underset{\rightarrow}{\mu}_{TFLD}, \underset{\rightarrow}{n}_{TFLD} \right\rangle$$

i.e. (transforming back to original space)

$$\left\langle \underset{\rightarrow}{X}^0, \left(\hat{\Sigma}^w\right)^{-1/2} \underset{\rightarrow}{n}_{TFLD} \right\rangle \geq \left\langle \left(\hat{\Sigma}^w\right)^{1/2} \underset{\rightarrow}{\mu}_{TFLD}, \left(\hat{\Sigma}^w\right)^{-1/2} \underset{\rightarrow}{n}_{TFLD} \right\rangle$$

$$\left\langle \underset{\rightarrow}{X}^0, \underset{\rightarrow}{n}_{FLD} \right\rangle \geq \left\langle \underset{\rightarrow}{\mu}_{FLD}, \underset{\rightarrow}{n}_{FLD} \right\rangle$$

where:

$$\underset{\rightarrow}{n}_{FLD} = \left(\hat{\Sigma}^w\right)^{-1/2} \underset{\rightarrow}{n}_{TFLD} = \left(\hat{\Sigma}^w\right)^{-1} \left(\overline{\underset{\rightarrow}{X}}^{(1)} - \overline{\underset{\rightarrow}{X}}^{(2)} \right)$$

$$\underset{\rightarrow}{\mu}_{FLD} = \left(\hat{\Sigma}^w\right)^{1/2} \underset{\rightarrow}{\mu}_{TFLD} = \left(\frac{1}{2} \overline{\underset{\rightarrow}{X}}^{(1)} + \frac{1}{2} \overline{\underset{\rightarrow}{X}}^{(2)} \right)$$

Fisher Linear Discrimination (cont.)

Thus (in original space) have separating hyperplane with:

Normal vector: $\underrightarrow{n}_{FLD}$

Intercept: $\underrightarrow{\mu}_{FLD}$

(lower right in [toy example])