Distance Weighted Discrimination &

Geometrical Representation of HDLSS data

by

J. S. Marron[1] , Peter Hall[2] and Michael Todd[3]

[1]Department of Statistics
University of North Carolina

[2]Centre for Mathematics and its Applications
Australian National University

[3]School of Operations Research and Industrial Engineering
& Department of Computer Science - Cornell University

Medical Imaging (serious FDA opportunity)

Early problems:

- Image denoising
- Registration
- Segmentation

More recent Problems:

- Understanding populations of “images”
- Discrimination (classification)
- Functional Data Analysis (generalized?)

Functional Data Analysis: A Personal View

Easy introduction via: The "atom" of the statistical analysis

<u>Statistical Context</u>	<u>Atom</u>
1st Course	Number
Multivar. Analysis	Vector
F. D. A.	Complex Object (curve, image, shape, …)

Data Representation

Object Space $\leftrightarrow$ Feature space

Curves

Images

Shapes

Vectors

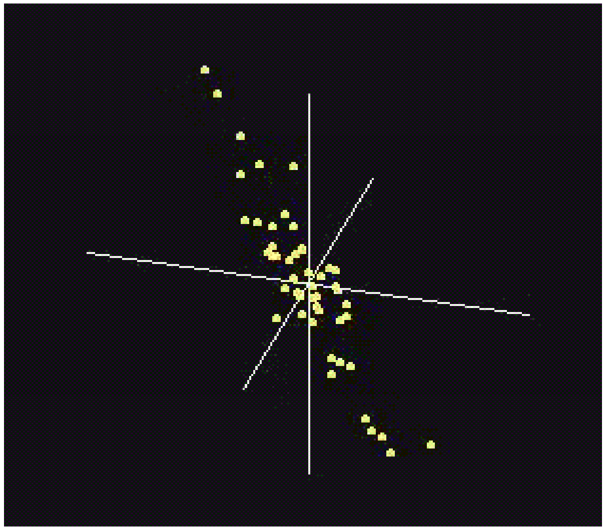
$$\begin{pmatrix} x_{1,1} \\ \vdots \\ x_{d,1} \end{pmatrix}, \cdots, \begin{pmatrix} x_{1,n} \\ \vdots \\ x_{d,n} \end{pmatrix}$$

Data Conceptualization

Feature space $\leftrightarrow$ Point Clouds

Vectors

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



Important Context

High Dimension Low Sample Size $d >> n$

(Personal) driving problems:

1. Medical imaging

   d high 10s – 100s, n 20s – 100s

2. Micro-arrays measuring gene expression

   d 100s – 10,000s, n 10s – 100s

3. Chemometric spectra

   d 1,000s, n 10s

A real data example

Genetic Micro-Arrays (thanks to C. M. Perou, et. al.):

Measures "expression" (activity) of many genes at once

Current Problem: "Batch effects" $(n = 49, \quad d = 2{,}452)$

(caused by production at different labs, g, h, j)

Visualization of Problem: PCA and 2-d scatterplot of proj'ns

- Serious problem, likely to affect subsequent analysis
- How to correct?

Batch Effect Adjustment

"Standard Approach": PCA (i.e. SVD), based on PC1

- Works well when PC1 is "in that direction" (Toy e.g.)

  (recall PC1 is in "direction of greatest variation")

- Otherwise (e.g. here) quite doubtful

Linear Model (+ Random Effects) Approaches

- "Interpretability"? (followed by exploratory data analysis??)

Proposed "New" Approach: Use discrimination methods

Discrimination

A.K.A. Classification (Two Class)

- Using "Training Data" from Class +1, and from Class –1
- Develop a "Rule", for assigning new data to a Class

Canonical Example: Disease Diagnosis

- New patients are either "healthy" or "ill"
- Determine on basis of measurements
- Based on preceding experience (training data)

Quick Overview of Discrimination

Toy Graphic i.i.d. $N(\mu, I)$, $\mu_{1,\pm} = \pm 2.2$, $n = 40$, $d = 50$

Classical Attempt: Fisher Linear Discrimination

Modern Approaches:

Support Vector Machine (toy graphic illustration)

Distance Weighted Discrimination

- Idea: "feel all of the data", not just "support vectors"
- Type into Google, to obtain paper
- Uses serious optimization (2^{nd} Order Cone Methods)

Application to Batch Effect Data

SVM Adjustment

- Looks reminiscent of above problem
- 2nd application to residuals still has gap?
- Must, since HDLSS, but "perhaps very small"?

DWD Adjustment

- Again reminiscent of above example
- 2nd application to residuals looks great!

Application to Batch Effect Data (cont.)

Careful: used different criteria for assessment

SVM adjustment, DWD assessment

- Now looks like similar results
- Reason for this? Geometrical Representation

Final result: Adjusted 2-d Scatterplots

- Applied Stepwise: 1. g vs. h & j, 2. h vs. j
- Great "mixing" of batches, i.e. successful adjustment

DWD vs. SVM Simulations

3 simulations: [Dist'n 1] [Dist'n 2] [Dist'n 3]

- Shows each method is sometimes best
- DWD is "usually near best" (i.e. "good overall")
- Note: all are closer together for higher $d = 1600$
- Explanation: Geometrical Representation

Some Simple "Paradoxes" of HDLSS data

For d dim'al "Standard Normal" dist'n:

$$\underset{\rightarrow}{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} \sim N\left(\underset{\rightarrow}{0}, I\right)$$

Euclidean Distance to Origin (as $d \rightarrow \infty$):

$$\left\| \underset{\rightarrow}{Z} \right\| = \sqrt{d} + O_p(1)$$

- Data lie roughly on surface of sphere of radius $\sqrt{d}$
- Yet origin is point of "highest density"???
- Paradox resolved by "density w. r. t. Lebesgue Measure"

Some Simple "Paradoxes" of HDLSS data (cont.)

For d dim'al "Standard Normal" dist'n:

$$\underset{\rightarrow}{Z}_1 \text{ indep. of } \underset{\rightarrow}{Z}_2 \sim N\left(\underset{\rightarrow}{0}, I\right)$$

Euclidean Distance between $\underset{\rightarrow}{Z}_1$ and $\underset{\rightarrow}{Z}_2$ (as $d \to \infty$):

$$\left\| \underset{\rightarrow}{Z}_1 - \underset{\rightarrow}{Z}_2 \right\| = \sqrt{2d} + O_p(1)$$

- Distance tends to *non-random* constant
- Can extend to $\underset{\rightarrow}{Z}_1, \ldots, \underset{\rightarrow}{Z}_n$
- Where do they all go??? (we can only perceive 3 dim'ns)

Some Simple "Paradoxes" of HDLSS data (cont.)

For d dim'al "Standard Normal" dist'n:

$$\underset{\rightarrow}{Z}_1 \text{ indep. of } \underset{\rightarrow}{Z}_2 \sim N\left(\underset{\rightarrow}{0}, I\right)$$

High dim'al Angles(as $d \rightarrow \infty$):

$$Angle(\underset{\rightarrow}{Z}_1, \underset{\rightarrow}{Z}_2) = 90^\circ + O_p\left(\frac{1}{\sqrt{d}}\right)$$

- "Everything is orthogonal"???
- Where do they all go??? (again our perceptual limitations)
- Again 1st order structure is *non-random*

Geometrical Representation of HDLSS data

Assume $\underset{\rightarrow}{Z}_1,...,\underset{\rightarrow}{Z}_n \sim N(0,I)$, $d >> n$, asymptotics as $d \to \infty$

1. Study Subspace Generated by Data

   a. Hyperplane through 0, of dimension n

   b. Points are "nearly equidistant to 0", & dist $\sim \sqrt{d}$

   c. Within plane, can "rotate towards $\sqrt{d} \times$ Unit Simplex"

   d. *All Gaussian data sets* are"near U. Simplex vertices"!!!

   e. "Randomness" *appears only in rotation* of simplex

[Two Point Toy Example]()

Geometrical Representation of HDLSS data (cont.)

Assume $\underset{\rightarrow}{Z}_1, \ldots, \underset{\rightarrow}{Z}_n \sim N(0, I)$, $d >> n$, asymptotics as $d \to \infty$

2. Study Hyperplane Generated by Data
   a. $n-1$ dimensional hyperplane
   b. Points are pair-wise equidistant, dist $\sim \sqrt{2d}$
   c. Points lie at vertices of $\sqrt{2d} \times$ "regular $n-$hedron"
   d. Again "randomness in data" is *only in rotation*
   e. Surprisingly rigid structure in data?

Three Point Toy Example

Geometrical Representation of HDLSS data (cont.)

Simulation View: shows "rigidity after rotation"

Straightforward Generalizations:

- non-Gaussian data: only need moments
- non-independent: use "mixing conditions"

⋮

All based on simple "Laws of Large Numbers"

Geometrical Representation of HDLSS data (cont.)

Explanation of Observed Behavior (Batch Effect & Simulations):

Recall “everything similar for very high d”

- 2 popn’s are 2 simplices
- everything is the same distance from the other class
- i.e. everything is a support vector
- i.e. all sensible directions show “data piling”
- so “sensible methods are all nearly the same”

Interesting Questions:

- Views on “Dimensionality Reduction”?

- Relation to “Curse of Dimensionality”???

Some Carry Away Lessons

- HDLSS contexts are worth more study
- DWD better than SVM for HDLSS data
- "Randomness" in HDLSS data is *only rotations*
- Modulo random rotation, have "constant simplex shape"
- How to put this new structure to serious work?