Object Oriented Data Analysis

J. S. Marron

Dept. of Statistics and Operations Research,

University of North Carolina

January 16, 2014

Object Oriented Data Analysis

What is the "atom" of a statistical analysis?

- 1st Course: Numbers
- Multivariate Analysis Course : Vectors
- Functional Data Analysis: Curves
- More generally: Data Objects

Object Oriented Data Analysis

Examples:

- Medical Image Analysis
  - Images as Data Objects?
  - Shape Representations as Objects
- Gene Expression (Microarrays – RNAseq)
  - Just multivariate analysis?

Principal Component Analysis

More Than *Dimensionality Reduction*:

- Visualization
  - Relationships Between Objects (Scores)
  - Drivers of Relationships (Loadings)

Principal Component Analysis

More Than *Dimensionality Reduction*:

- Visualization
  - Relationships Between Objects (Scores)
  - Drivers of Relationships (Loadings)

But ∃ Limitations (good to know about)

Principal Component Analysis

Visualization Limitation:

Finds Directions of Maximal Variation

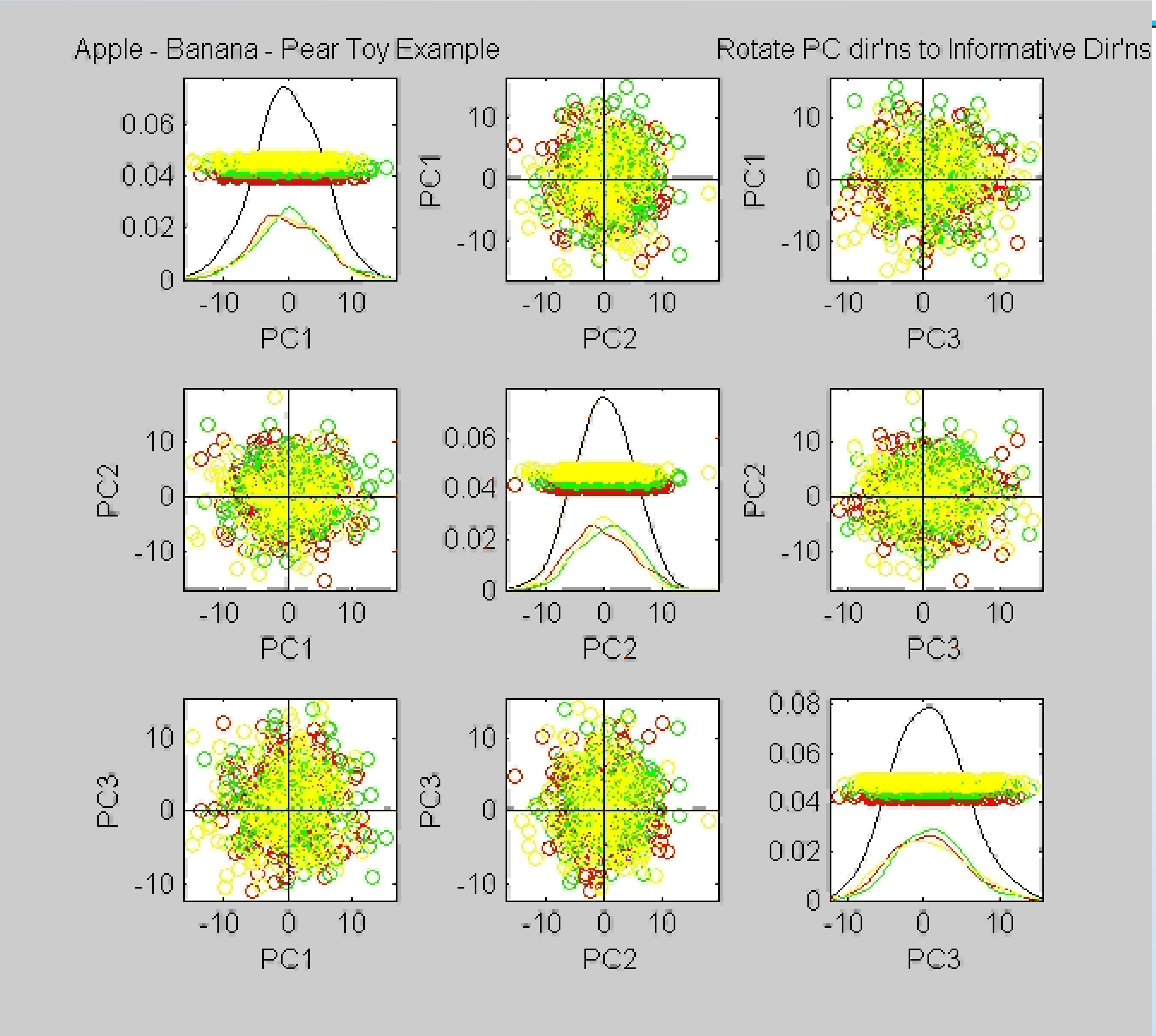
Principal Component Analysis

Visualization Limitation:

Finds Directions of Maximal Variation

- Apple – Banana – Pear Example (6-d)

Apple – Banana - Pear

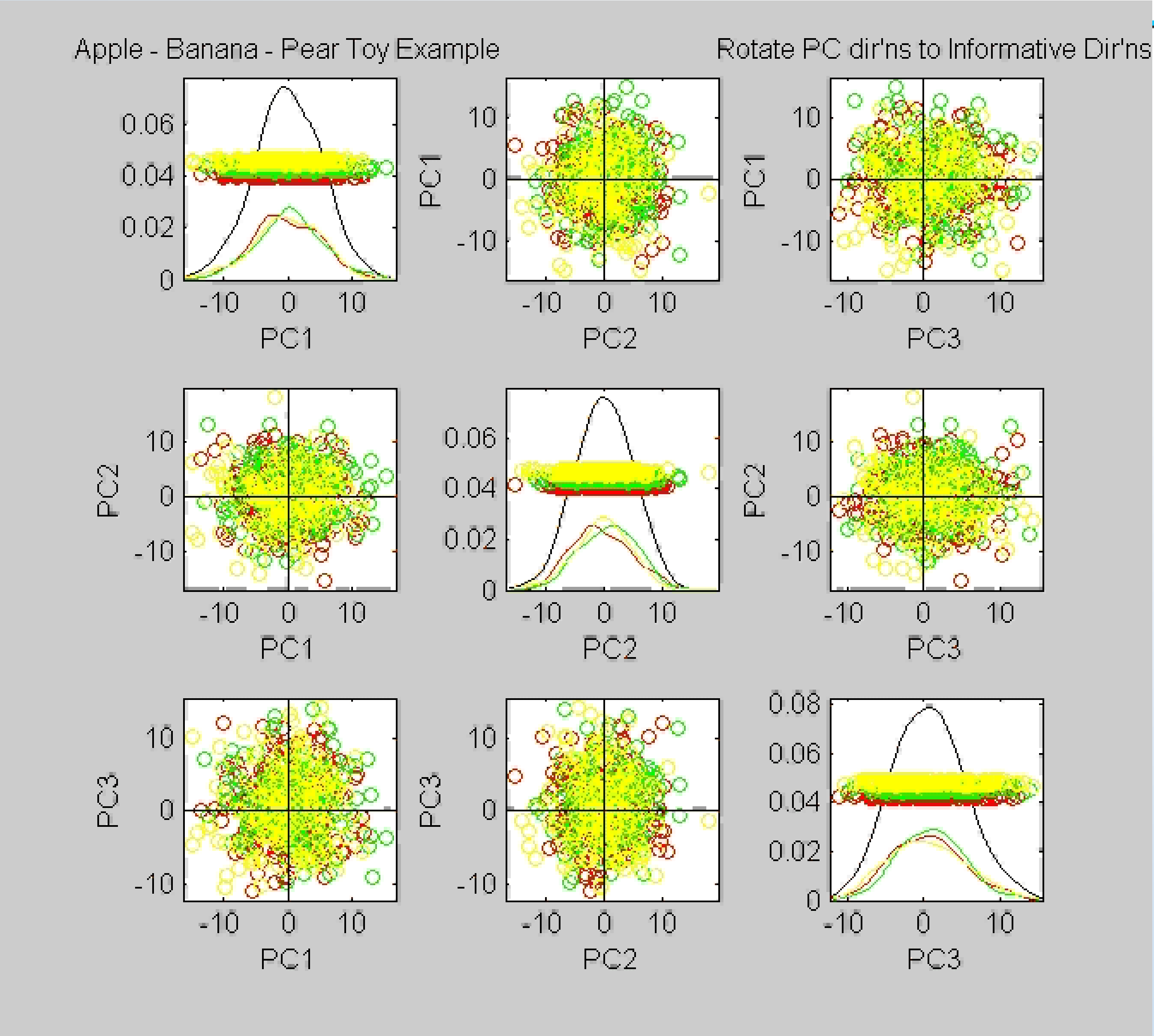


Apple – Banana - Pear

- Structure in Data Obscured
- 1st 3 PC Dir'ns are Pure Noise
- Rotate Axes to Find Structure

UNC, Stat & OR

Apple – Banana - Pear



Principal Component Analysis

Visualization Limitation:

Finds Directions of Maximal Variation

- Apple – Banana – Pear Example (6-d)
- Often Doesn't Separate Subgroups

HDLSS Classification (i.e. Discrimination)

Background: Two Class (Binary) version:

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 "Healthy" or "Ill"
- Determined based on measurements

UNC, Stat & OR

HDLSS Classification (Cont.)

- Ineffective Methods:
  - Fisher Linear Discrimination
  - Gaussian Likelihood Ratio

- Less Useful Methods:
  - Nearest Neighbors
  - Neural Nets

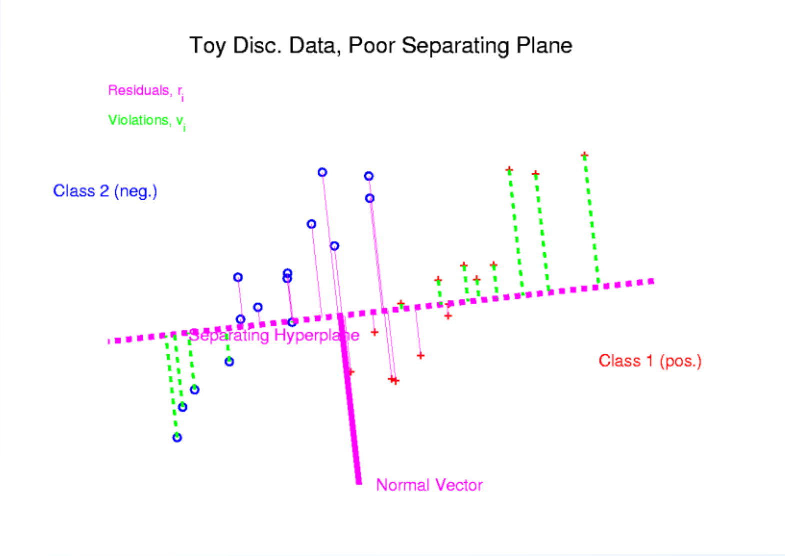
("black boxes", no "directions" or intuition)

HDLSS Classification (Cont.)

- Currently Fashionable Methods:
  - Support Vector Machines
  - Trees Based Approaches
- New High Tech Method
  - Distance Weighted Discrimination (DWD)
    - Specially designed for HDLSS data
    - Avoids "data piling" problem of SVM
    - Solves more suitable optimization problem

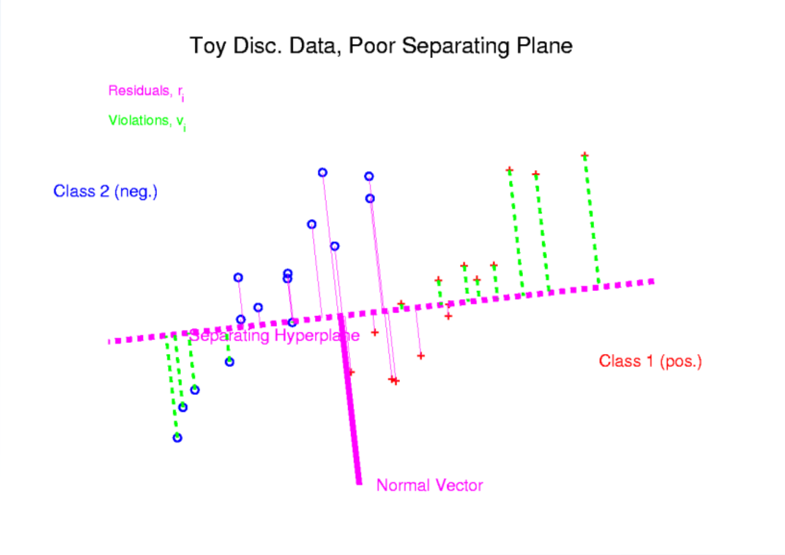
HDLSS Classification (Cont.)

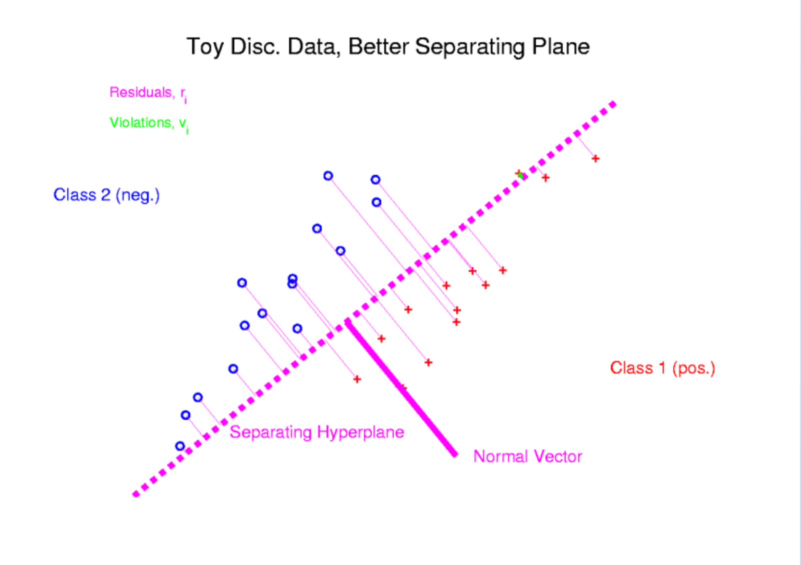
- Currently Fashionable Methods:
  - Trees Based Approaches
  - Support Vector Machines:



HDLSS Classification (Cont.)

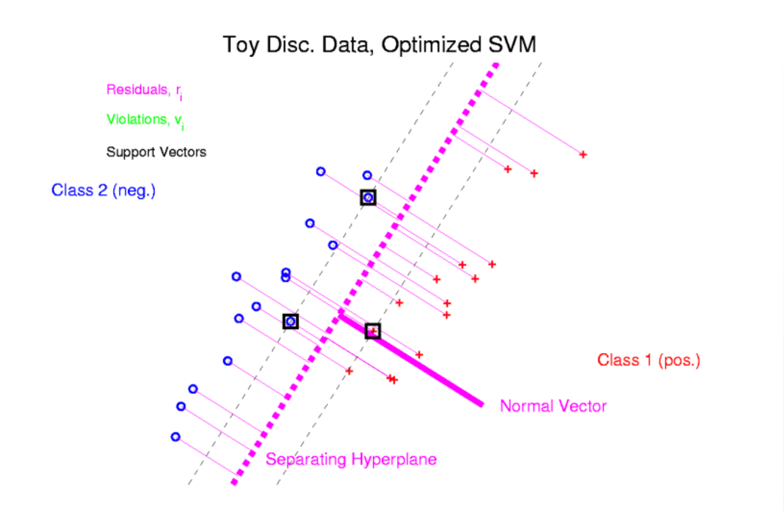
- Currently Fashionable Methods:
  - Trees Based Approaches
  - Support Vector Machines:

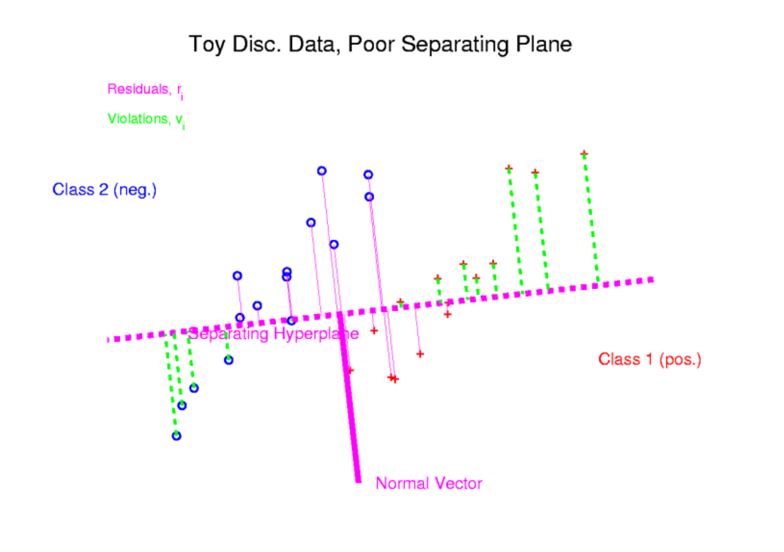


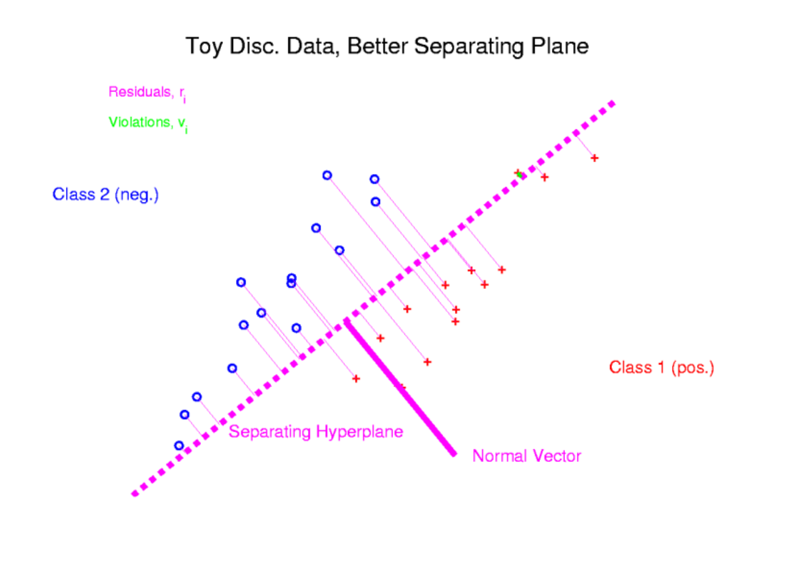


HDLSS Classification (Cont.)

- Currently Fashionable Methods:
  - Trees Based Approaches
  - Support Vector Machines:







Kernel Embedding Idea

Aizerman, Braverman, Rozoner (1964)

Make data *linearly separable*

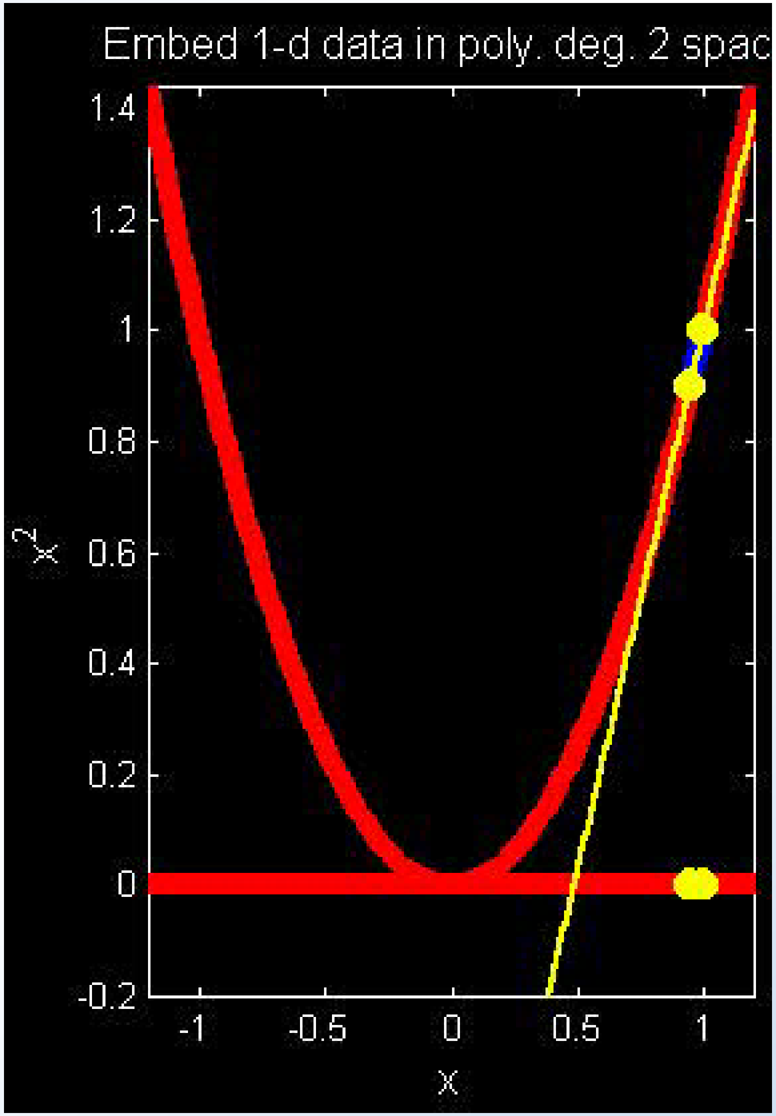
by embedding in

higher dimensional space

Kernel Embedding Idea

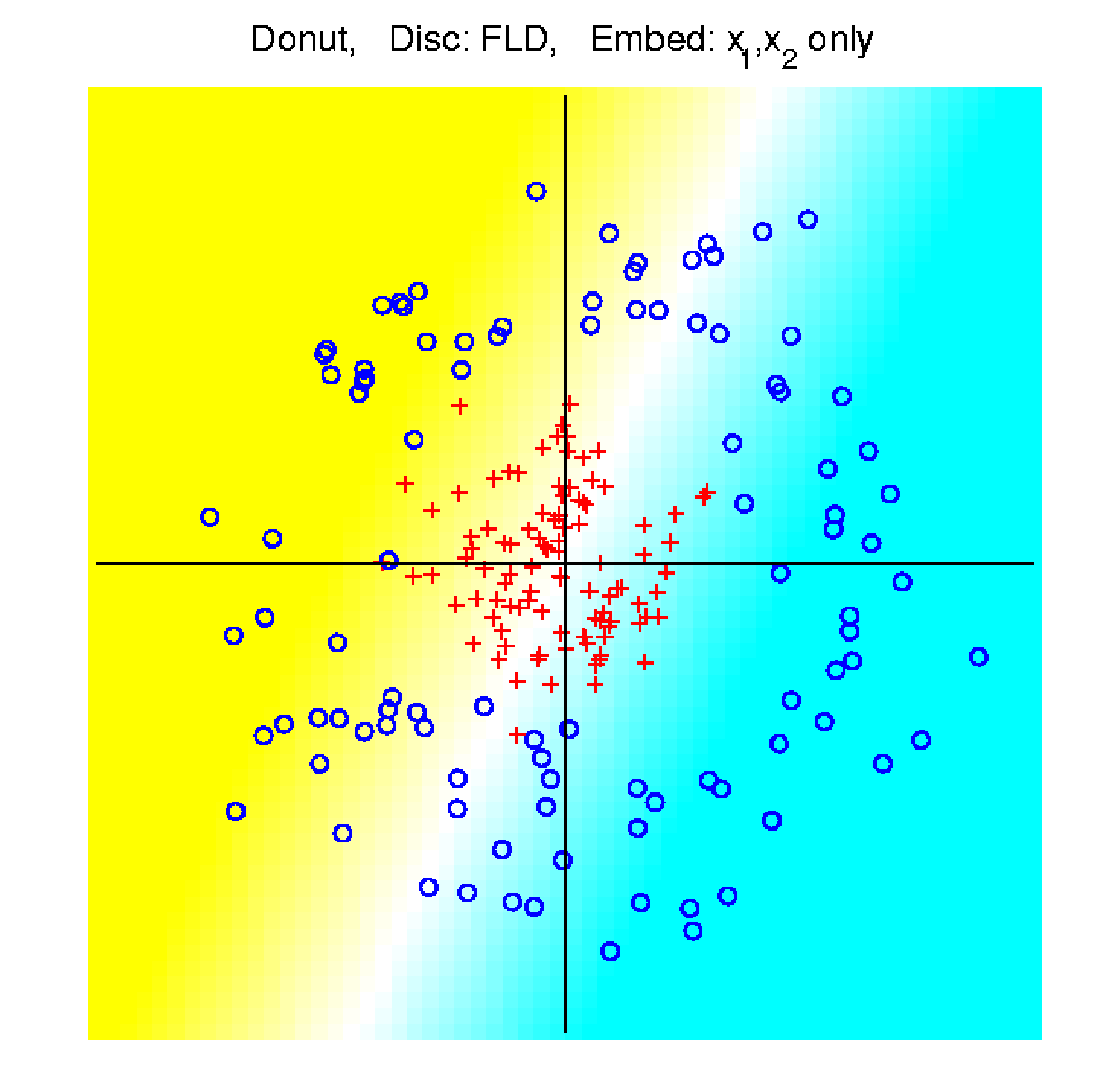
UNC, Stat & OR

Linearly separable by embedding in <u>higher dimensions</u>



Kernel Embedding Idea

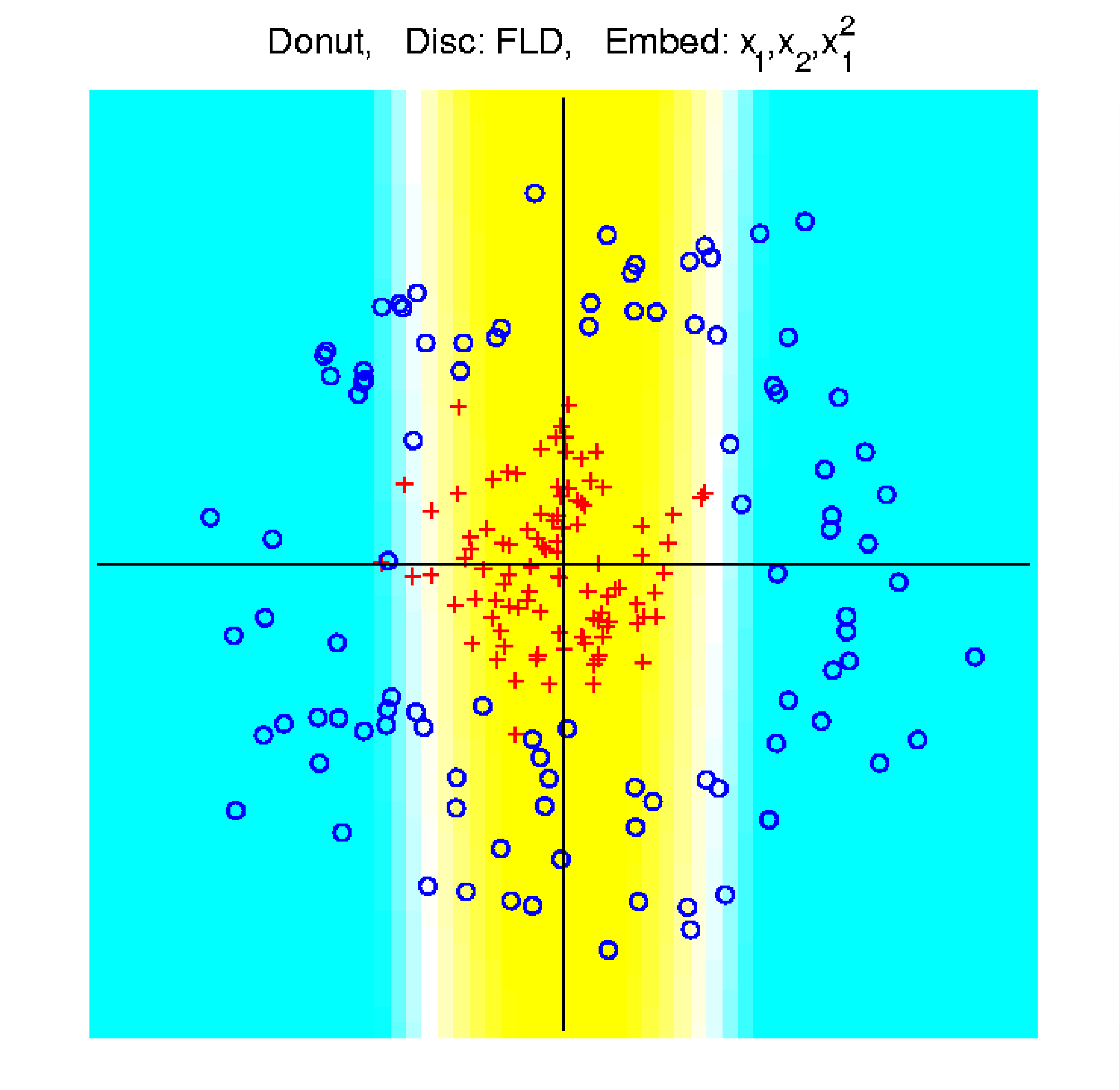
Linearly separable by embedding in <u>higher dimensions</u>



Kernel Embedding Idea

Linearly separable

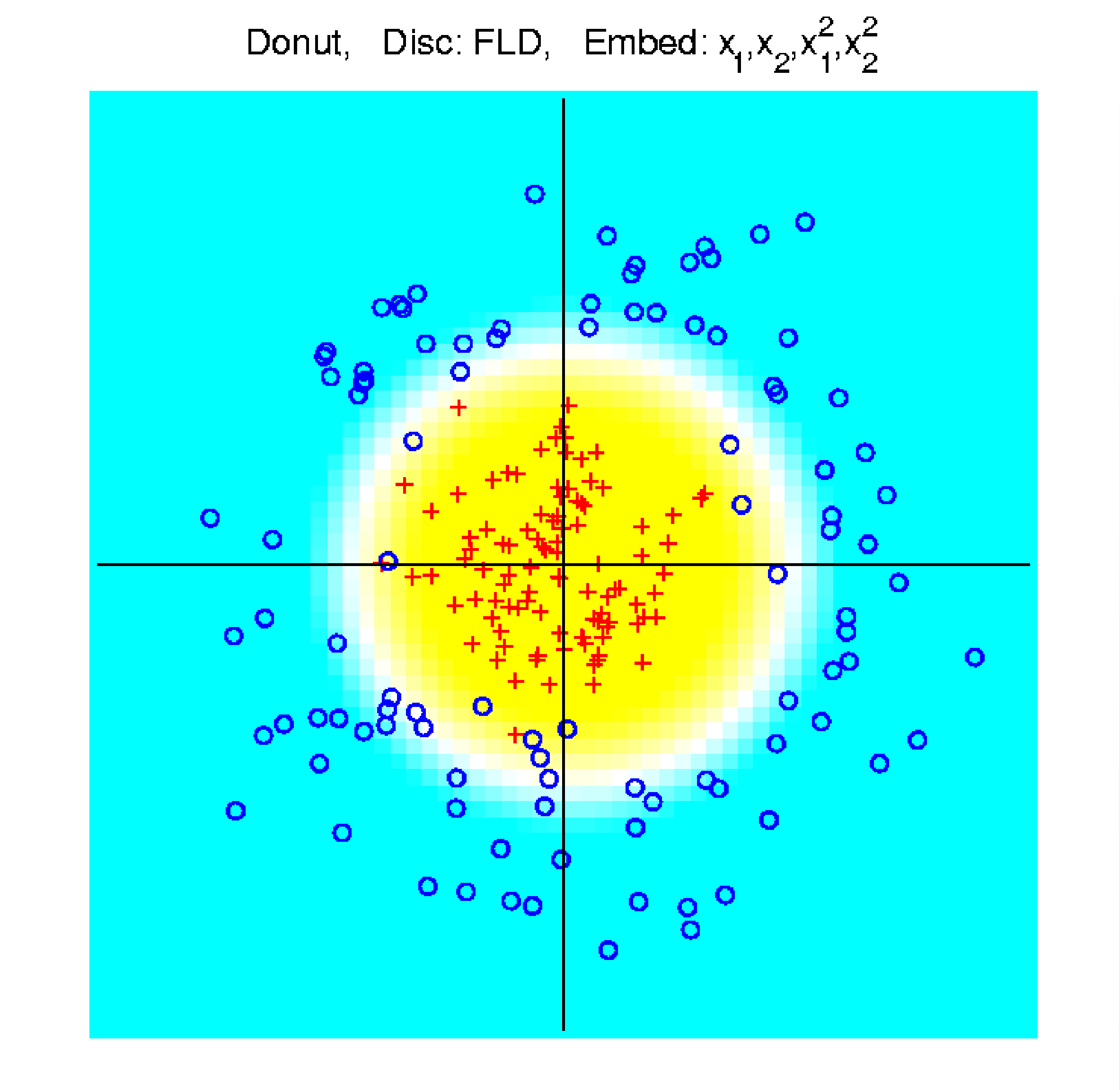
by embedding in <u>higher dimensions</u>



UNC, Stat & OR

Kernel Embedding Idea

Linearly separable
by
embedding
in
<u>higher</u>
<u>dimensions</u>



Kernel Embedding Idea

Linearly separable by embedding in <u>higher dimensions</u>

Distributional Assumptions in Embedded Space?

||
V

Support Vector Machine

HDLSS Classification (Cont.)

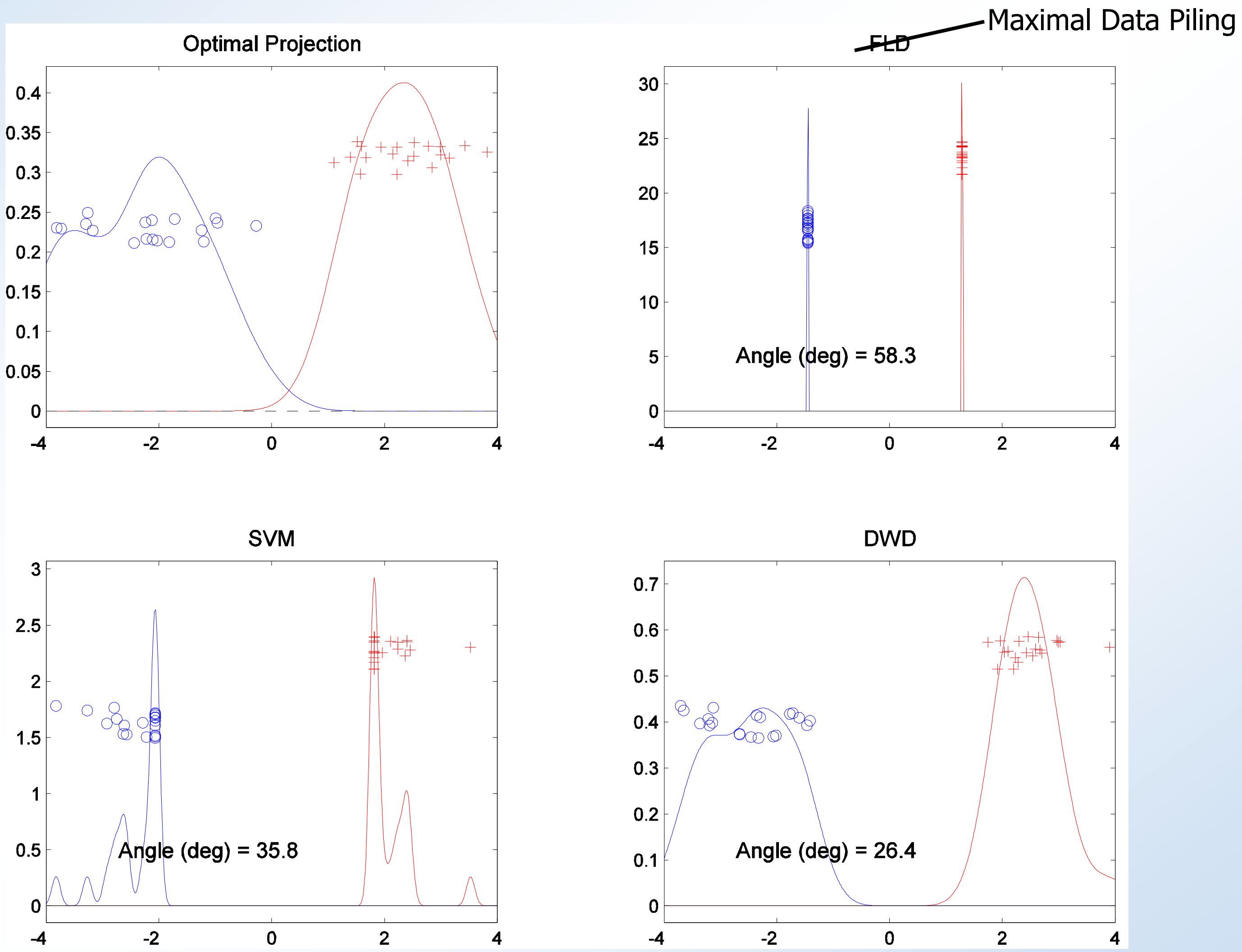
Comparison of Linear Methods (toy data):

$$N_d\left(\mu, I\right),\ \mu_{1,\pm} = \pm 2.2,\ n_1 = n_2 = 20,\ d = 50$$

- Optimal Direction
  - Excellent, but need dir'n in dim = 50
- Maximal Data Piling (J. Y. Ahn, D. Peña)
  - *Great separation*, but generalizability???
- Support Vector Machine
  - More separation, gen'ity, but some data piling?
- Distance Weighted Discrimination
  - Avoids data piling, good gen'ity, Gaussians?

Distance Weighted Discrimination

UNC, Stat & OR



Distance Weighted Discrimination

Based on Optimization Problem: $$\min_{w,b} \sum_{i=1}^{n} \frac{1}{r_i}$$

More precisely work in appropriate penalty for violations

Optimization Method (Michael Todd):

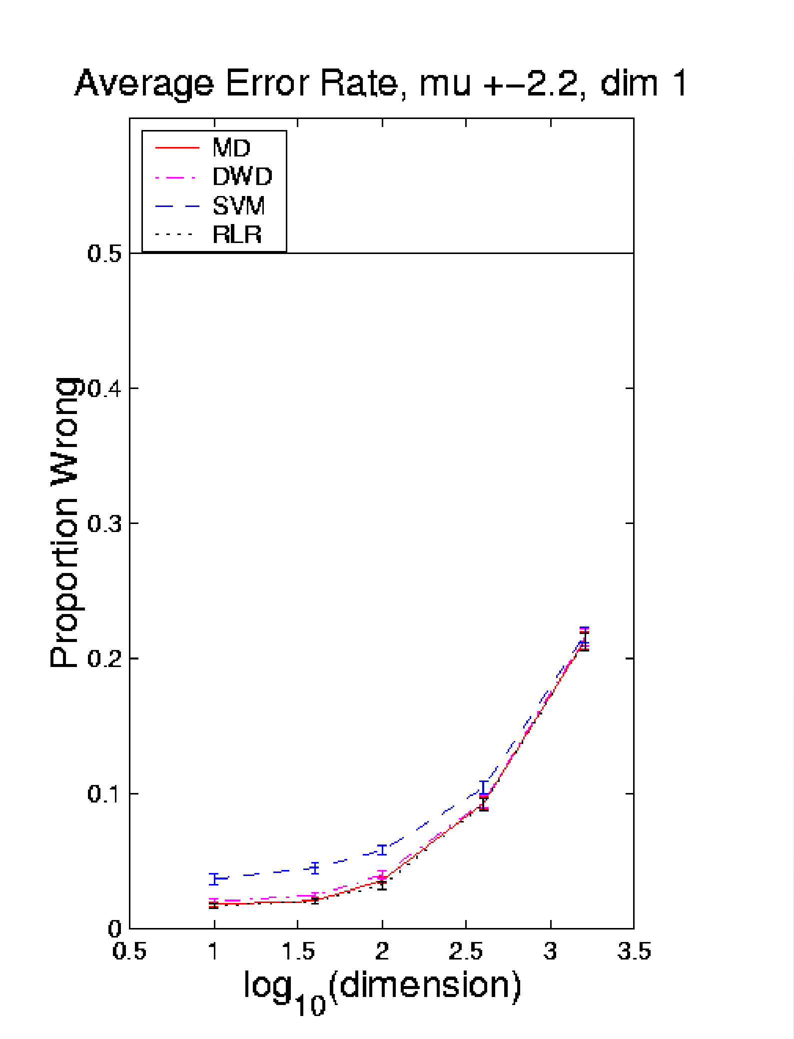
Second Order Cone Programming

- Still Convex gen'tion of quadratic prog'ing
- Fast greedy solution
- Can use existing software

Simulation Comparison

E.G. Above Gaussians:

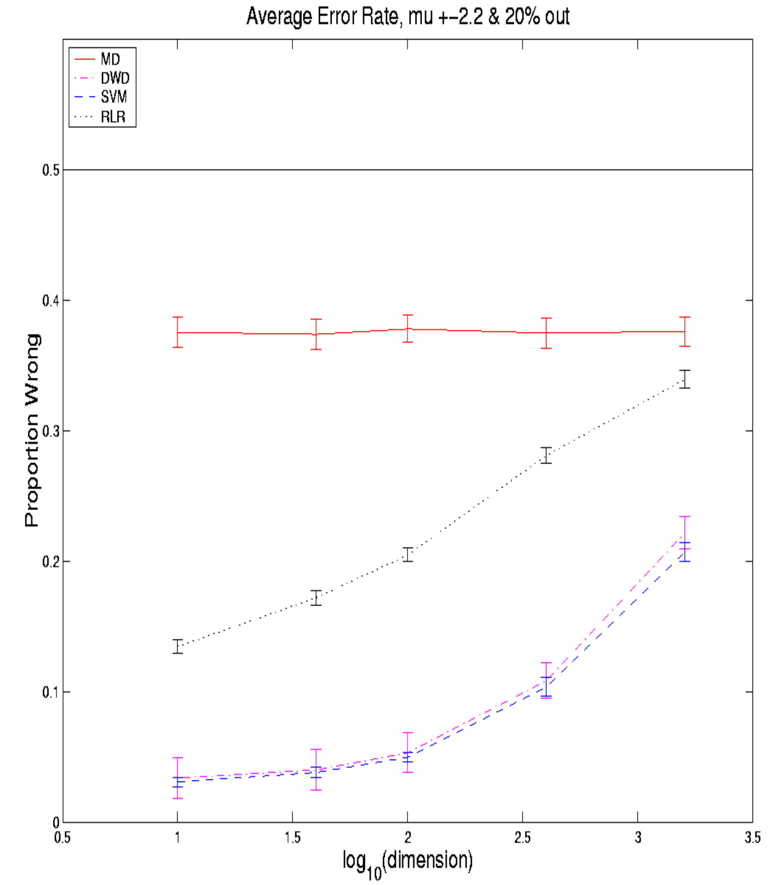
- Wide array of dim's
- SVM Subst'ly worse
- MD – Bayes Optimal
- DWD close to MD



Simulation Comparison

E.G. Outlier Mixture:

- Disaster for MD
- SVM & DWD much more solid
- Dir'ns are "robust"
- SVM & DWD similar



Simulation Comparison

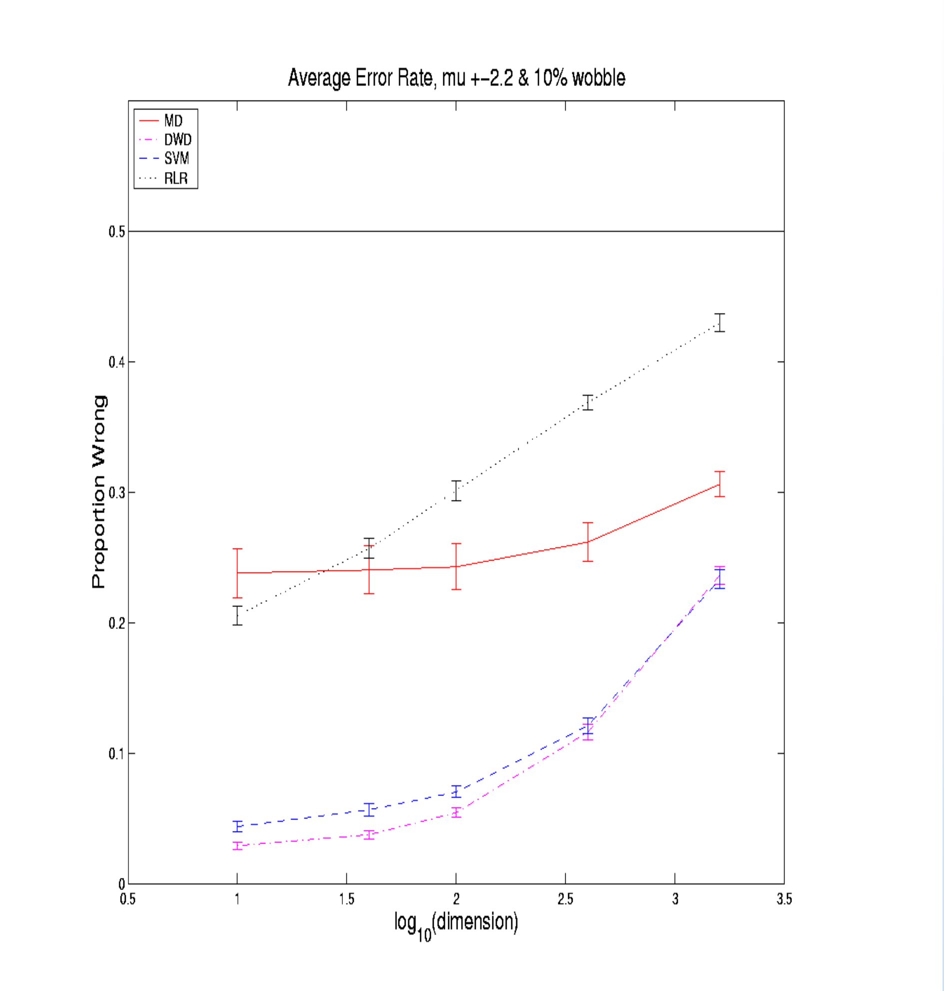
E.G. Wobble Mixture:

- Disaster for MD
- SVM less good
- DWD slightly better

Note: All methods

come together for

larger d ???



DWD Bias Adjustment for Microarrays

Microarray data:

- Simult. Measur'ts of "gene expression"
- Intrinsically HDLSS
  - Dimension $d \sim$ 1,000s – 10,000s
  - Sample Sizes $n \sim$ 10s – 100s

My view:

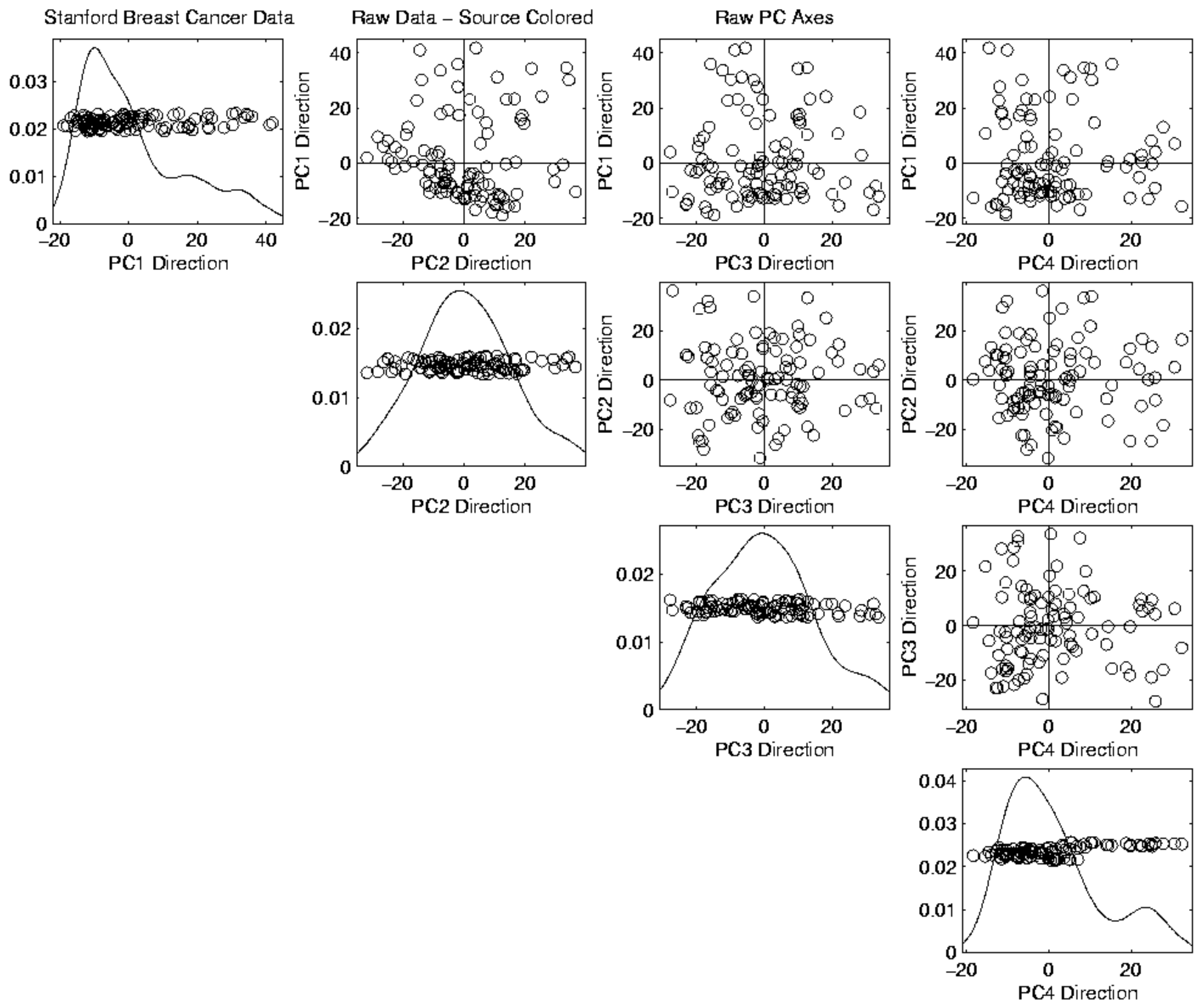
Each array is "point in cloud"

DWD Batch and Source Adjustment

- For Perou’s Stanford Breast Cancer Data
- Analysis in Benito, et al (2004) *Bioinformatics*

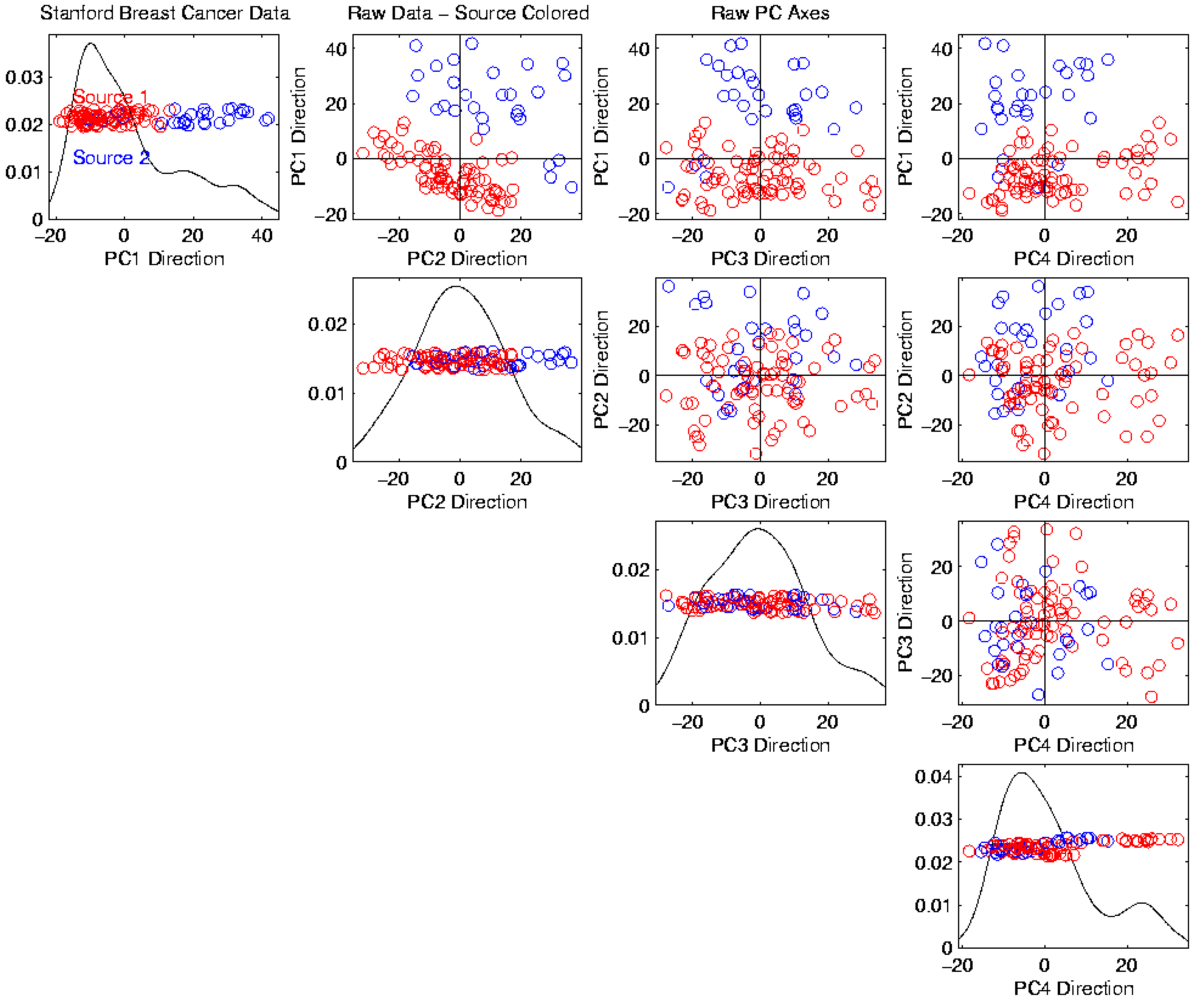
  https://genome.unc.edu/pubsup/dwd/
- Adjust for Source Effects
  - Different sources of mRNA
- Adjust for Batch Effects
  - Arrays fabricated at different times

DWD Adj: Raw Breast Cancer data

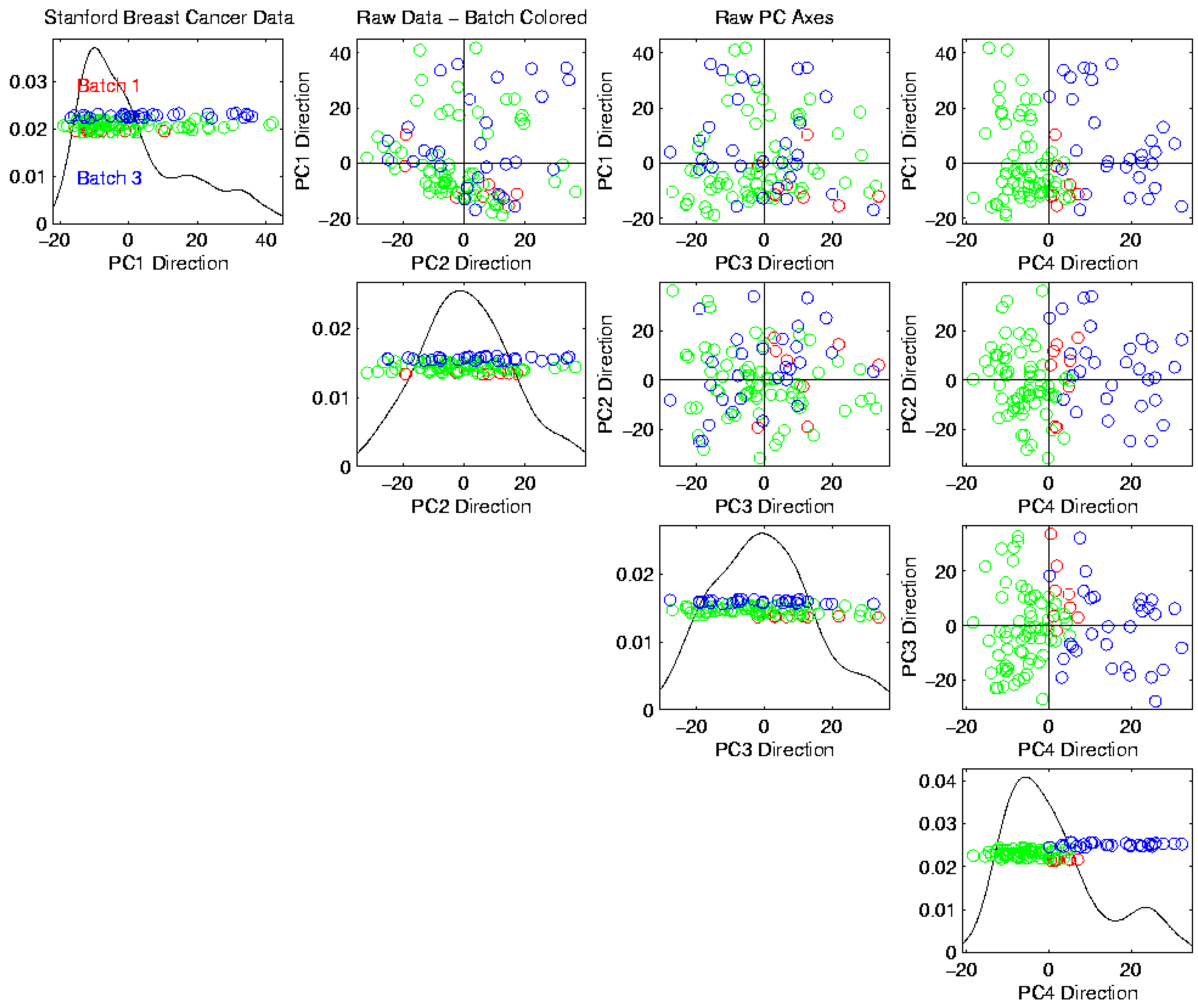


DWD Adj: Source Colors

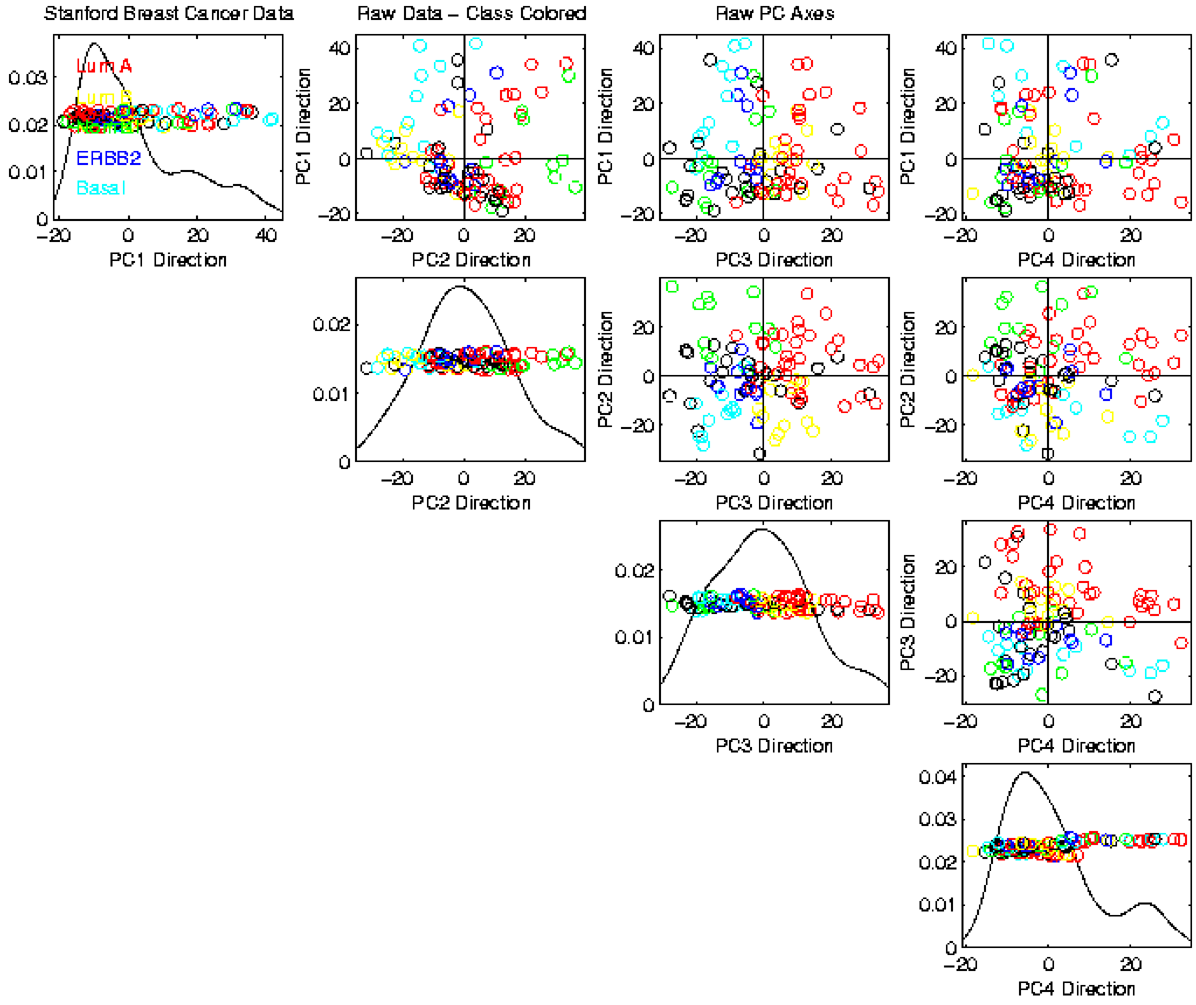
UNC, Stat & OR



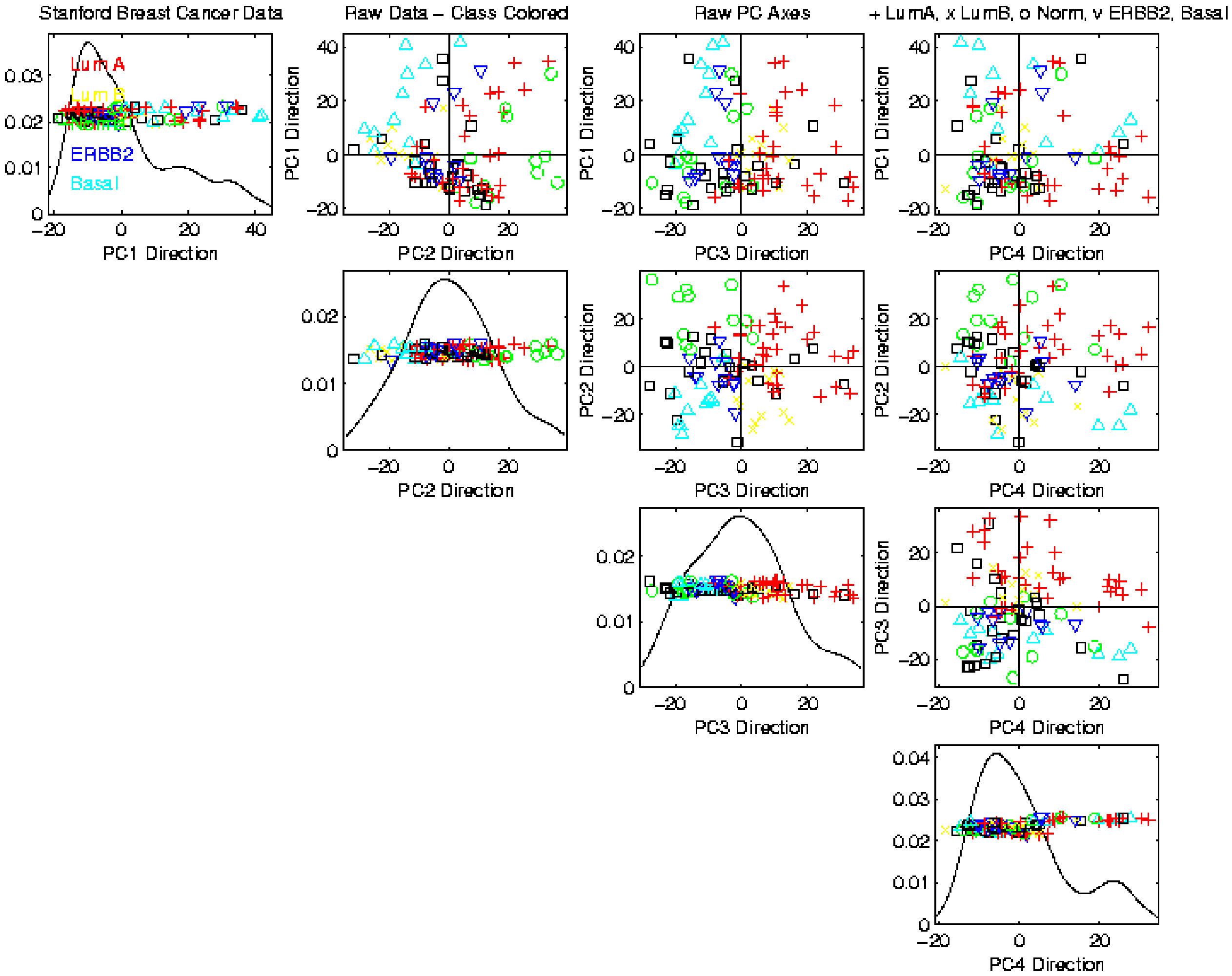
DWD Adj: Batch Colors



DWD Adj: Biological Class Colors

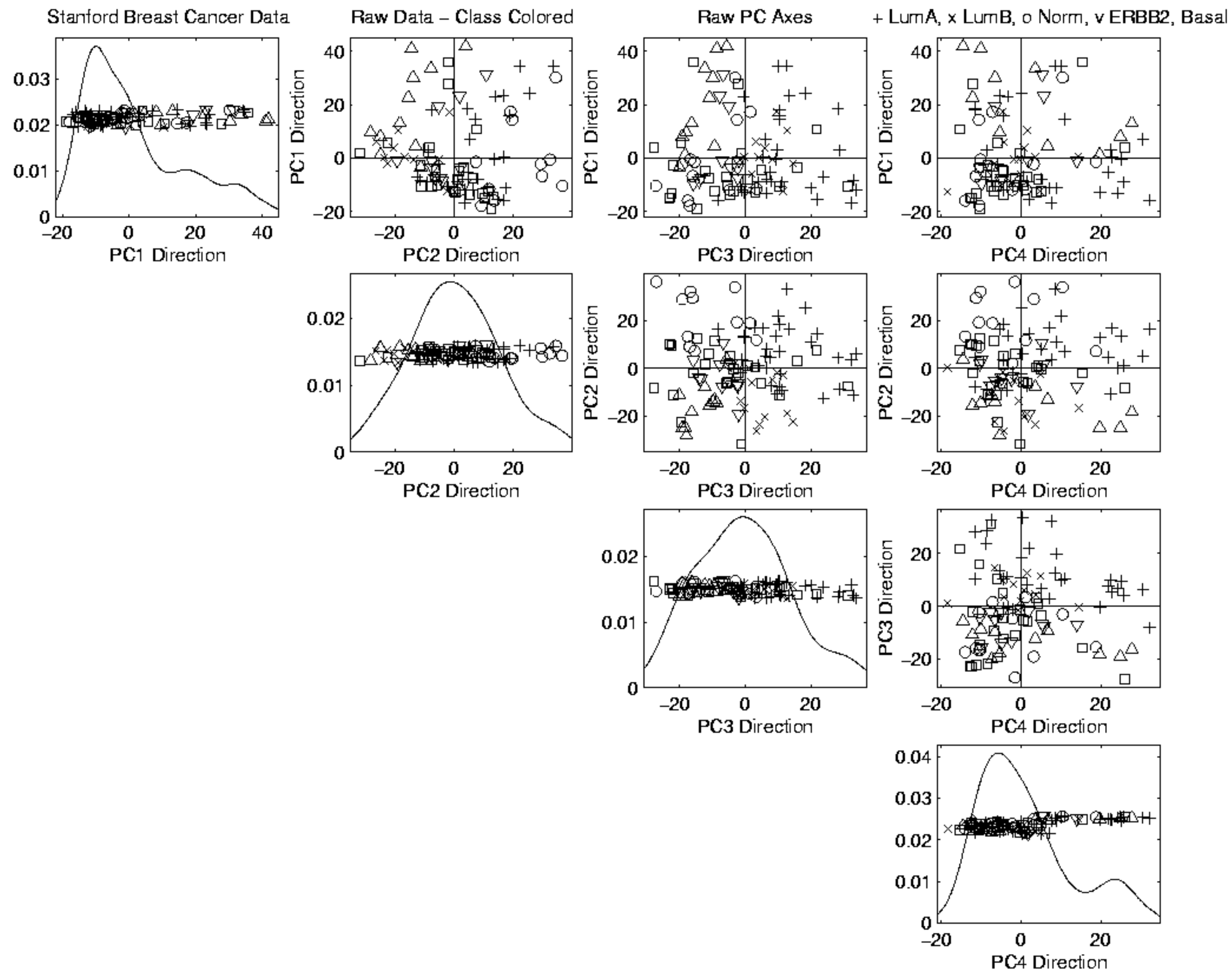


DWD Adj: Biological Class Colors & Symbols



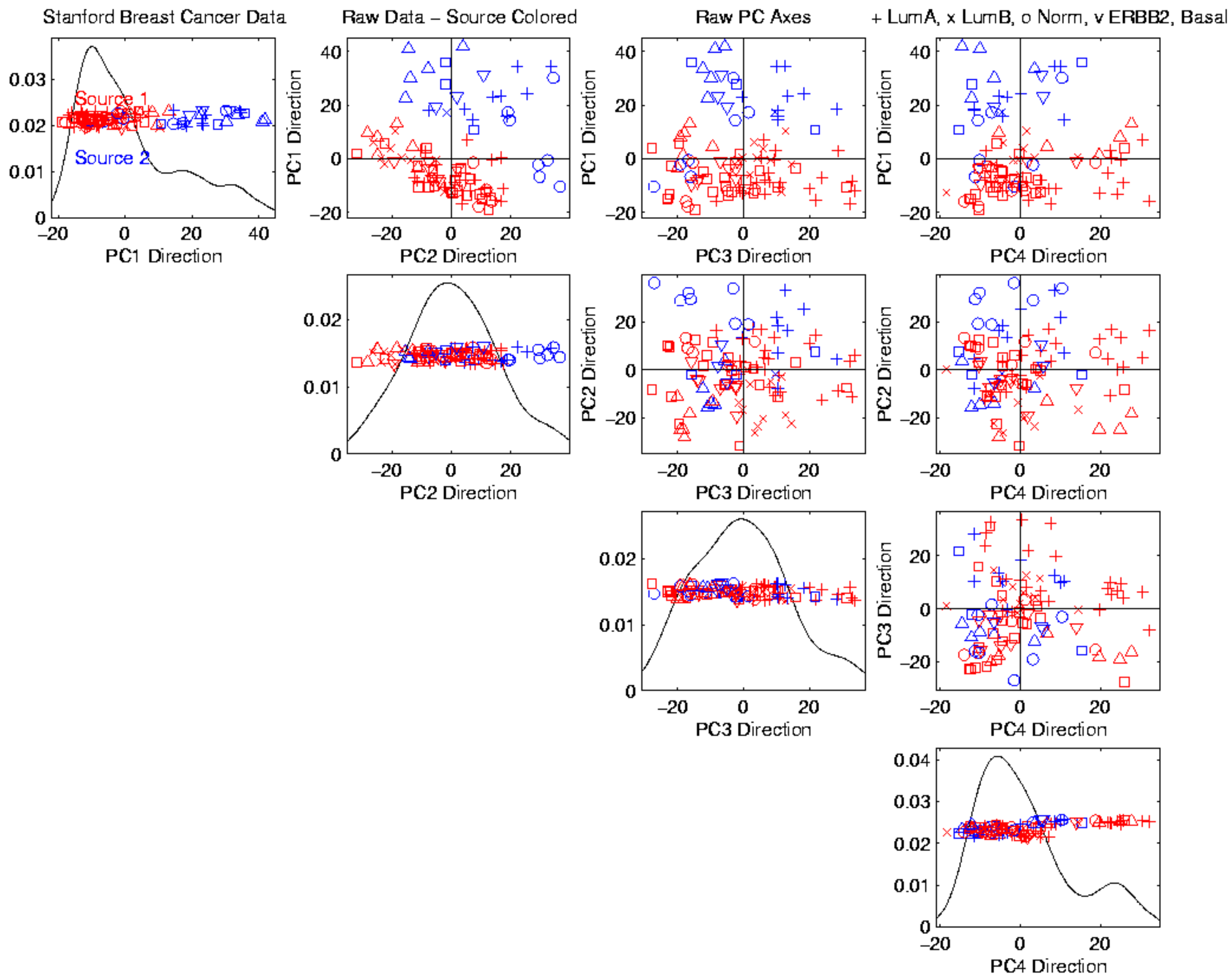
DWD Adj: Biological Class Symbols

UNC, Stat & OR



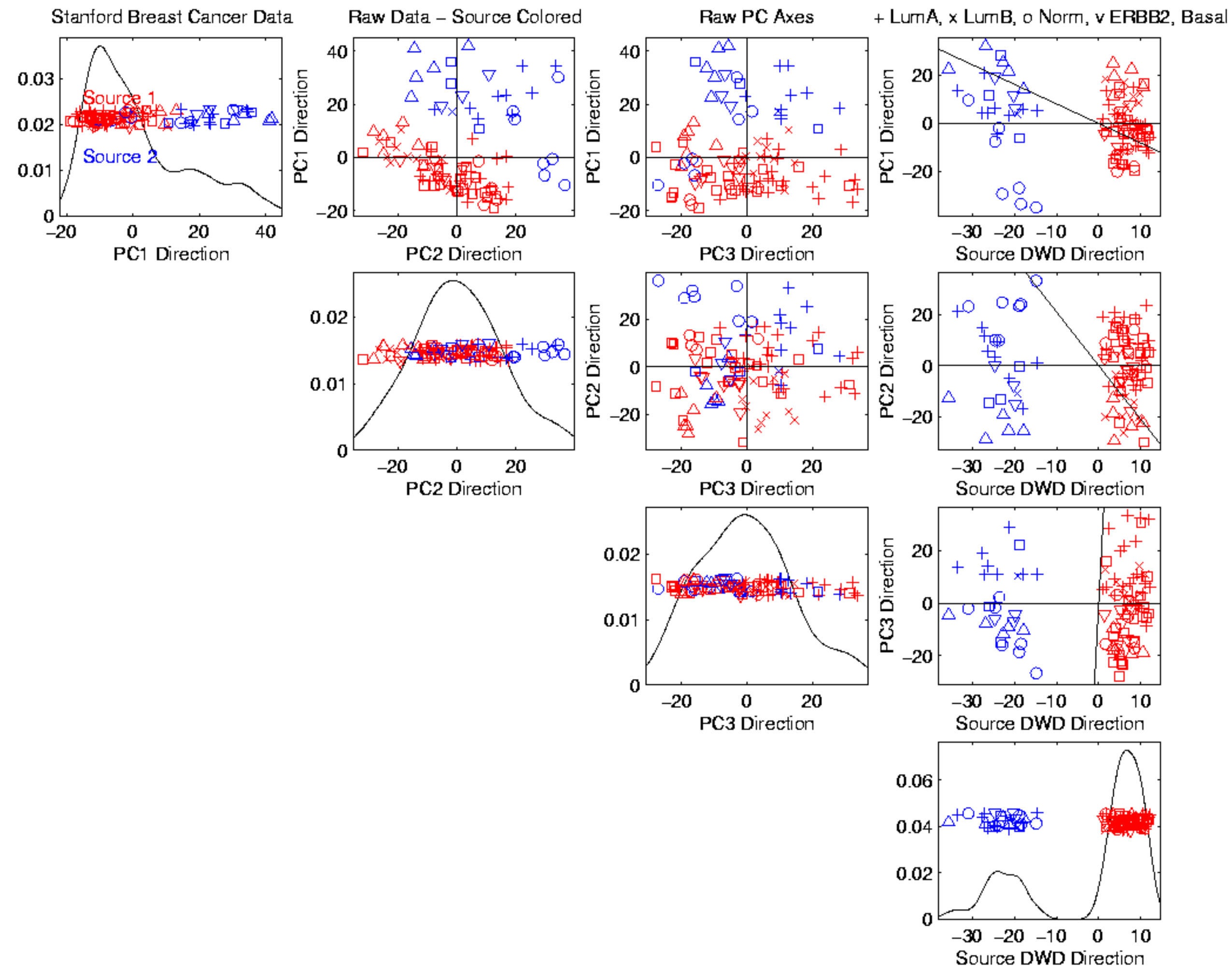
UNC, Stat & OR

DWD Adj: Source Colors

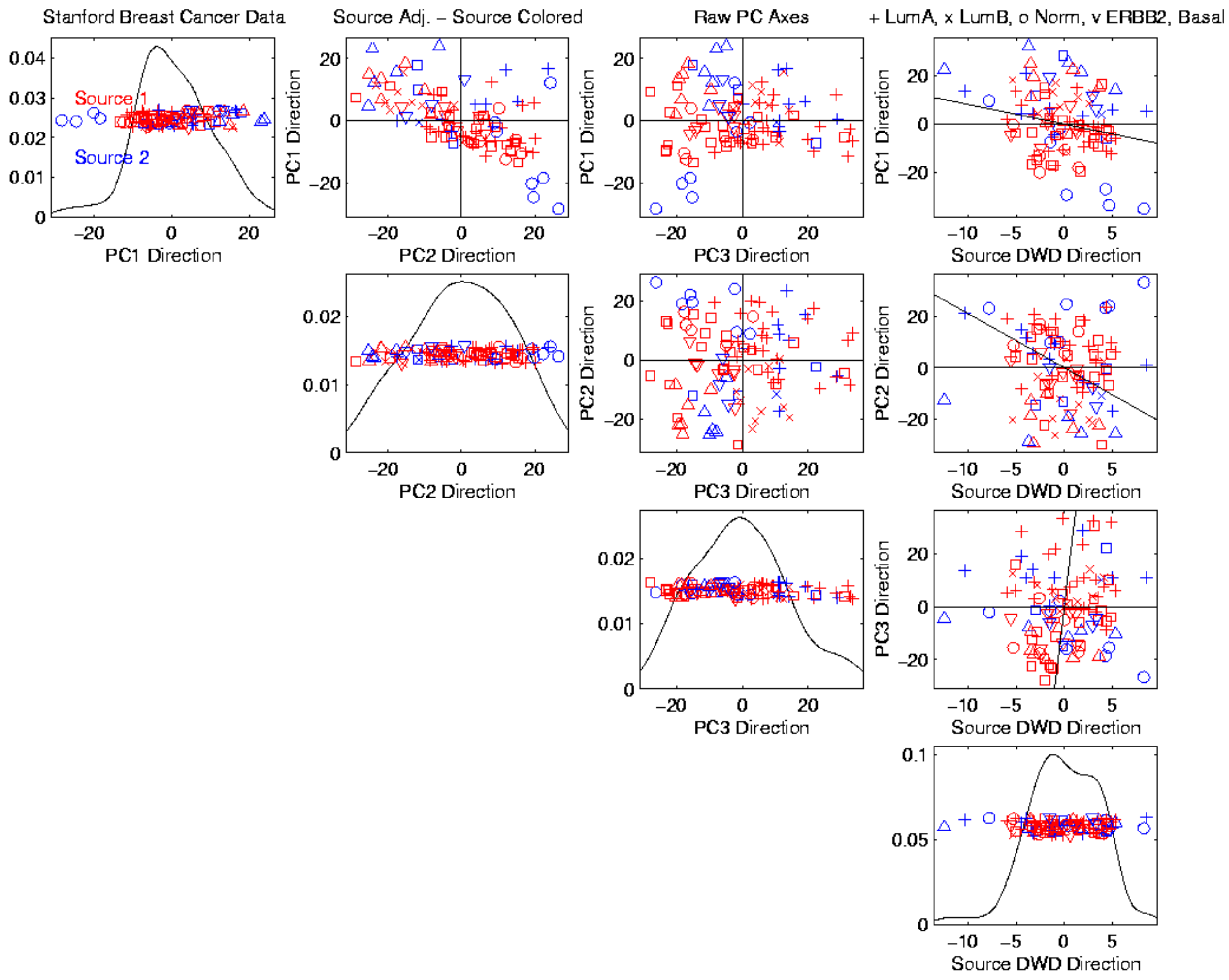


DWD Adj: PC 1-2 & DWD direction

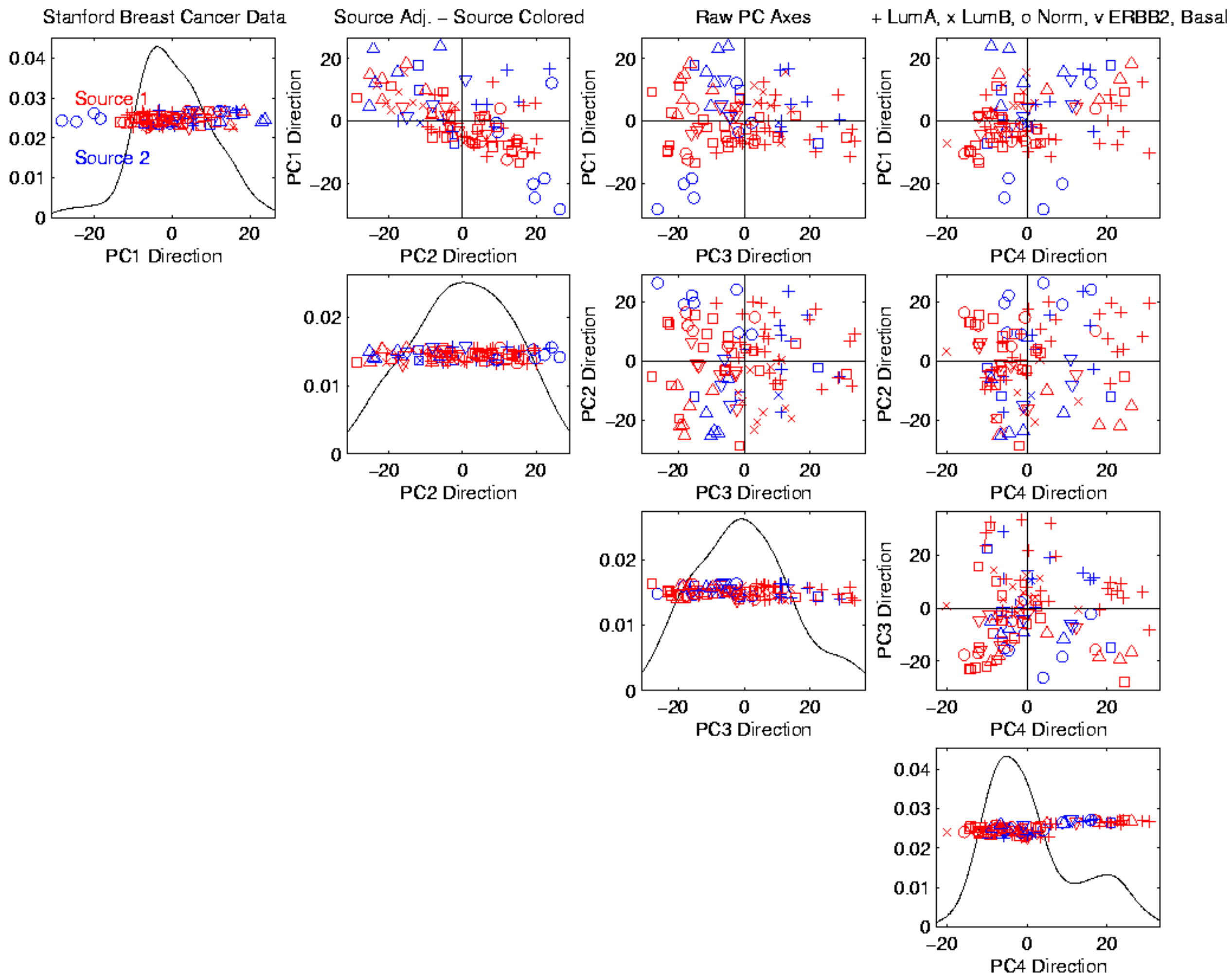
UNC, Stat & OR



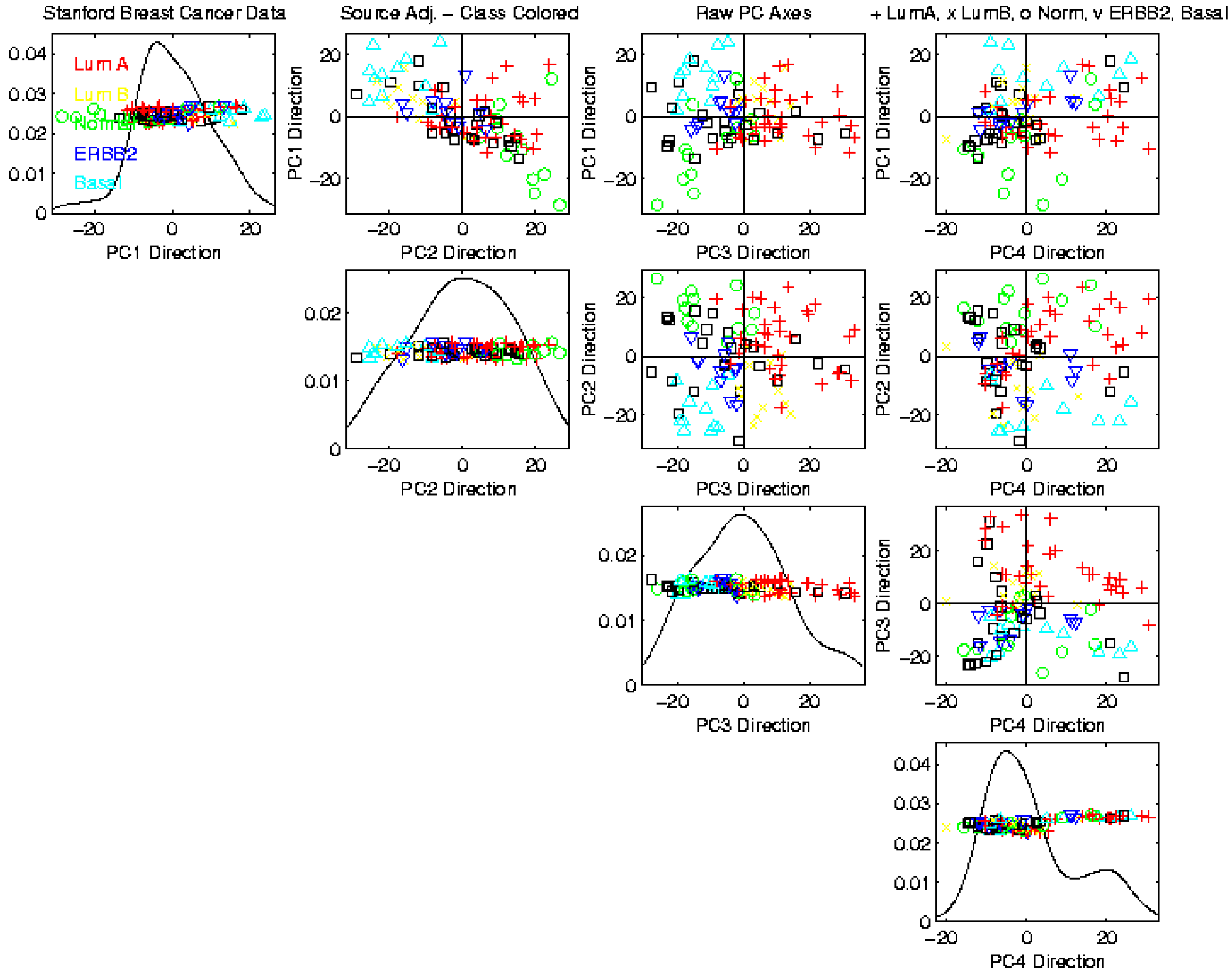
DWD Adj: DWD Source Adjustment



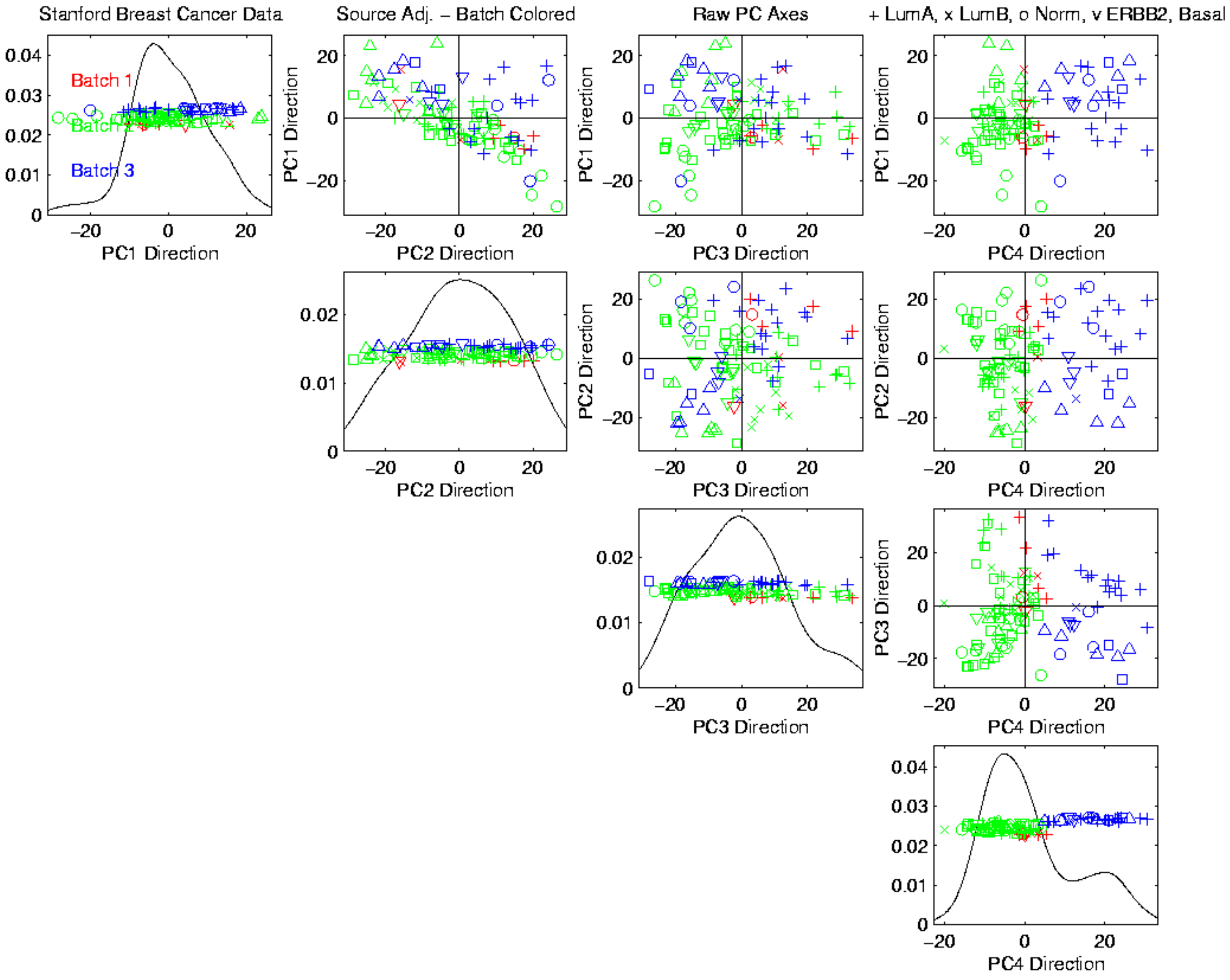
DWD Adj: Source Adj'd, PCA view



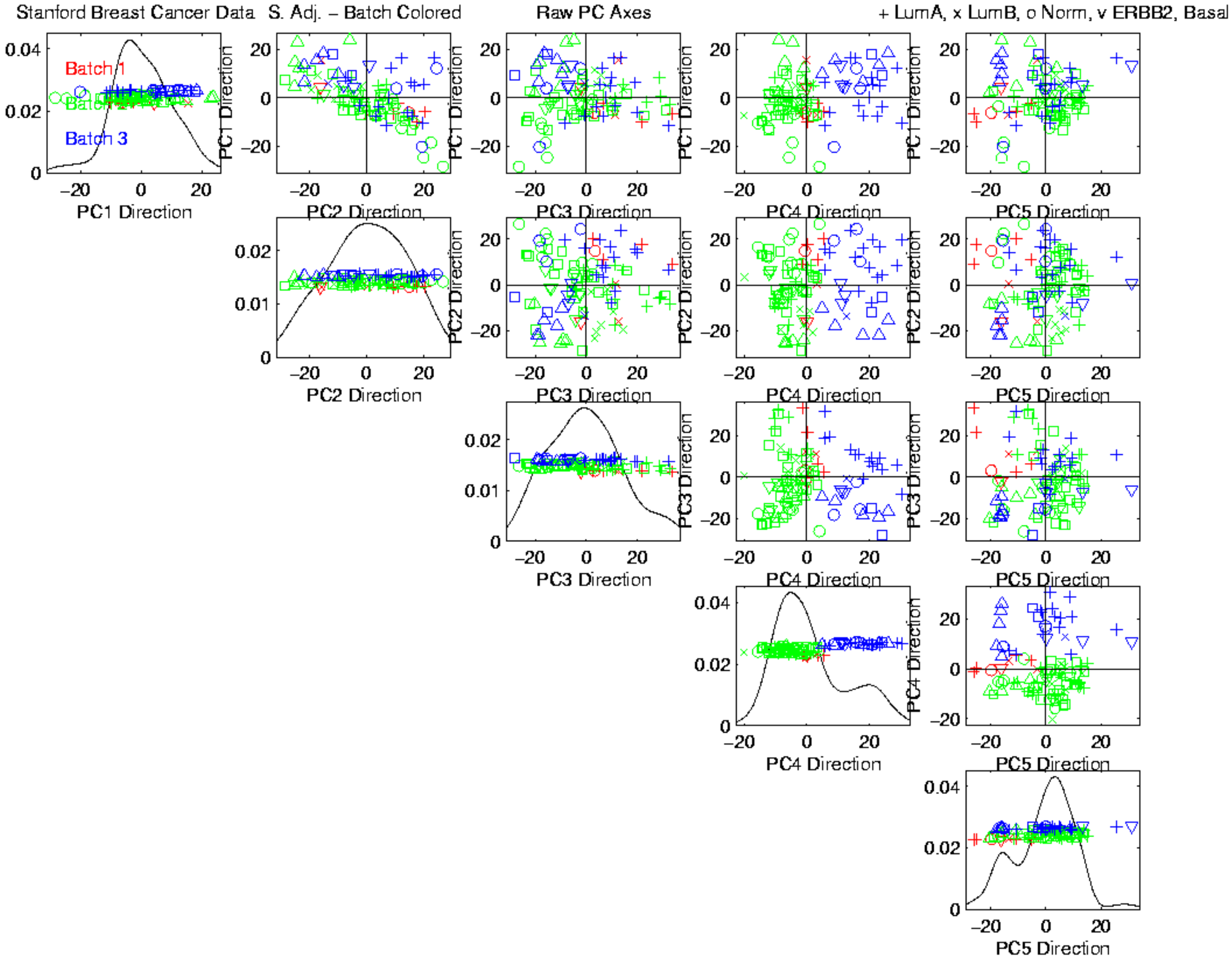
DWD Adj: Source Adj'd, Class Colored



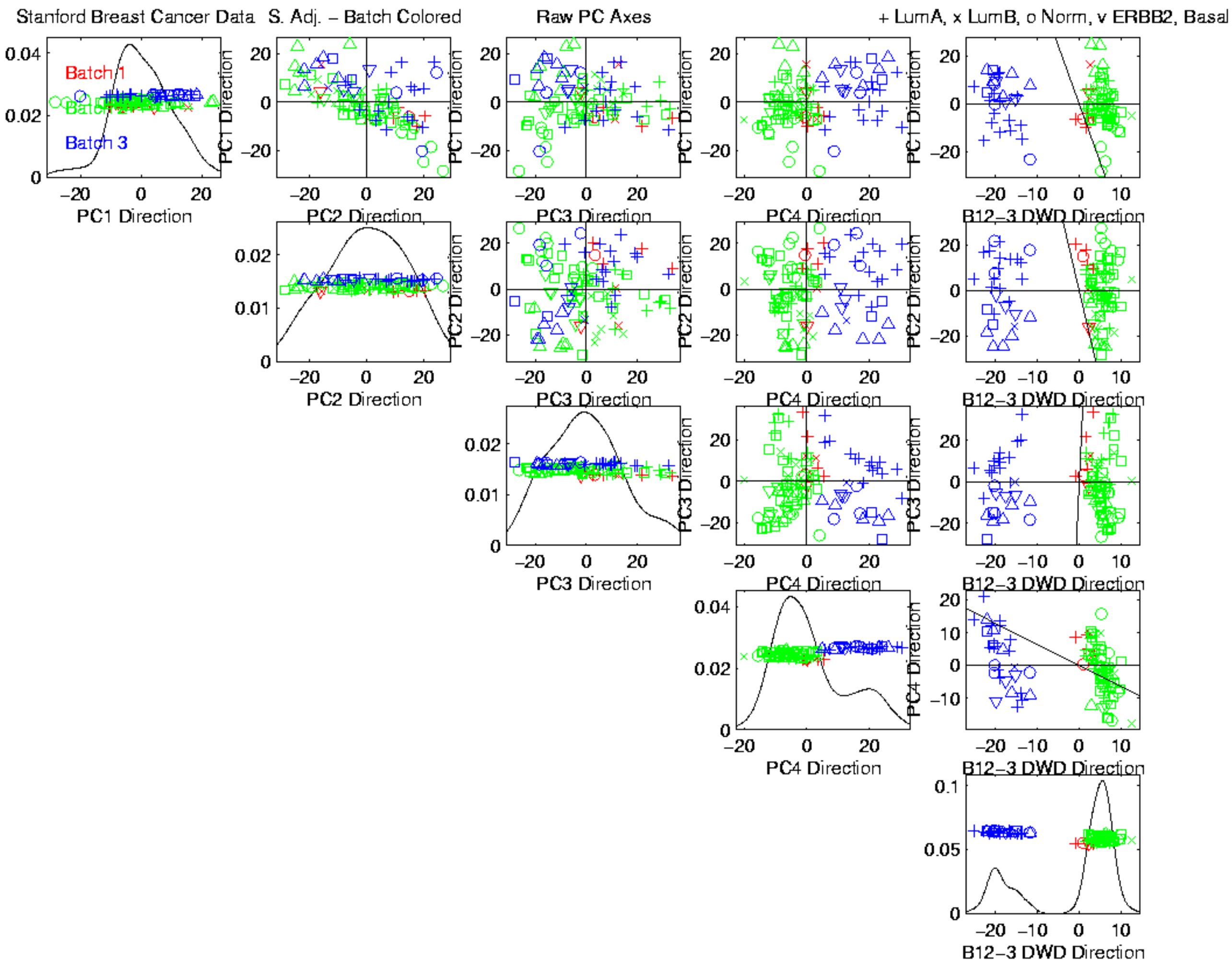
DWD Adj: Source Adj'd, Batch Colored



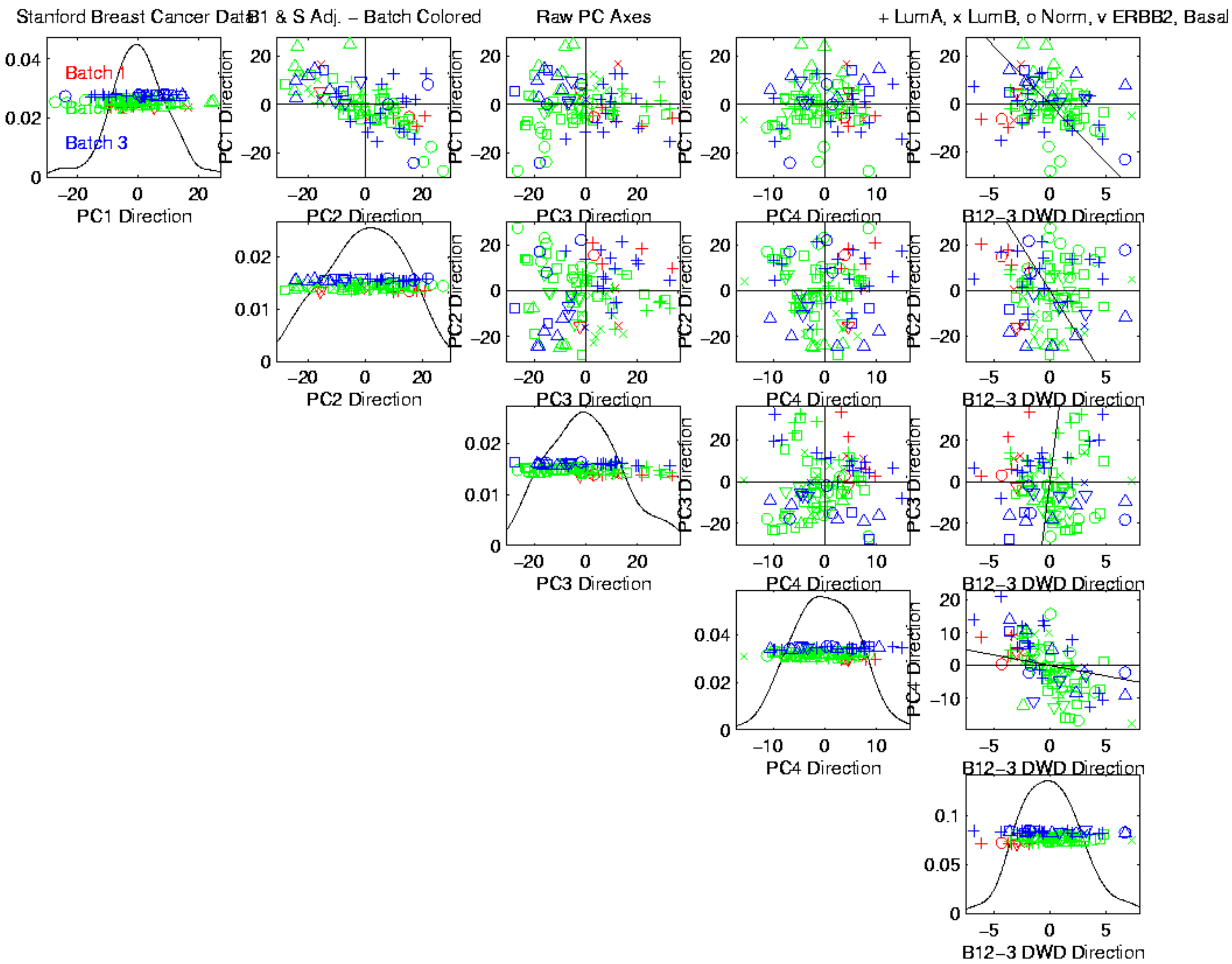
DWD Adj: Source Adj'd, 5 PCs



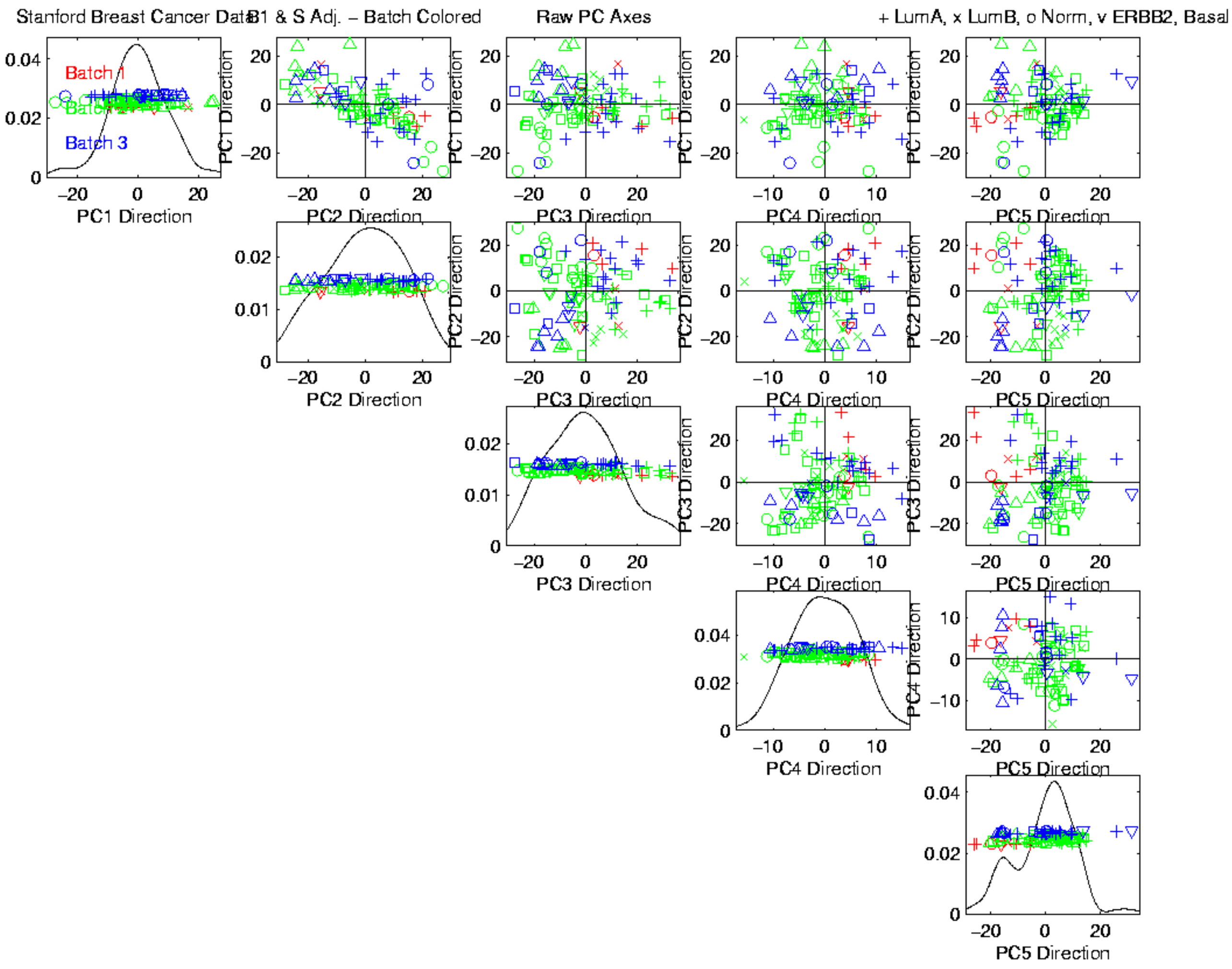
DWD Adj: S. Adj'd, Batch 1,2 vs. 3 DWD



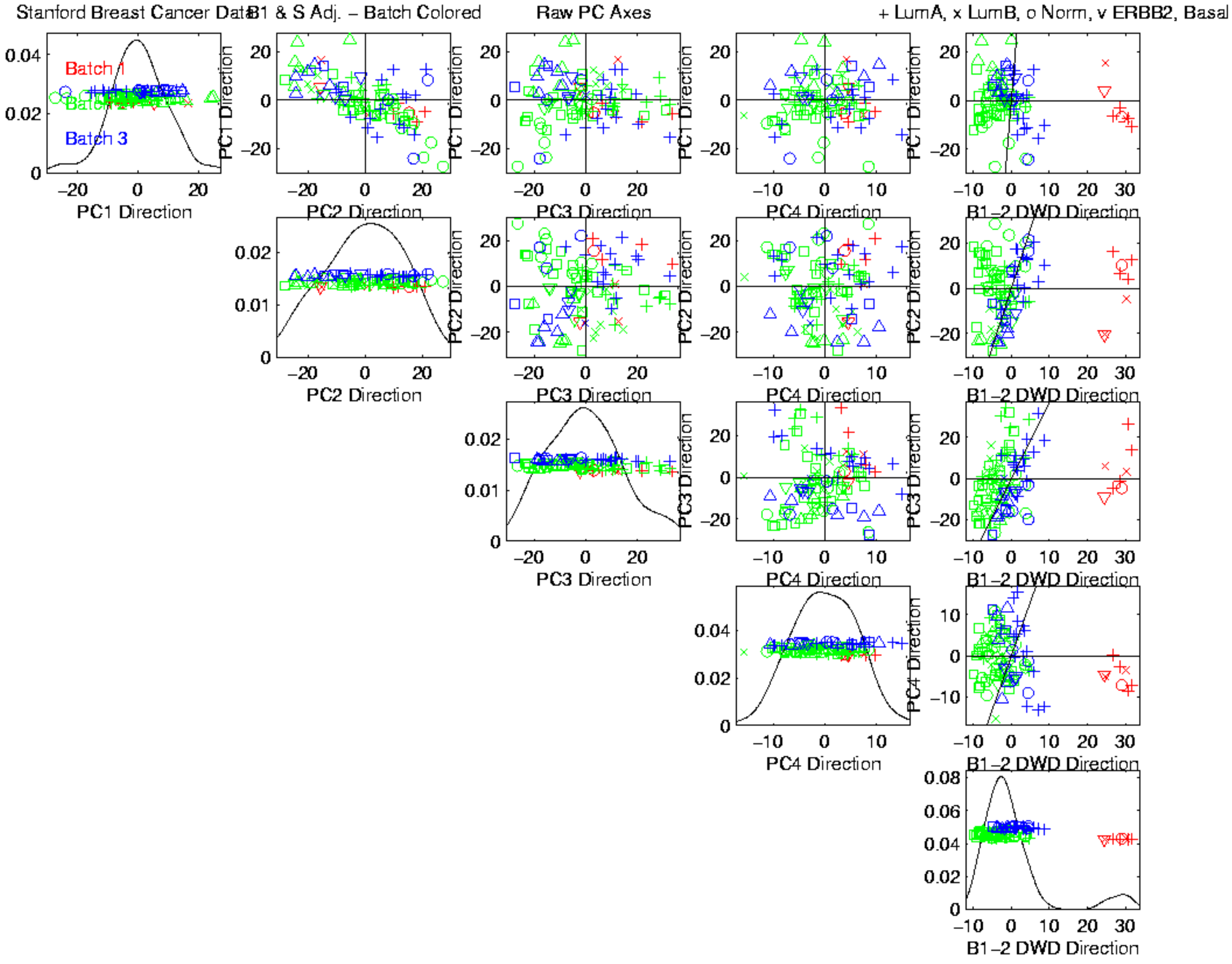
DWD Adj: S. & B1,2 vs. 3 Adjusted



DWD Adj: S. & B1,2 vs. 3 Adj'd, 5 PCs

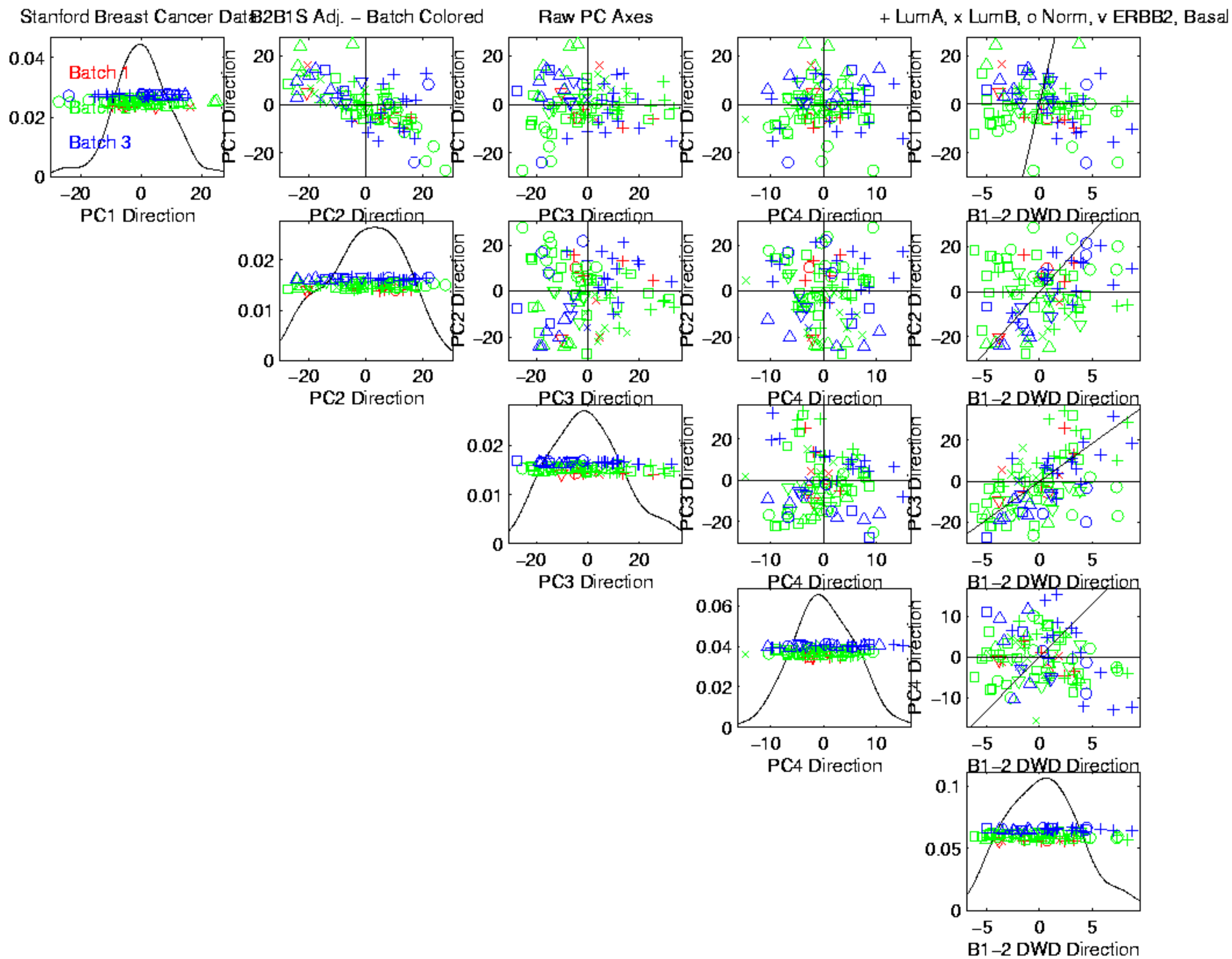


DWD Adj: S. & B Adj'd, B1 vs. 2 DWD

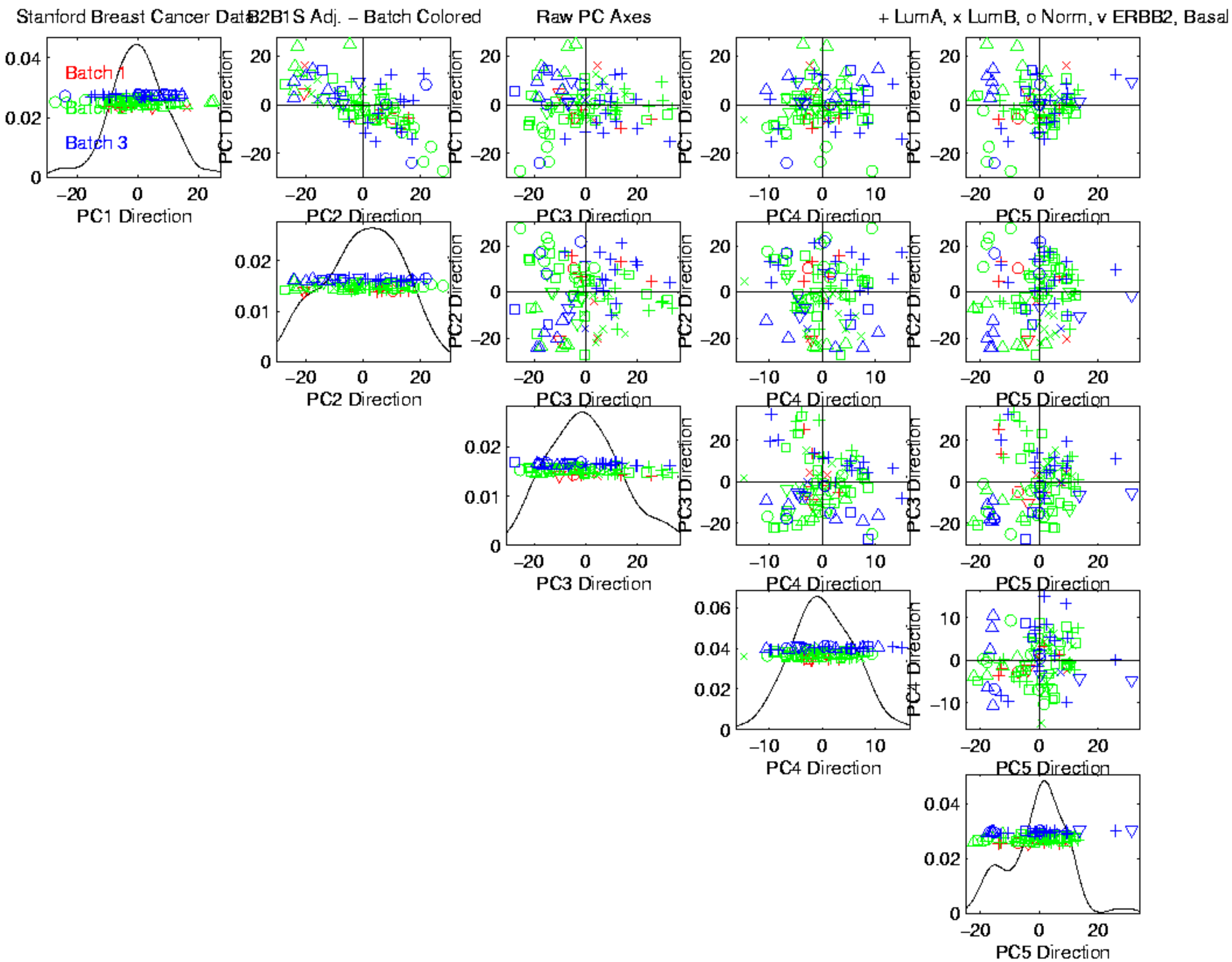


DWD Adj: S. & B Adj'd, B1 vs. 2 Adj'd

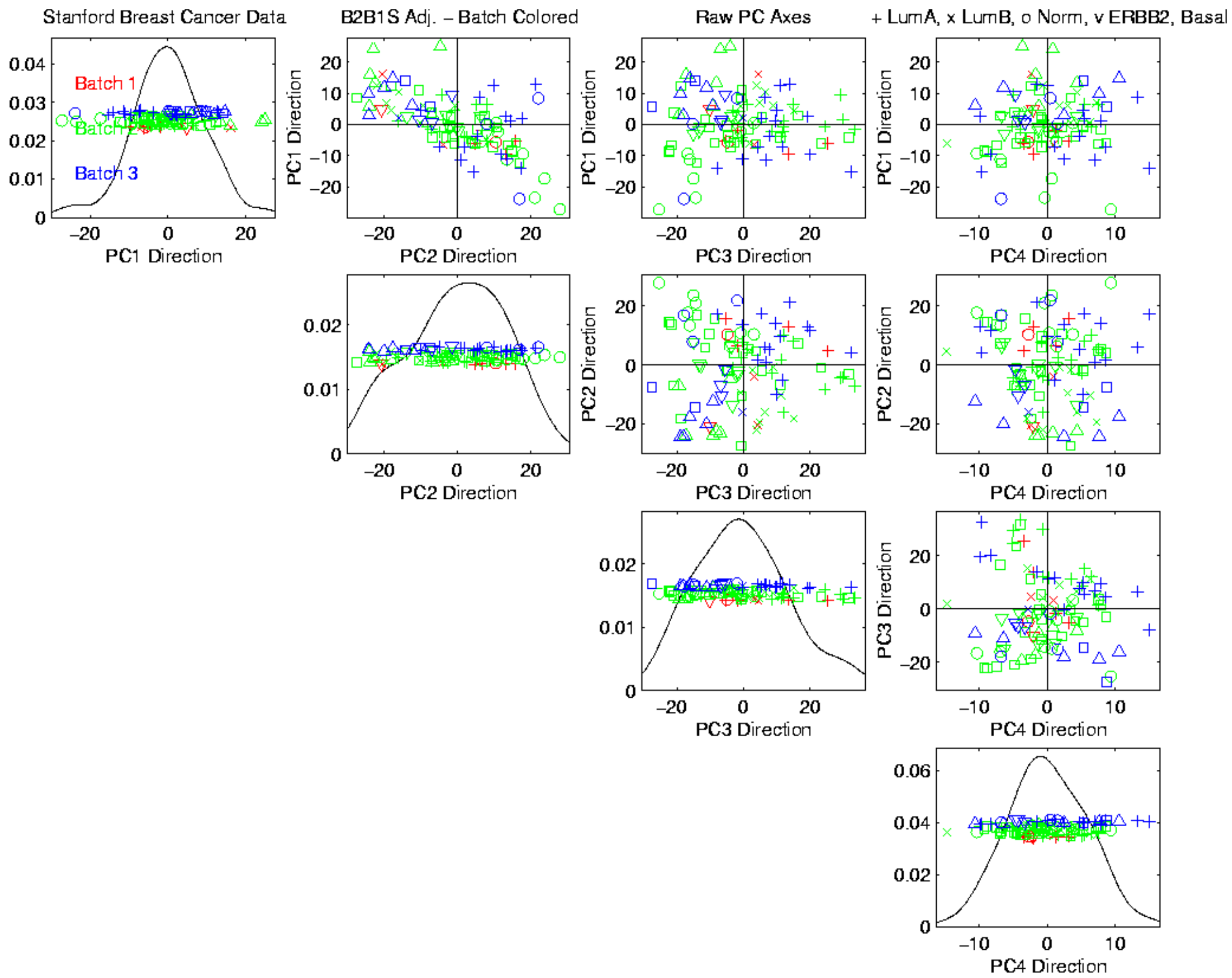
UNC, Stat & OR



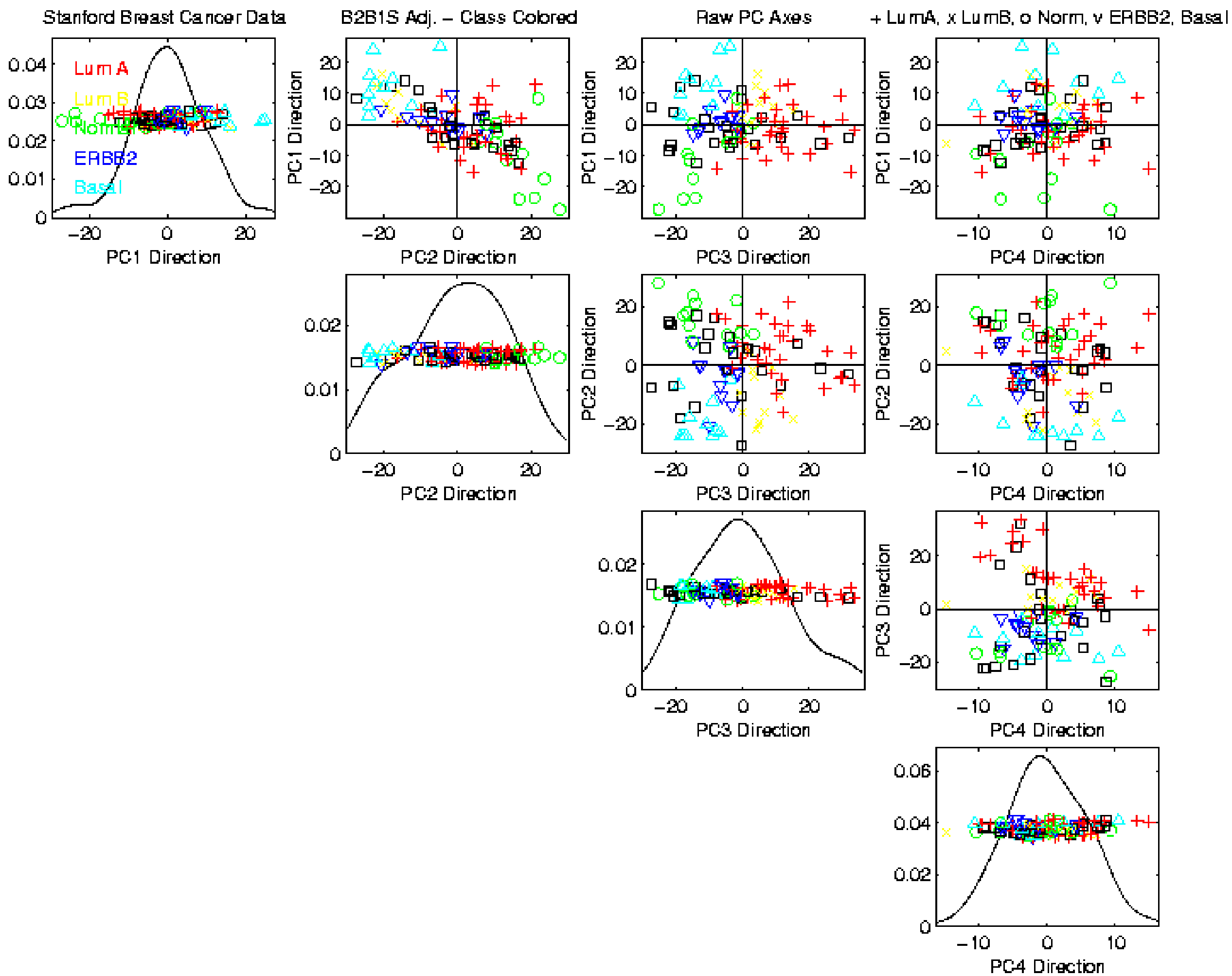
DWD Adj: S. & B Adj'd, 5 PC view



DWD Adj: S. & B Adj'd, 4 PC view

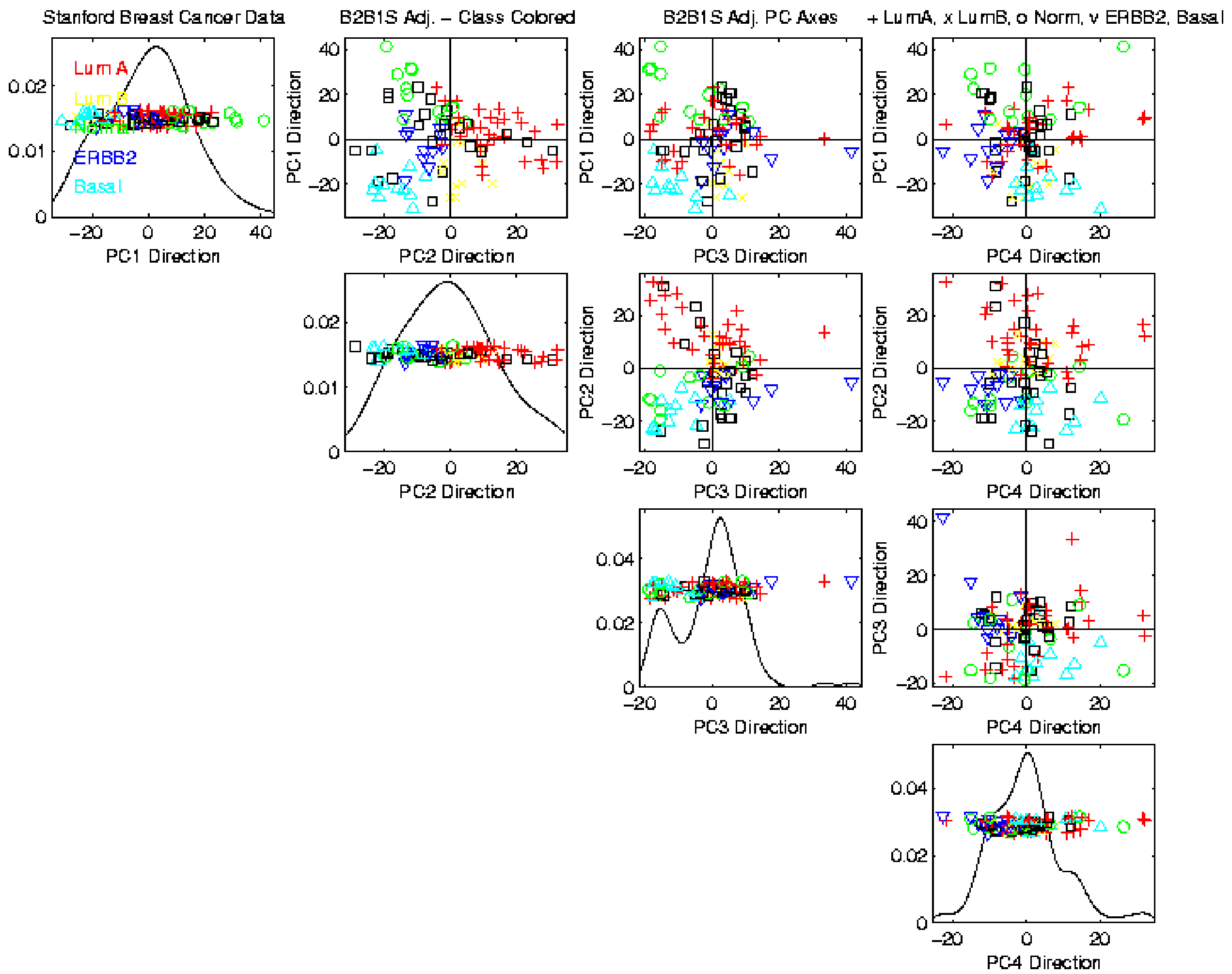


DWD Adj: S. & B Adj'd, Class Colors



DWD Adj: S. & B Adj'd, Adj'd PCA

UNC, Stat & OR



DWD Bias Adjustment for Microarrays

- Effective for Batch and Source Adj.
- Also works for *cross-platform Adj.*
  - E.g. cDNA & Affy
  - Despite literature claiming contrary

"Gene by Gene" vs. "Multivariate" views

- Funded as part of caBIG

"Cancer BioInformatics Grid"

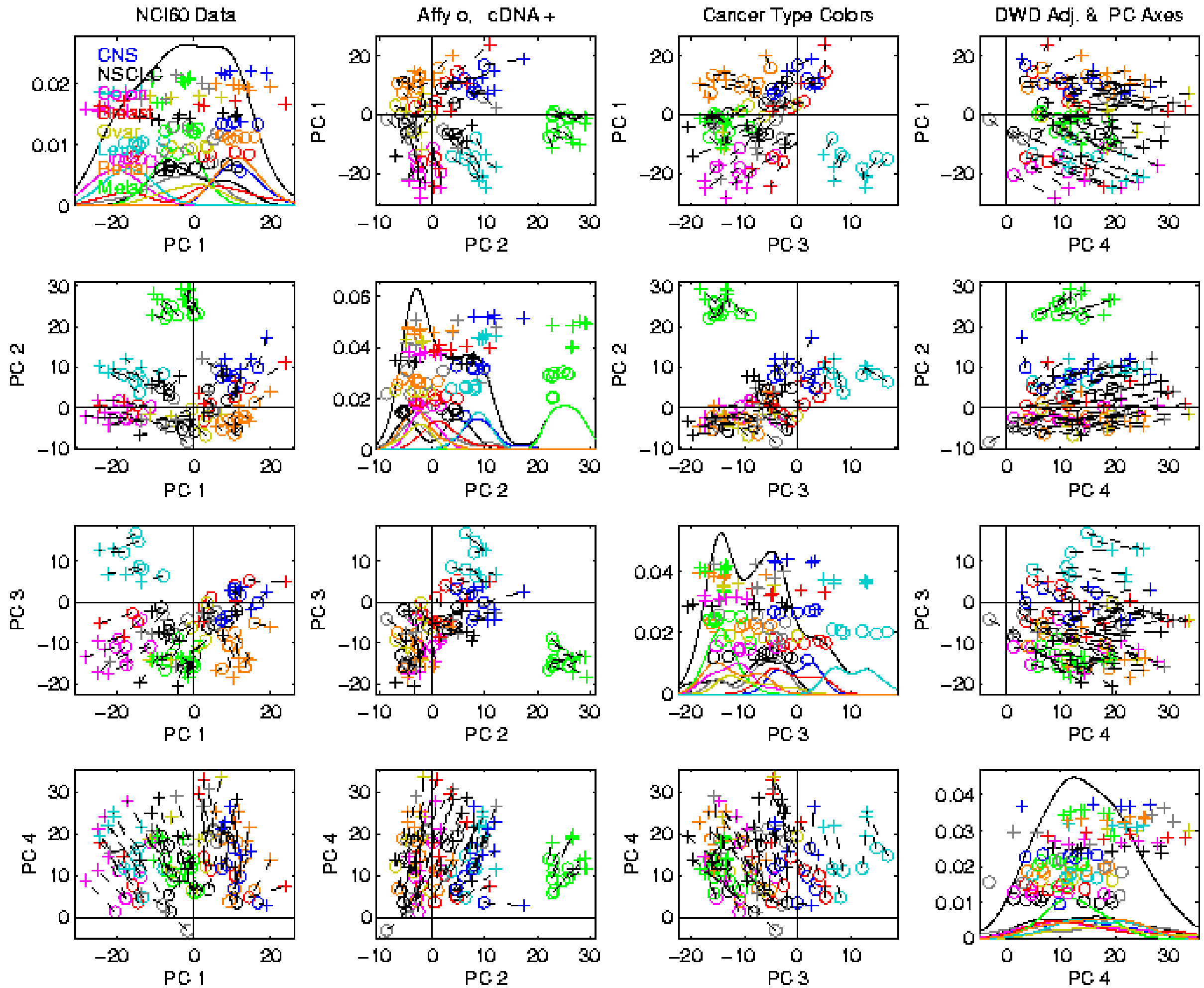
  - "Data Combination Effort" of NCI

Interesting Benchmark Data Set

- NCI 60 Cell Lines
  - Interesting benchmark, since *same* cells
  - Data Web available:

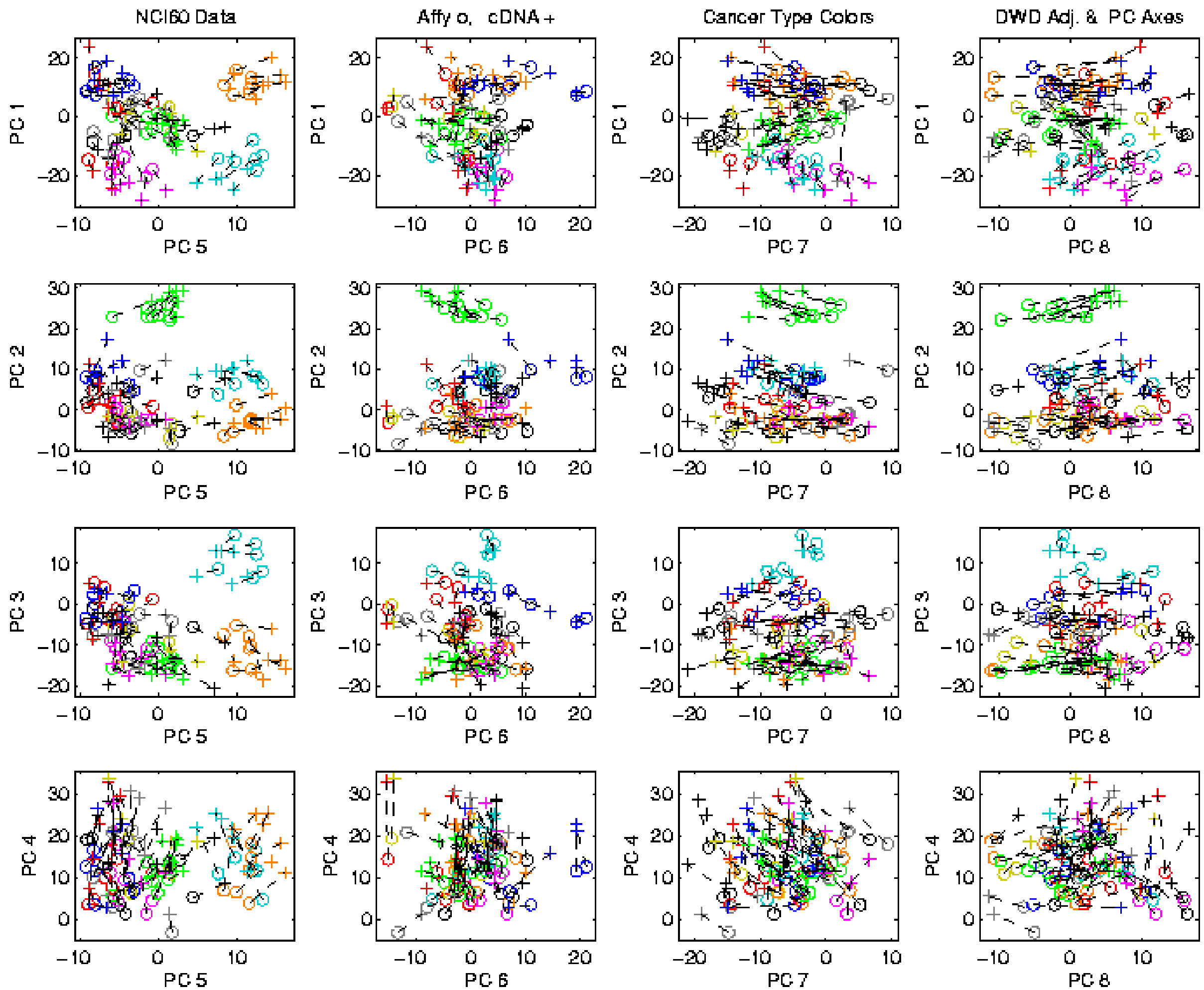
  http://discover.nci.nih.gov/datasetsNature2000.jsp
  - *Both* cDNA and Affymetrix Platforms
- 8 Major cancer subtypes
- Use DWD now for *visualization*

NCI 60: PCA 1-4 View & Subtype Colors

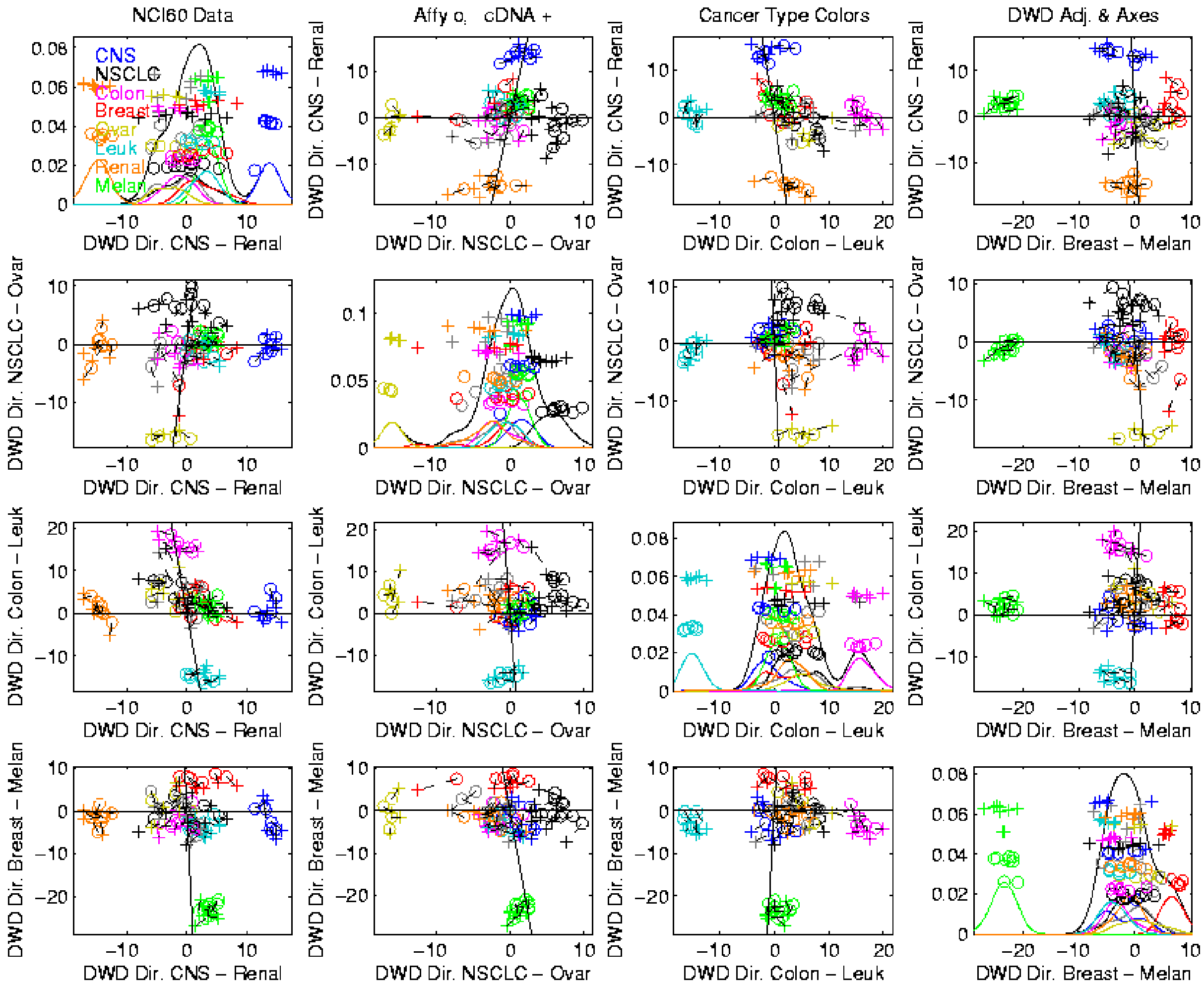


NCI 60: PCA 1-4 vs. 5-8 View & Subtype Colors

UNC, Stat & OR



NCI 60: Views using DWD Dir'ns (focus on biology)



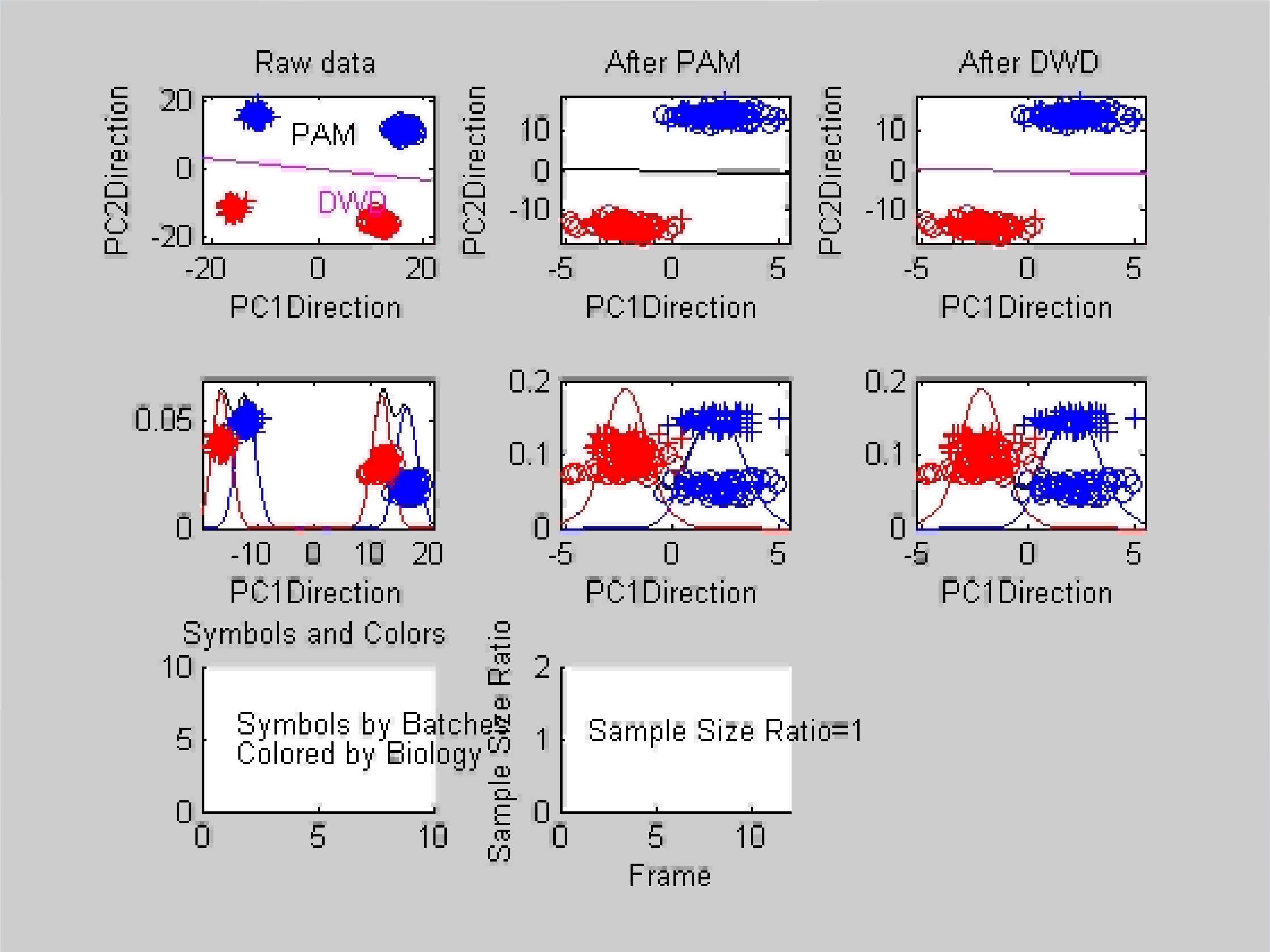
Why not adjust by means?

- DWD is complicated: value added?
- Xuxin Liu example…
- Key is sizes of biological subtypes
- Differing ratio trips up mean
- But DWD more robust

(although still not perfect)

UNC, Stat & OR

Twiddle ratios of subtypes



DWD in Face Recognition, I

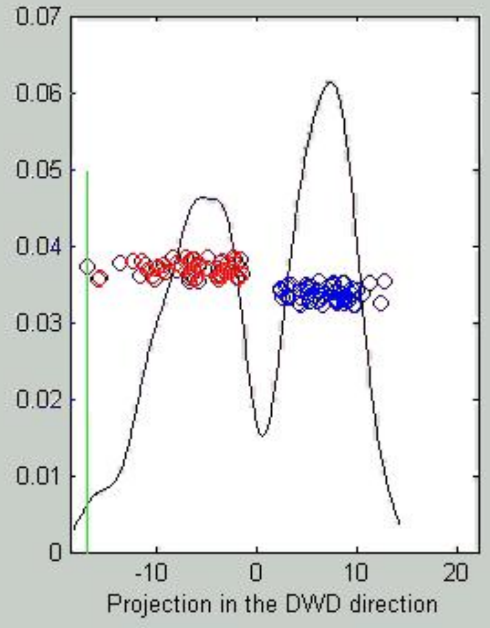
- Face Images as Data

  (with M. Benito & D. Peña)
- Registered using landmarks
- Male – Female Difference?
- Discrimination Rule?

DWD in Face Recognition, II

- DWD Direction
- Good separation
- Images "make sense"
- Garbage at ends?

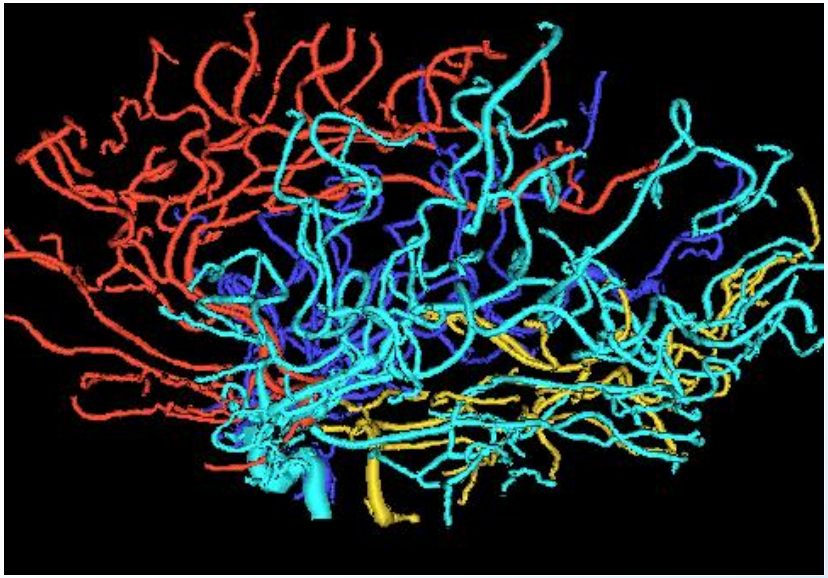
(extrapolation effects?)



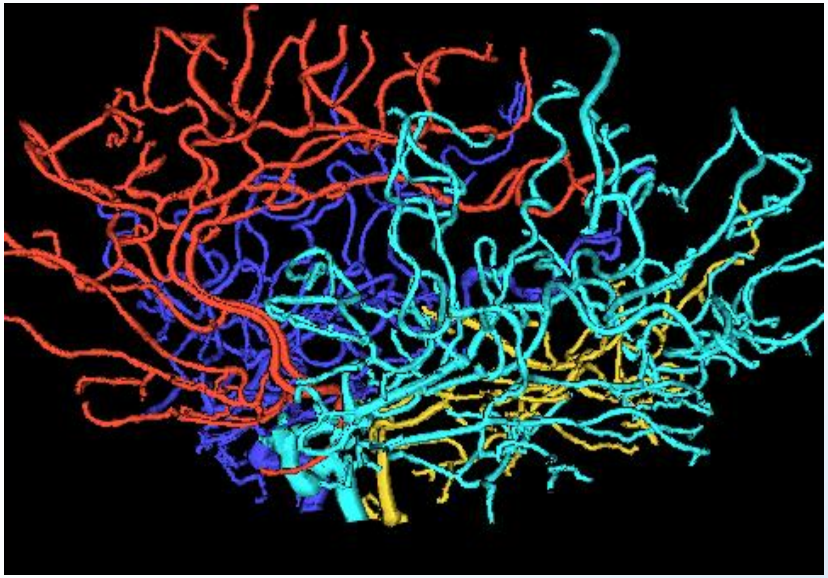
Blood vessel tree data

Marron's brain:

- Segmented from MRA
- Reconstruct trees
- in 3d
- Rotate to view



Blood vessel tree data



Marron’s brain:

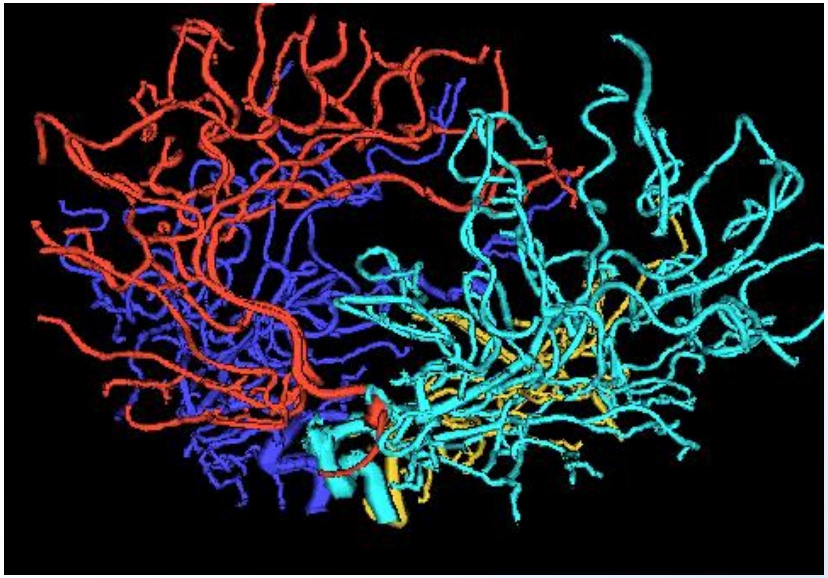
- Segmented from MRA
- Reconstruct trees
- in 3d
- Rotate to view

Blood vessel tree data

UNC, Stat & OR

Marron's brain:

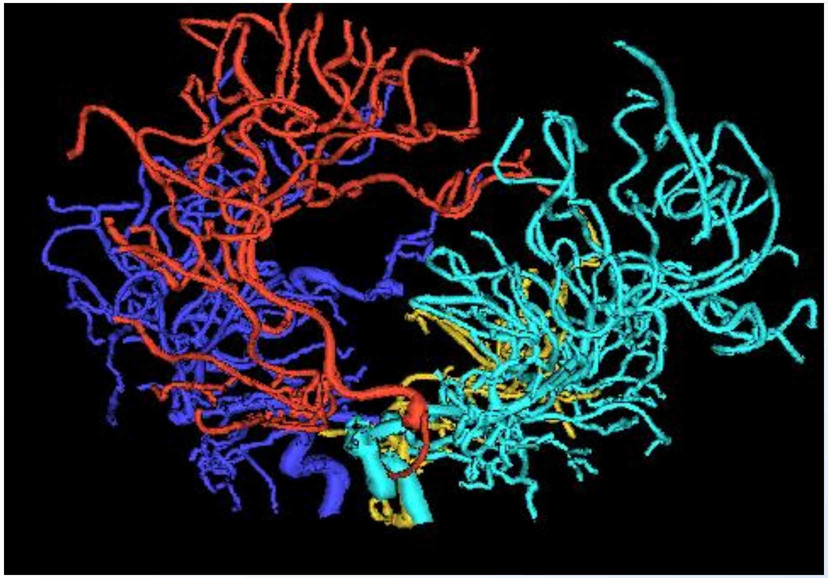
- Segmented from MRA
- Reconstruct trees
- in 3d
- Rotate to view



Blood vessel tree data

Marron's brain:

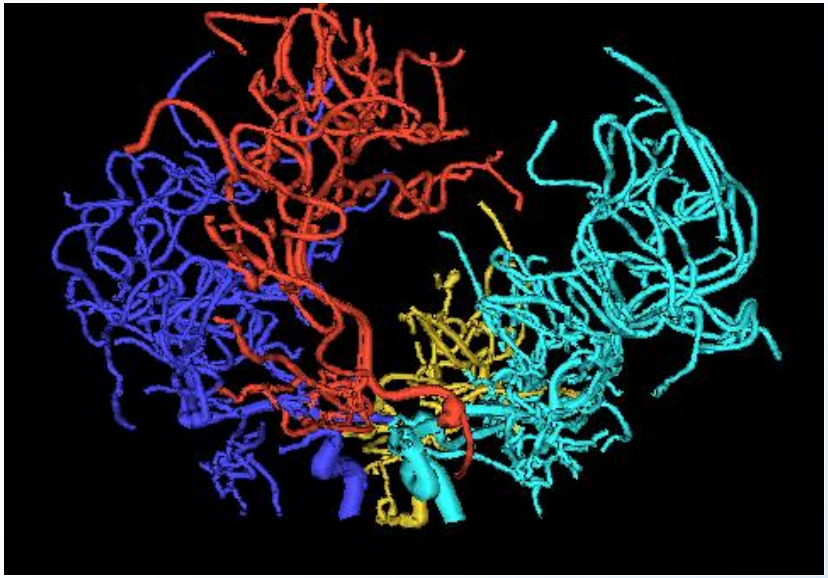
- Segmented from MRA
- Reconstruct trees
- in 3d
- Rotate to view



Blood vessel tree data

Marron's brain:

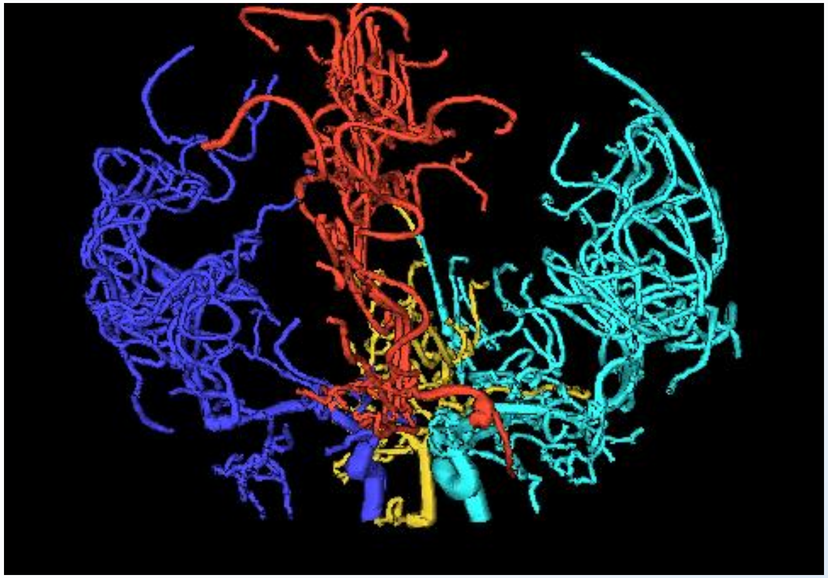
- Segmented from MRA
- Reconstruct trees
- in 3d
- Rotate to view



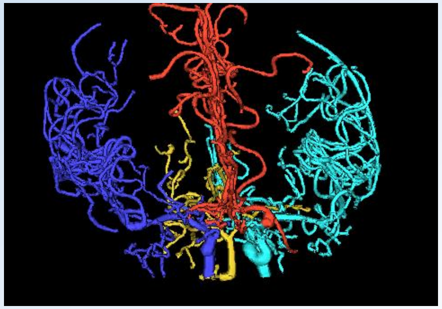
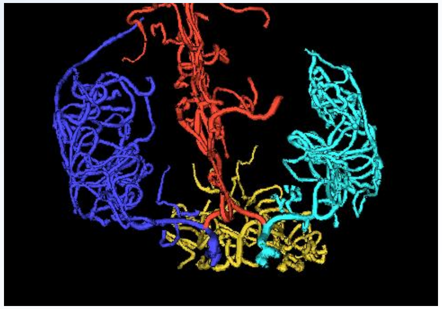
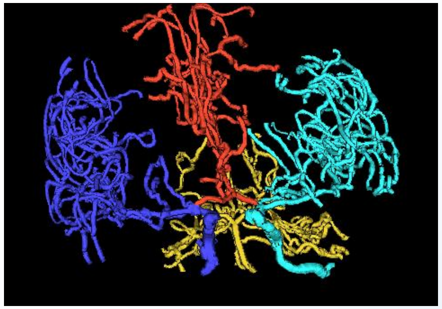
Blood vessel tree data

Marron's brain:

- Segmented from MRA
- Reconstruct trees
- in 3d
- Rotate to view



Blood vessel tree data

 ,  , … , 

Now look over many people (data objects)

Structure of population (understand variation?)

PCA in <u>strongly non-Euclidean</u> Space???

Blood vessel tree data

Big Picture: 4 Approaches

1. Purely Combinatorial

2. Euclidean Orthant

3. Harris Correspondence

4. Persistent Homologies

Time Series of Data Objects

Mortality Data Illustrates an Important Point:

OODA is more than a “framework”

It Provides a Focal Point

Highlights Pivotal Choice:

What should be the Data Objects?

Time Series of Data Objects

Another Interesting Data Set:

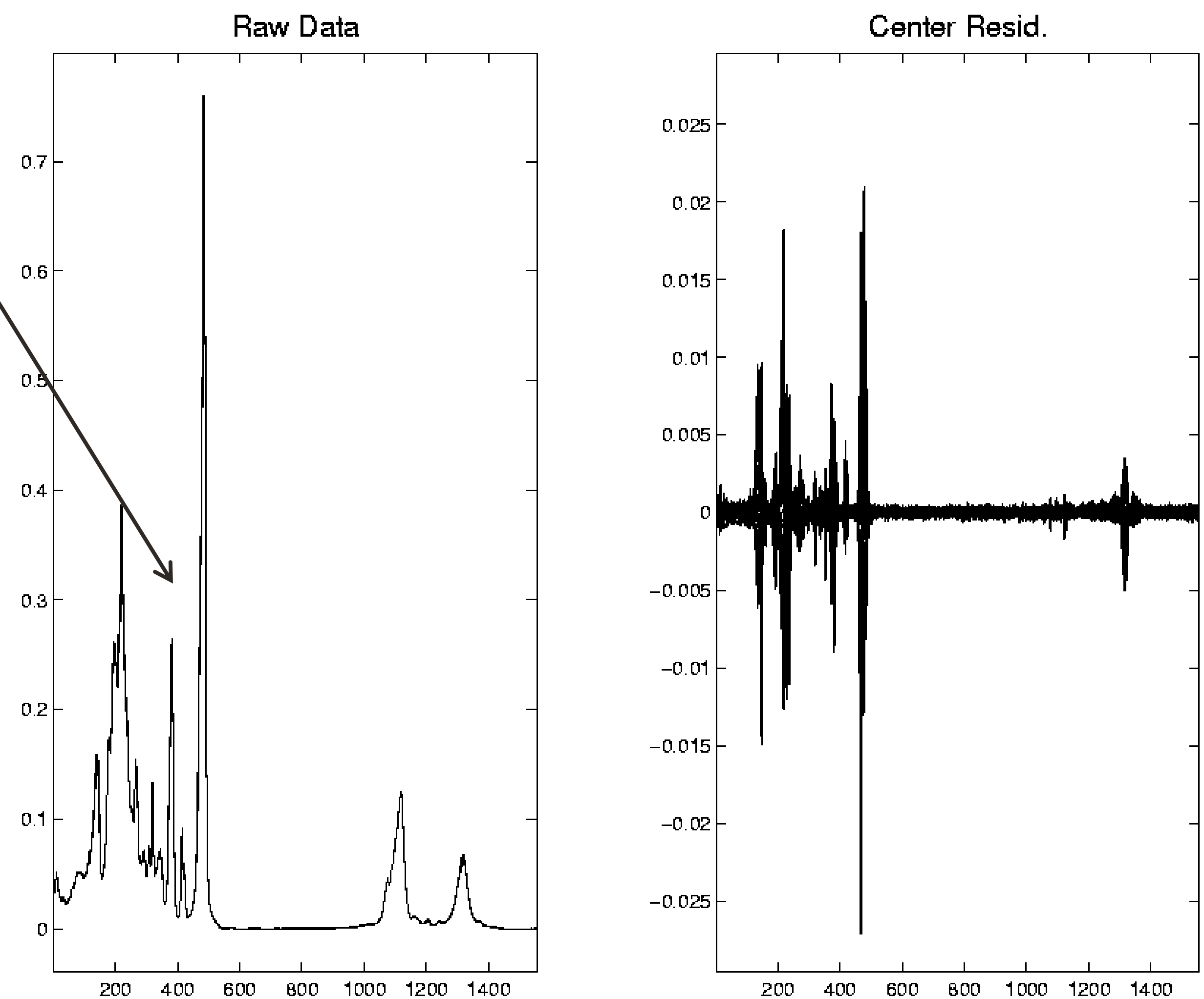
- Chemical Spectra
- Evolving over time
- Studying aging of compounds
- Under different conditions
- From Ed Kober, LANL

Time Series of Chemical Spectra

77 Spectra

Hard to See Them

(because of small differences and large dynamic range)

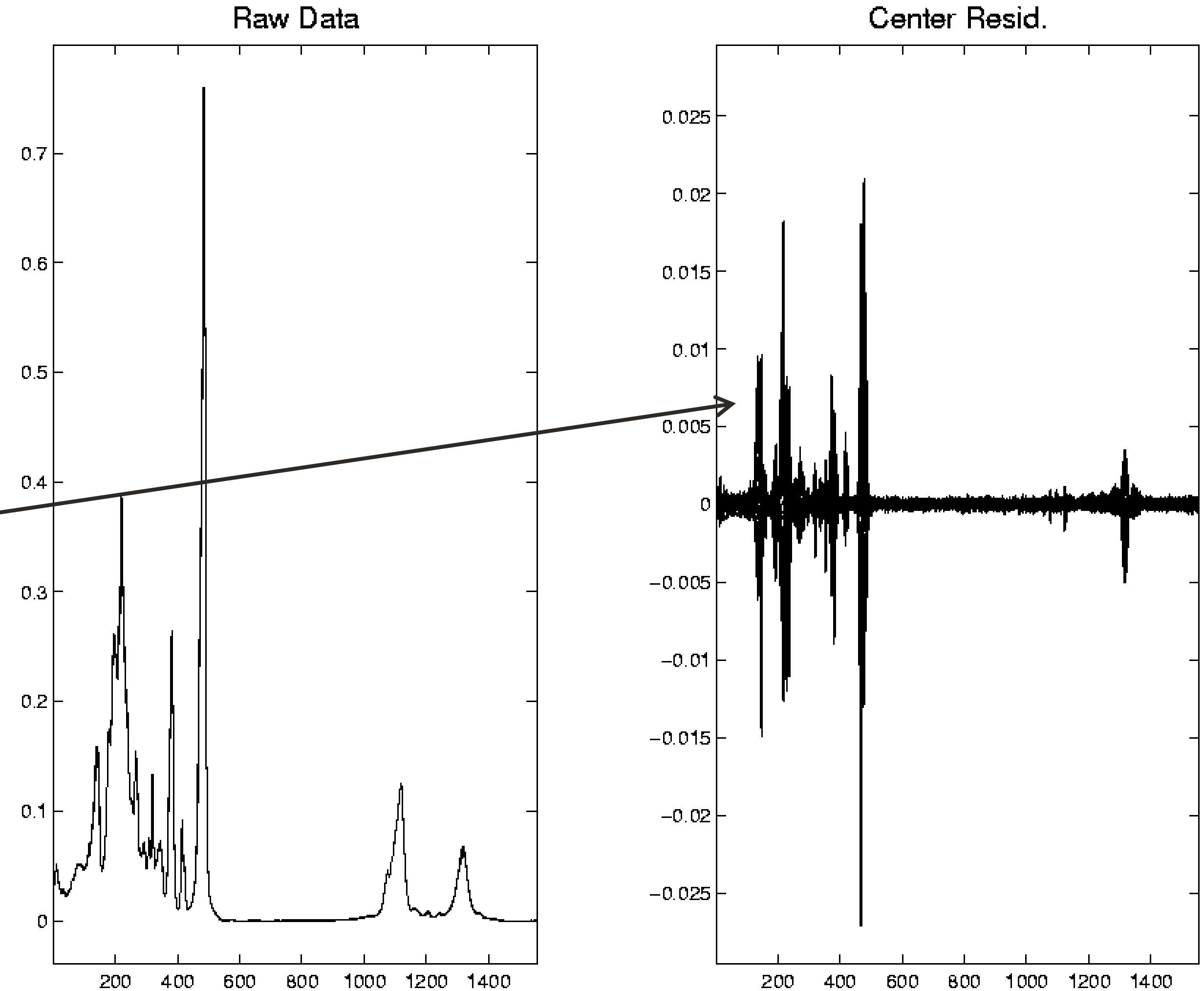


Time Series of Chemical Spectra

77 Spectra

Looking at Mean Residuals Helps

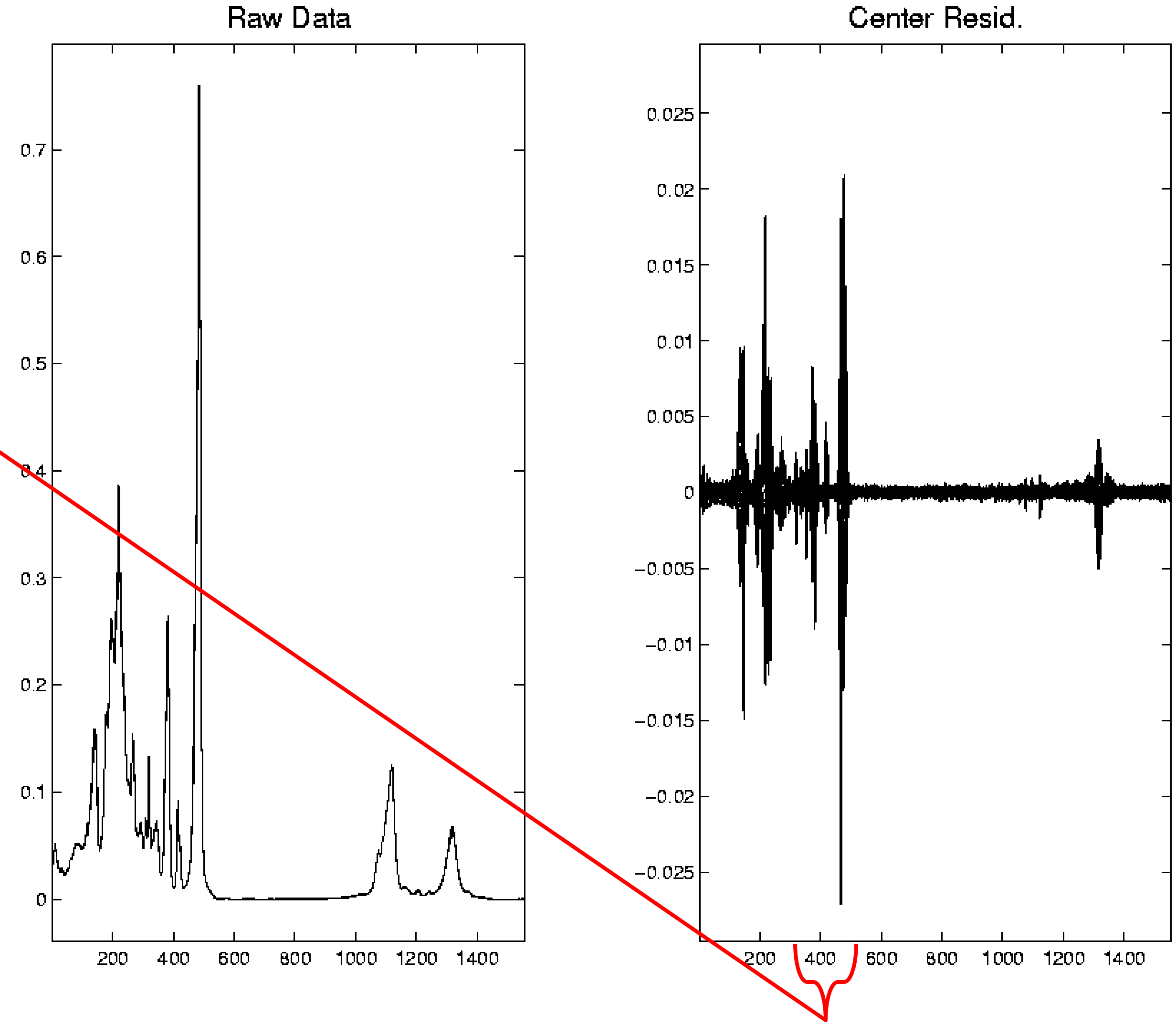
(But Not Much)



Time Series of Chemical Spectra

UNC, Stat & OR

Try Zooming In On an "Interesting Window"

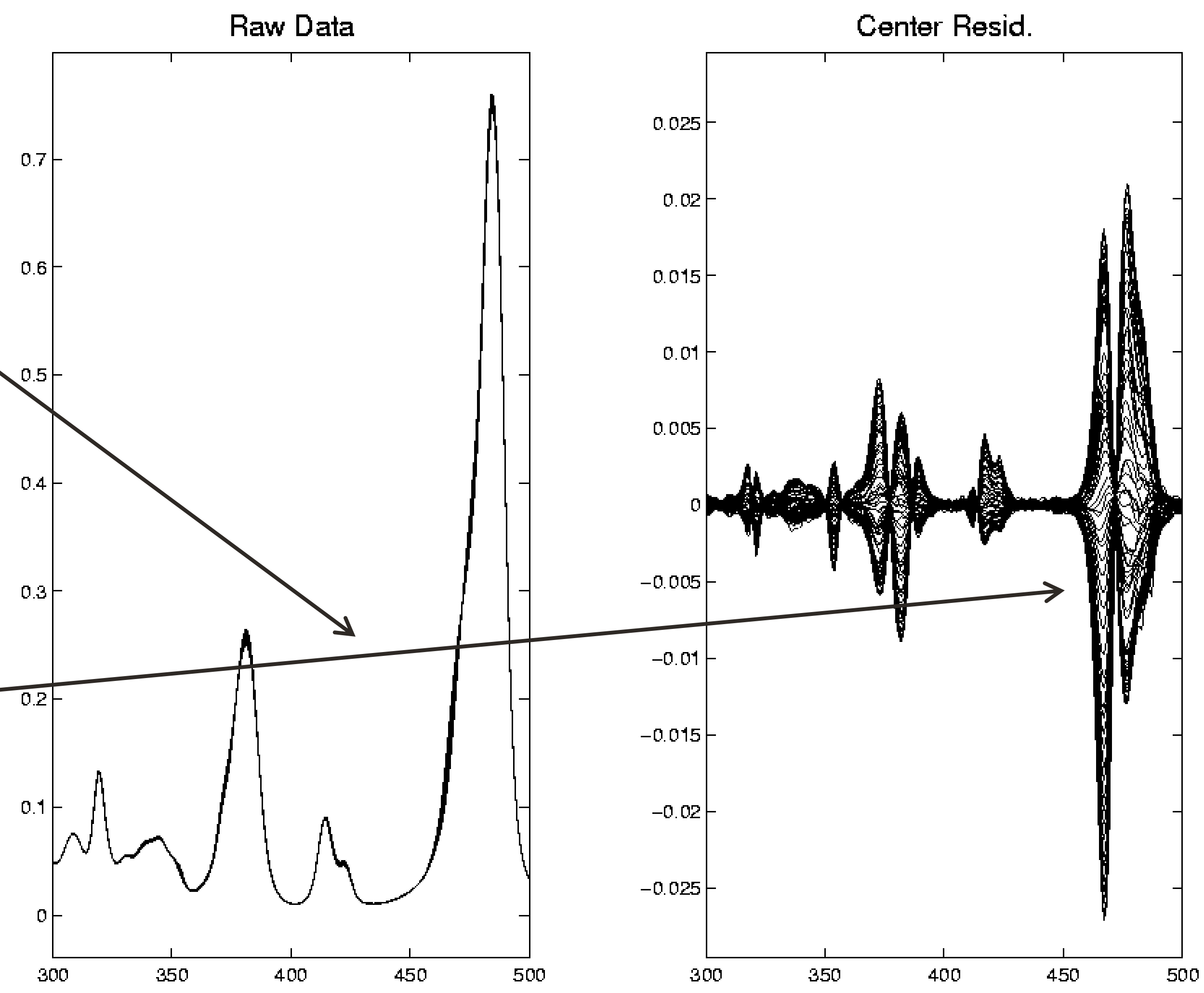


Time Series of Chemical Spectra

77 Spectra

Structure Still Lost in Dynamic Range

But Visible In Mean Residuals

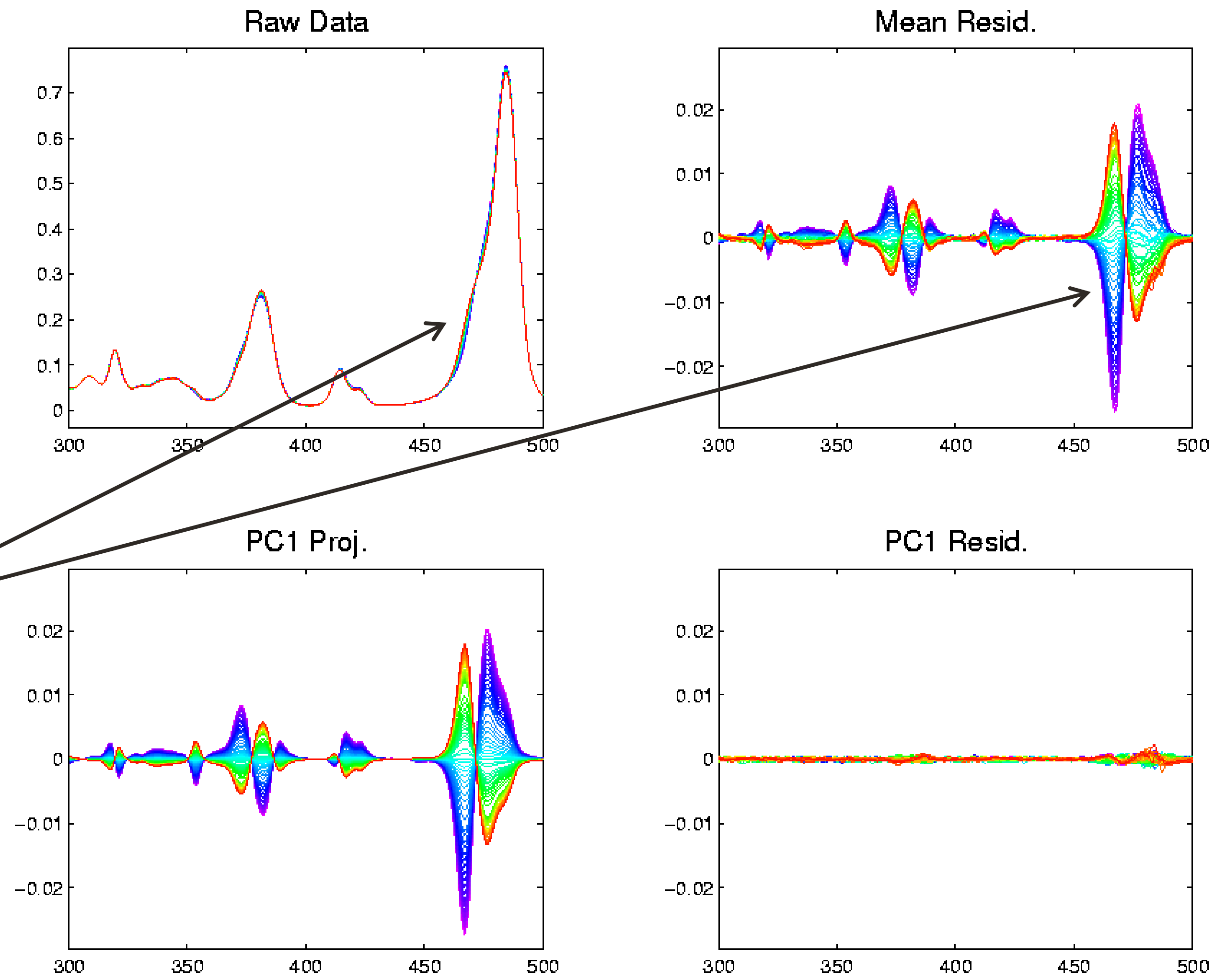


Time Series of Chemical Spectra

Time Colors Again Very Helpful

Shows Up & Down Behavior

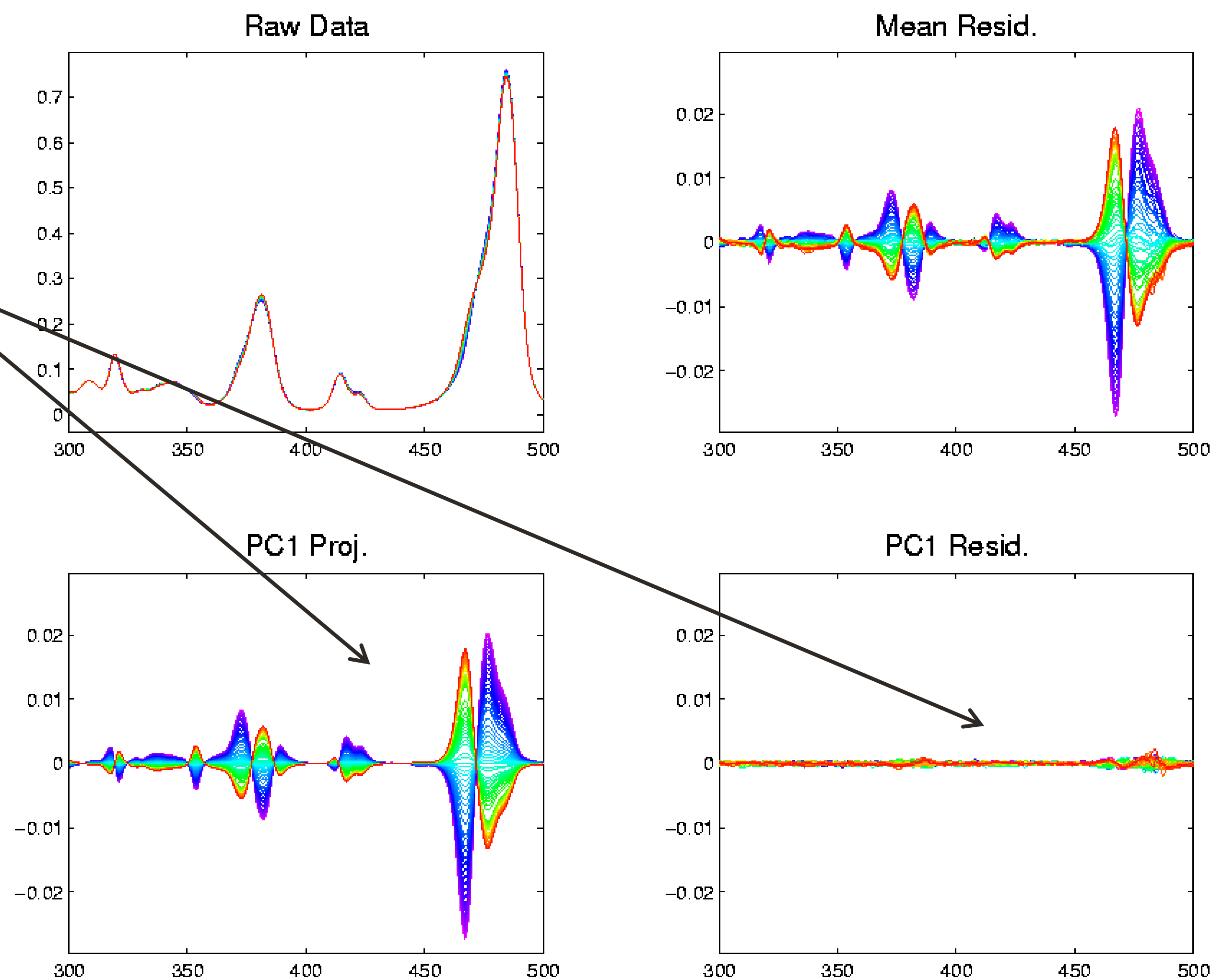
(Movement Of Mass)



Time Series of Chemical Spectra

Note PC1
Is Most Of
Variation

(Mostly
Single
Reaction)

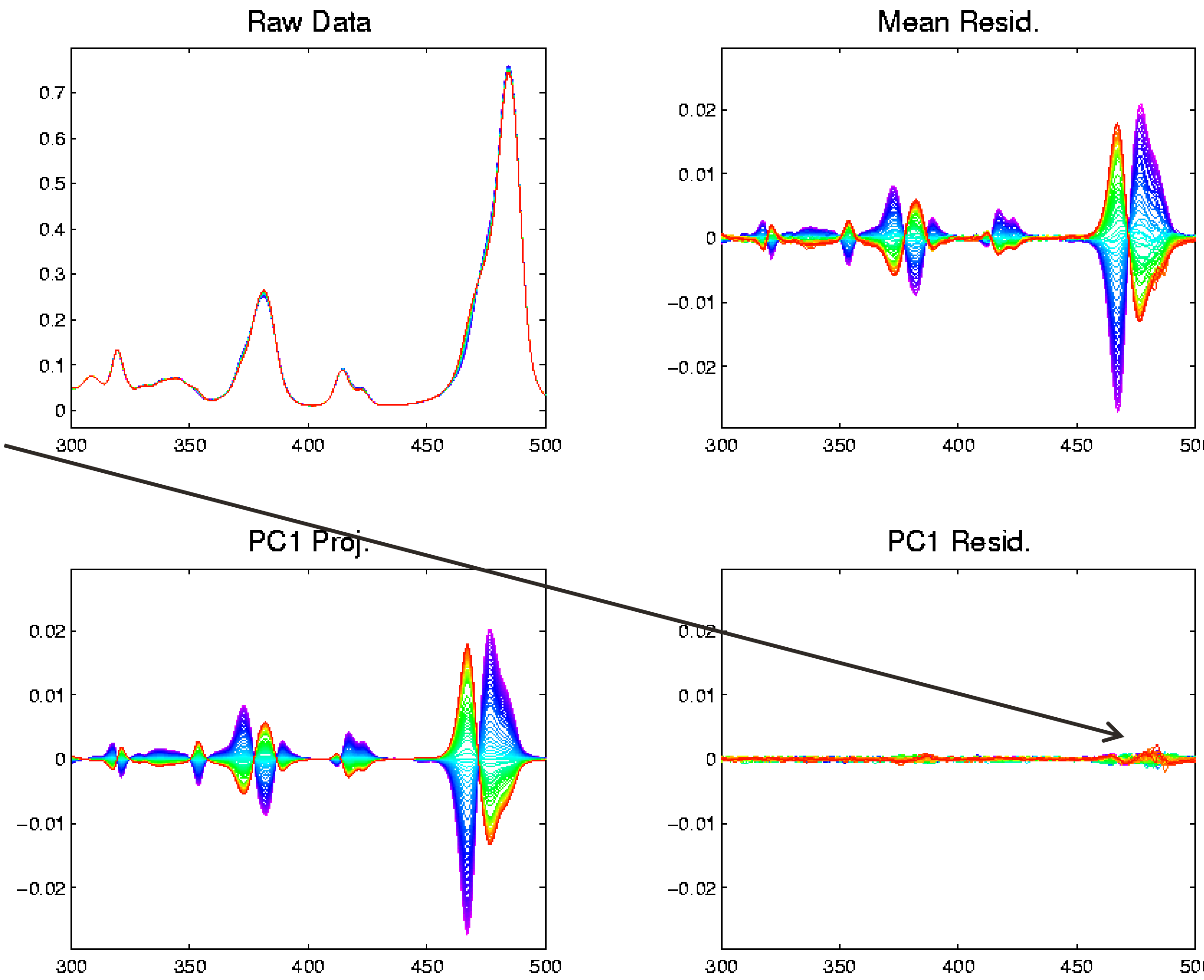


Time Series of Chemical Spectra

UNC, Stat & OR

Anything Important Beyond This?

Study Scores Plot

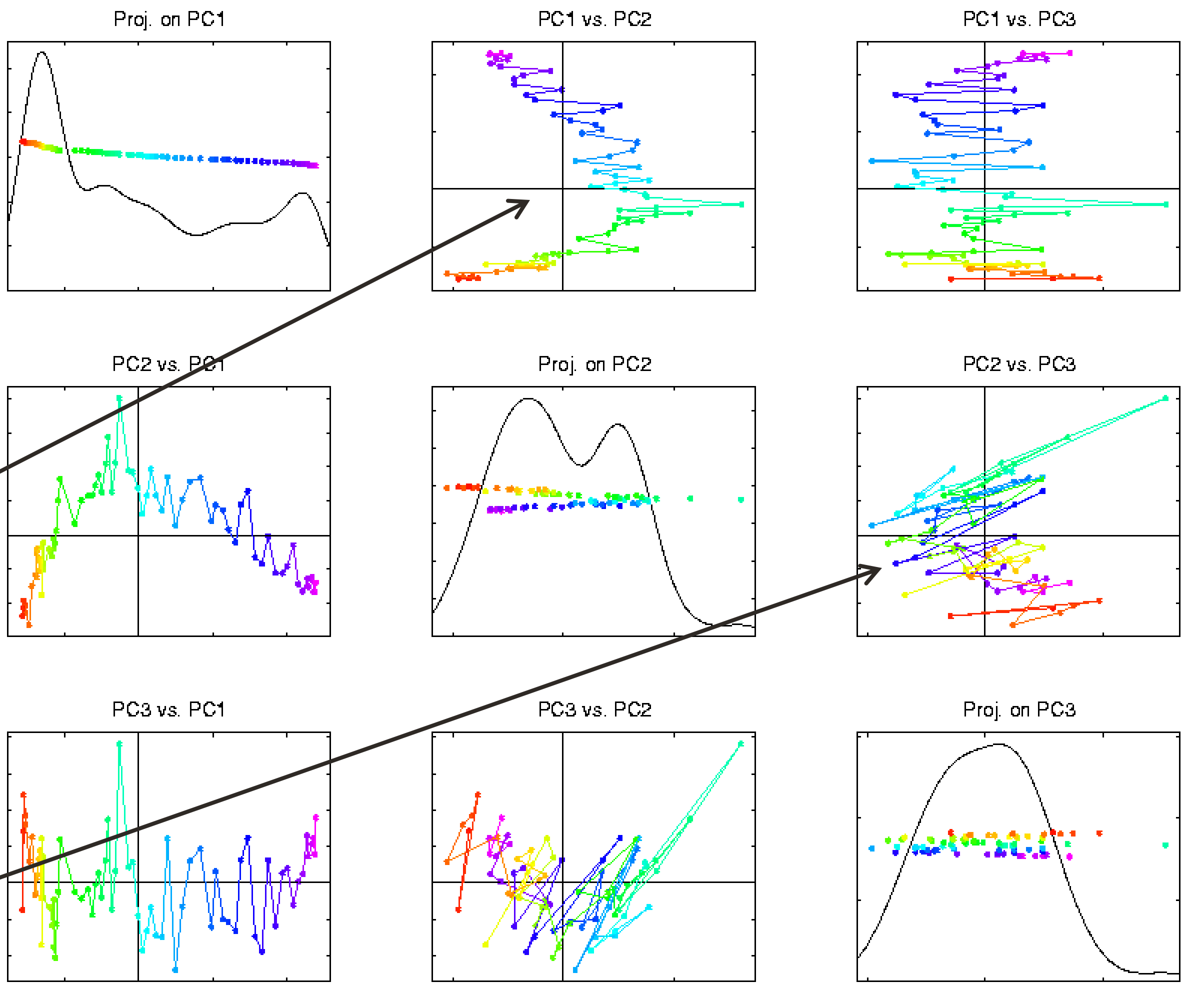


Time Series of Chemical Spectra

Study Scores (Feature Space Point Cloud)

PC2: (mostly) Systematic Variation

PC3: (mostly) Noise Driven?

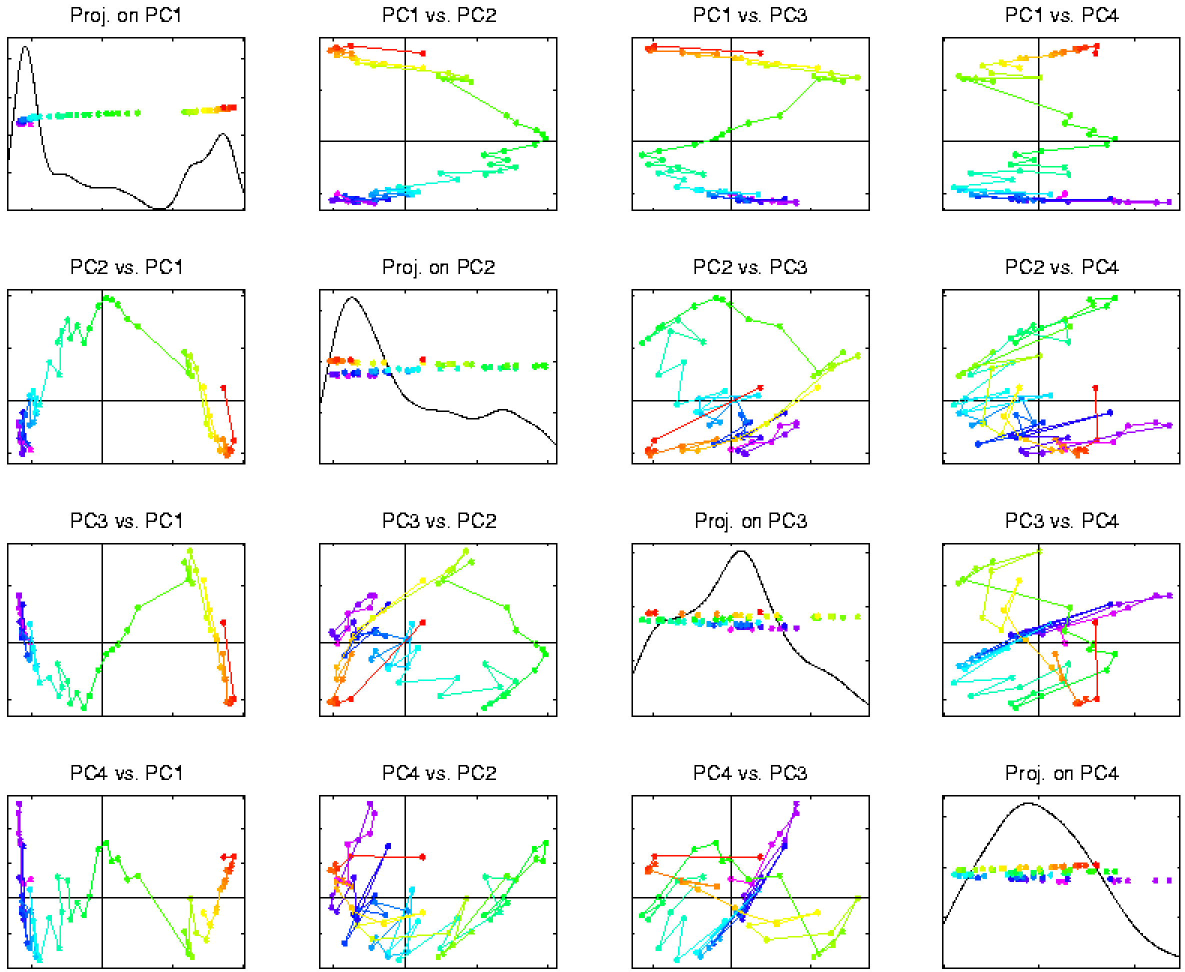


Important Trade-Off: Signal vs. Noise

Time Series of Chemical Spectra

Another Experiment

(Different Signal vs. Noise Balance)



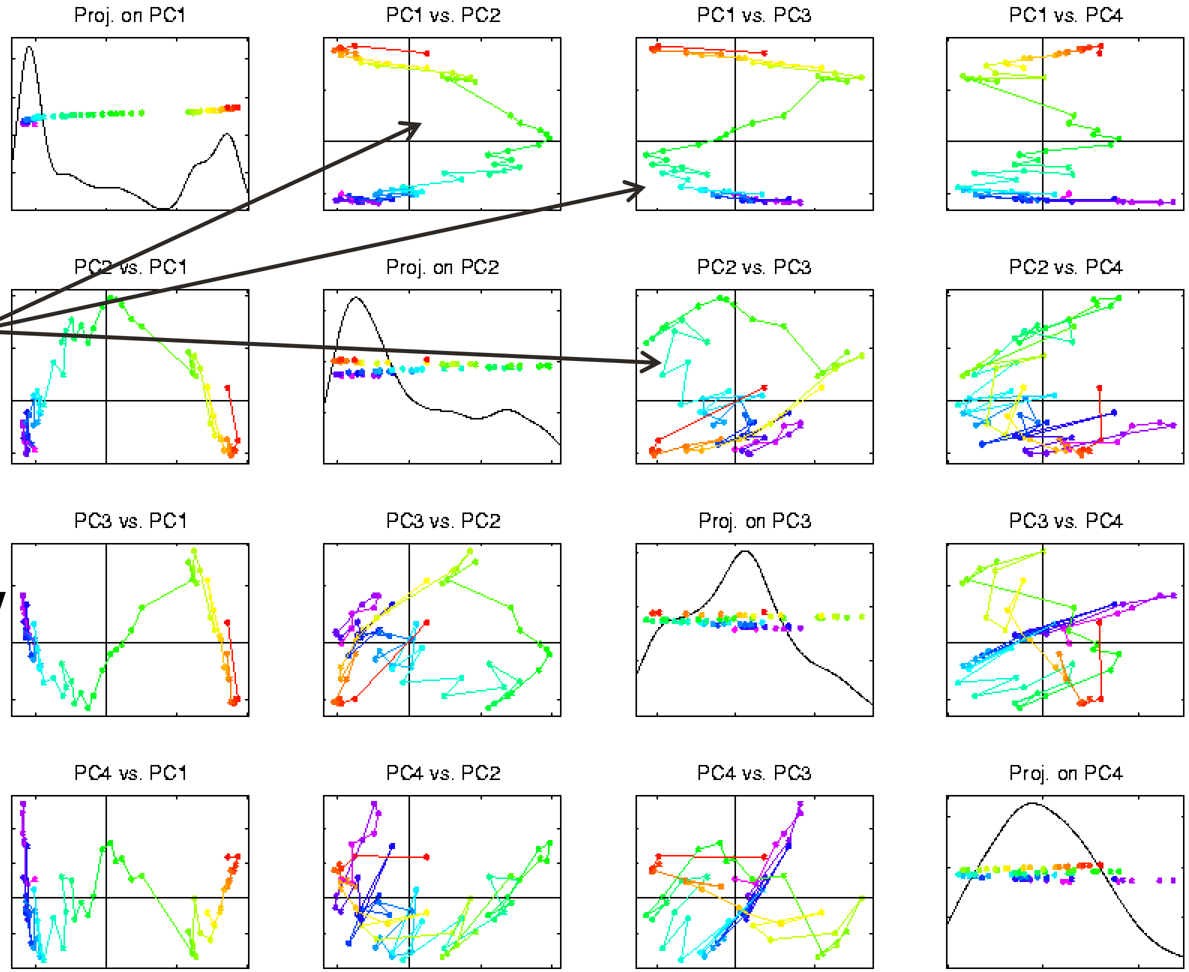
Time Series of Chemical Spectra

UNC, Stat & OR

PC1, PC2, PC3
All Suggest
Important
Patterns

(Think *Bending Curve* in
Feature Space)

Noise Is
Lower Order



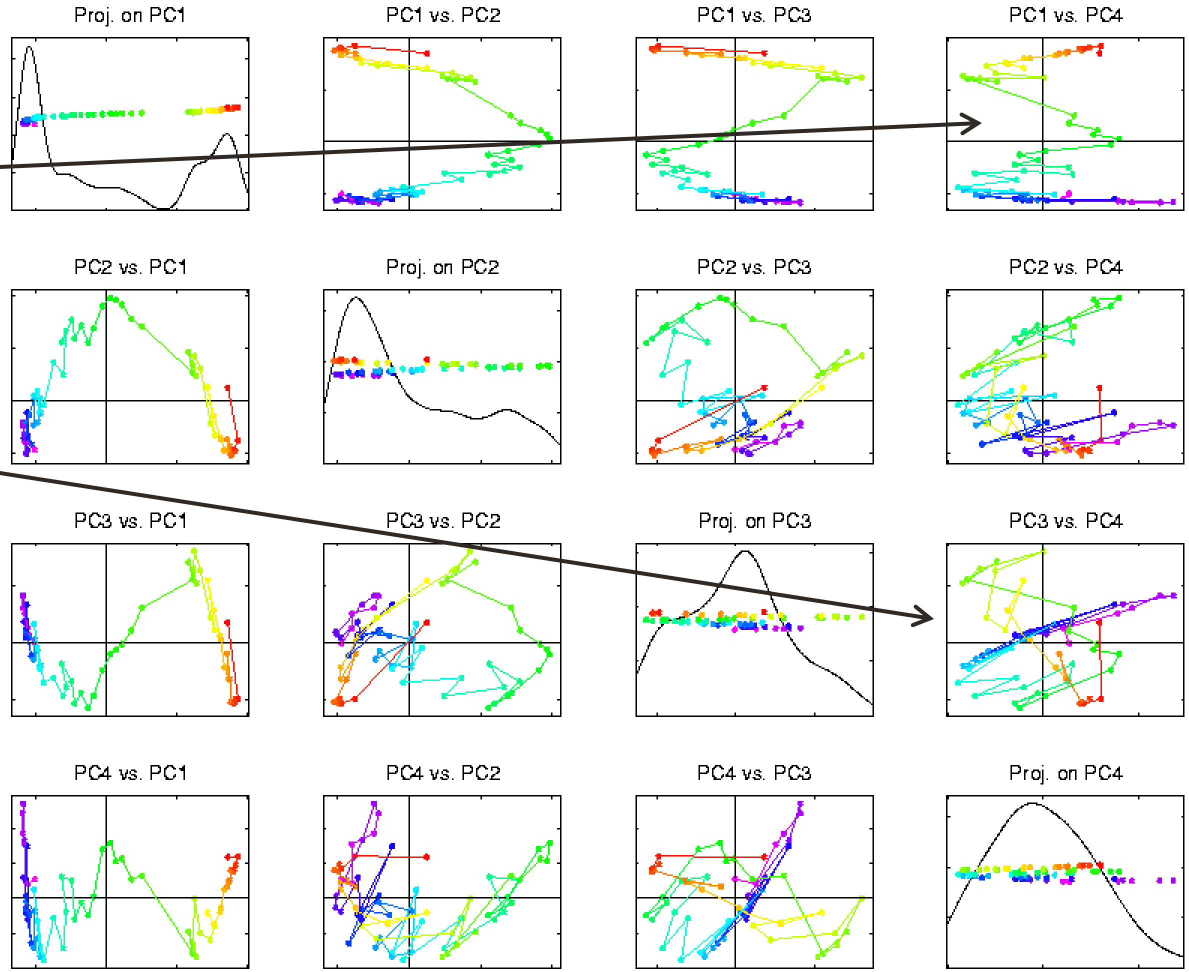
Time Series of Chemical Spectra

UNC, Stat & OR

PC 4?
Systematic?

Or Noise
Dominated?

This Pattern
Appears
Very
General



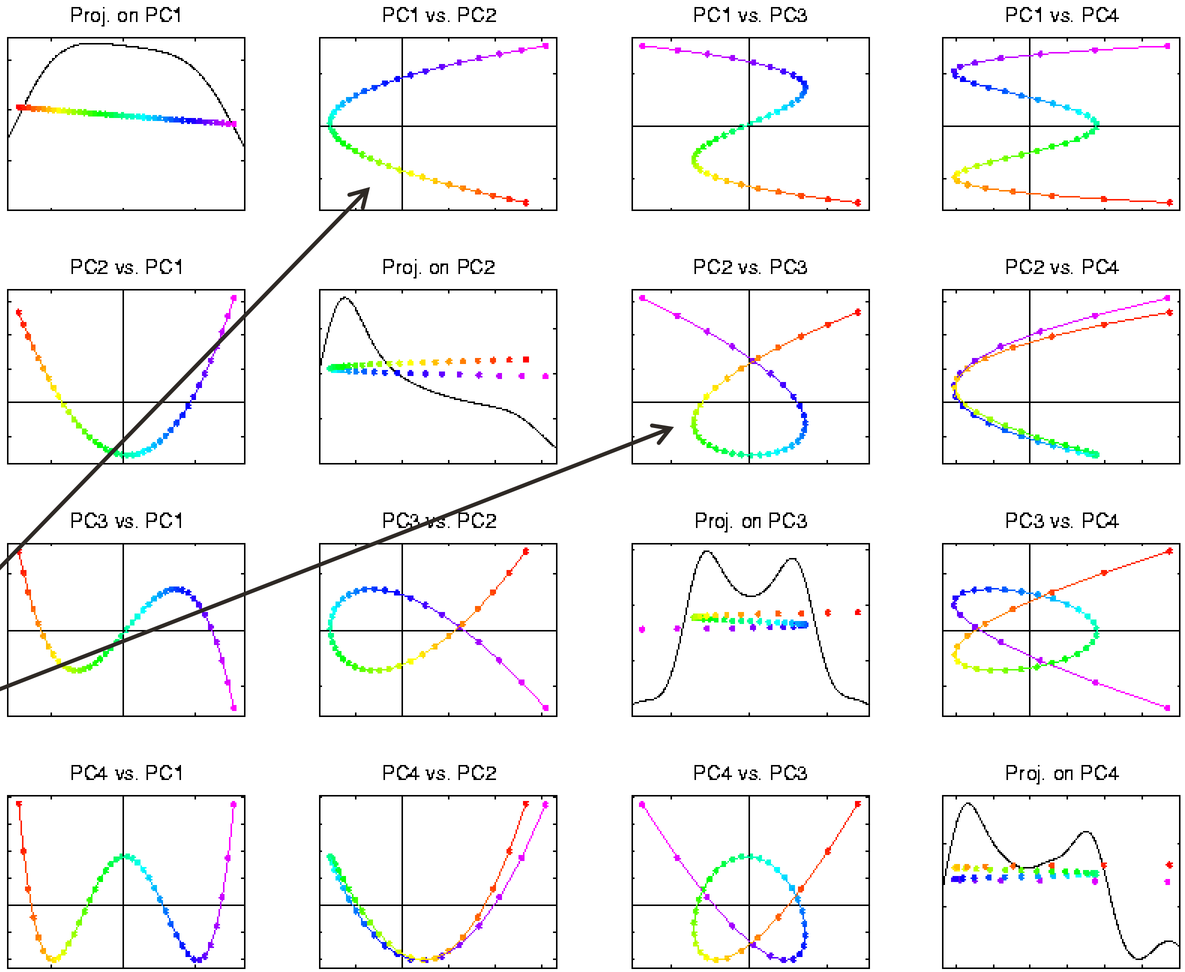
Interesting Mathematical Question: Why?

Time Series of Chemical Spectra

Simulated Chemical Experiment

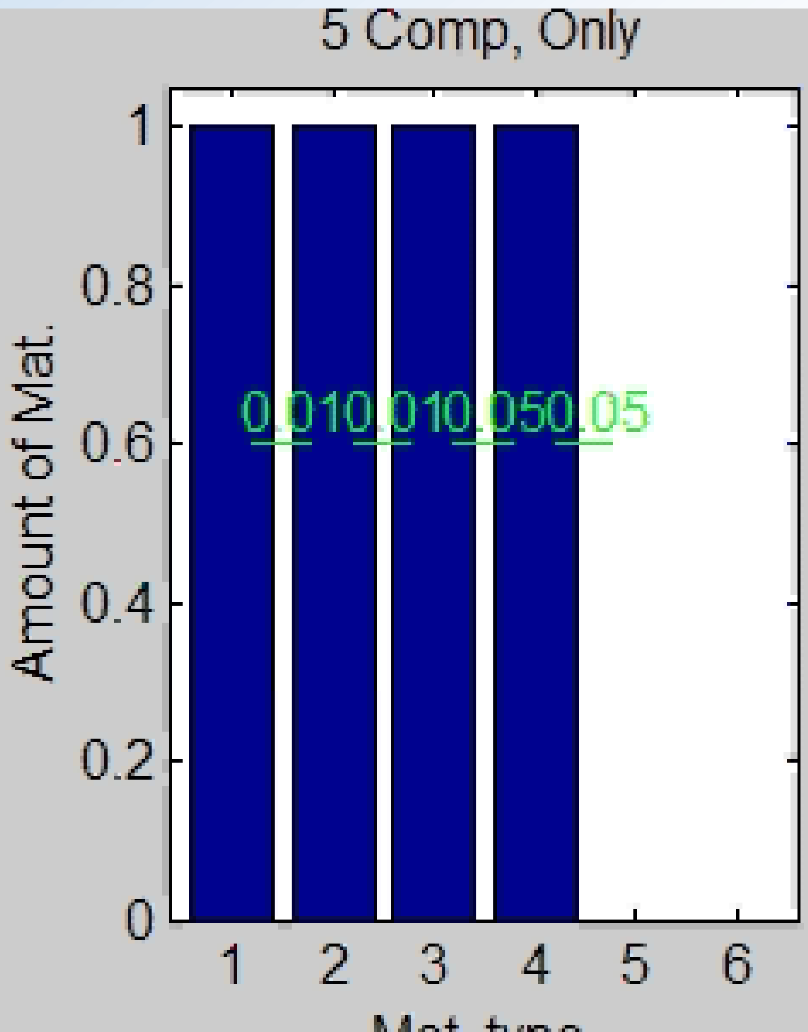
All Signal, No Noise

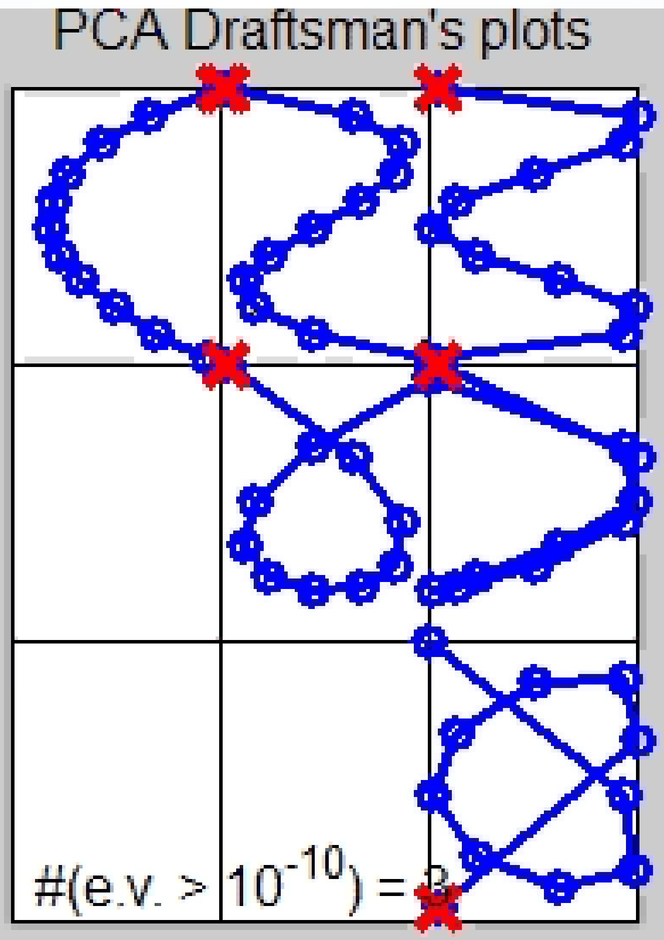
Note Observed PC Patterns



Time Series of Chemical Spectra

Simulated Chemical Experiment

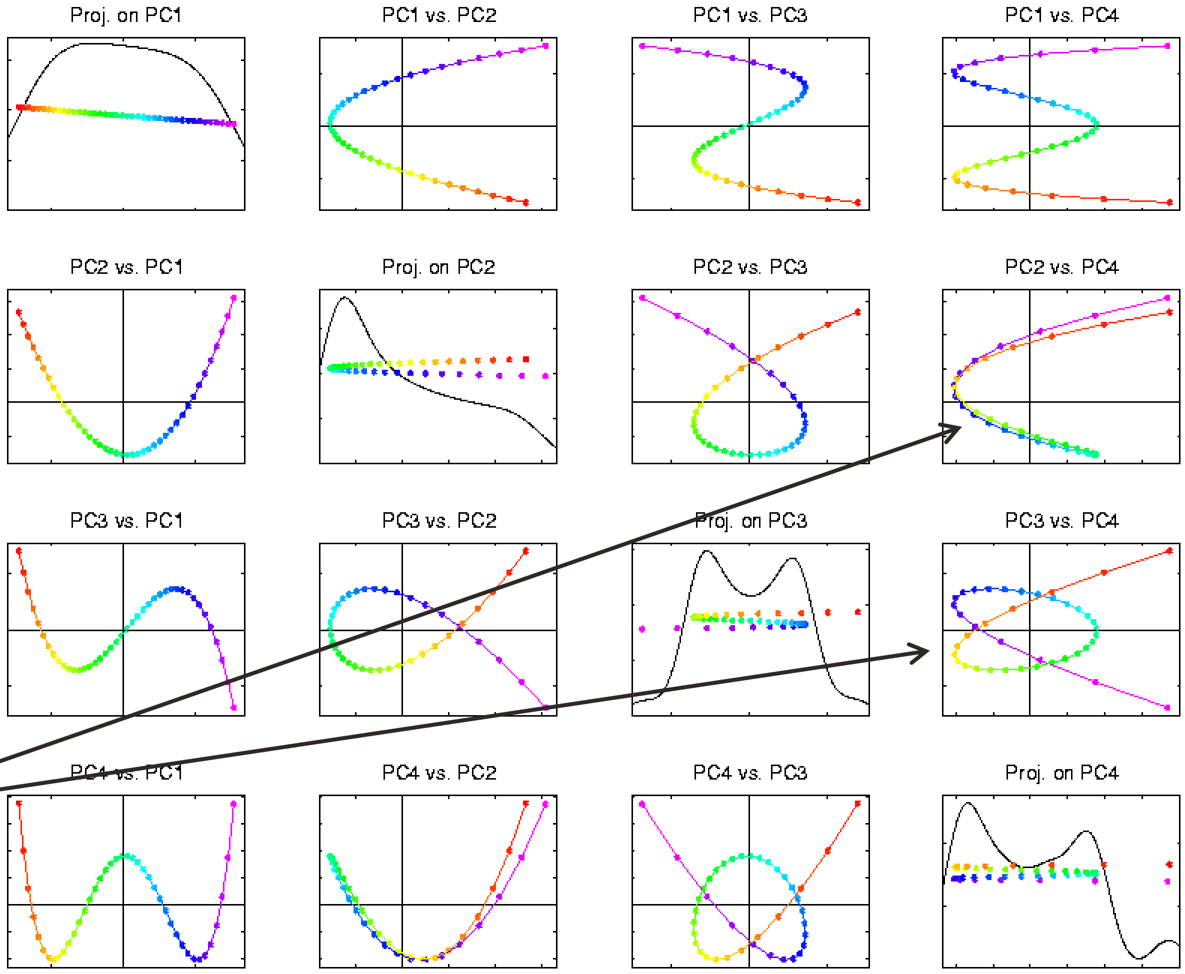




Time Series of Chemical Spectra

UNC, Stat & OR

Simulated Chemical Experiment



Higher Order PC Patterns Now Clear

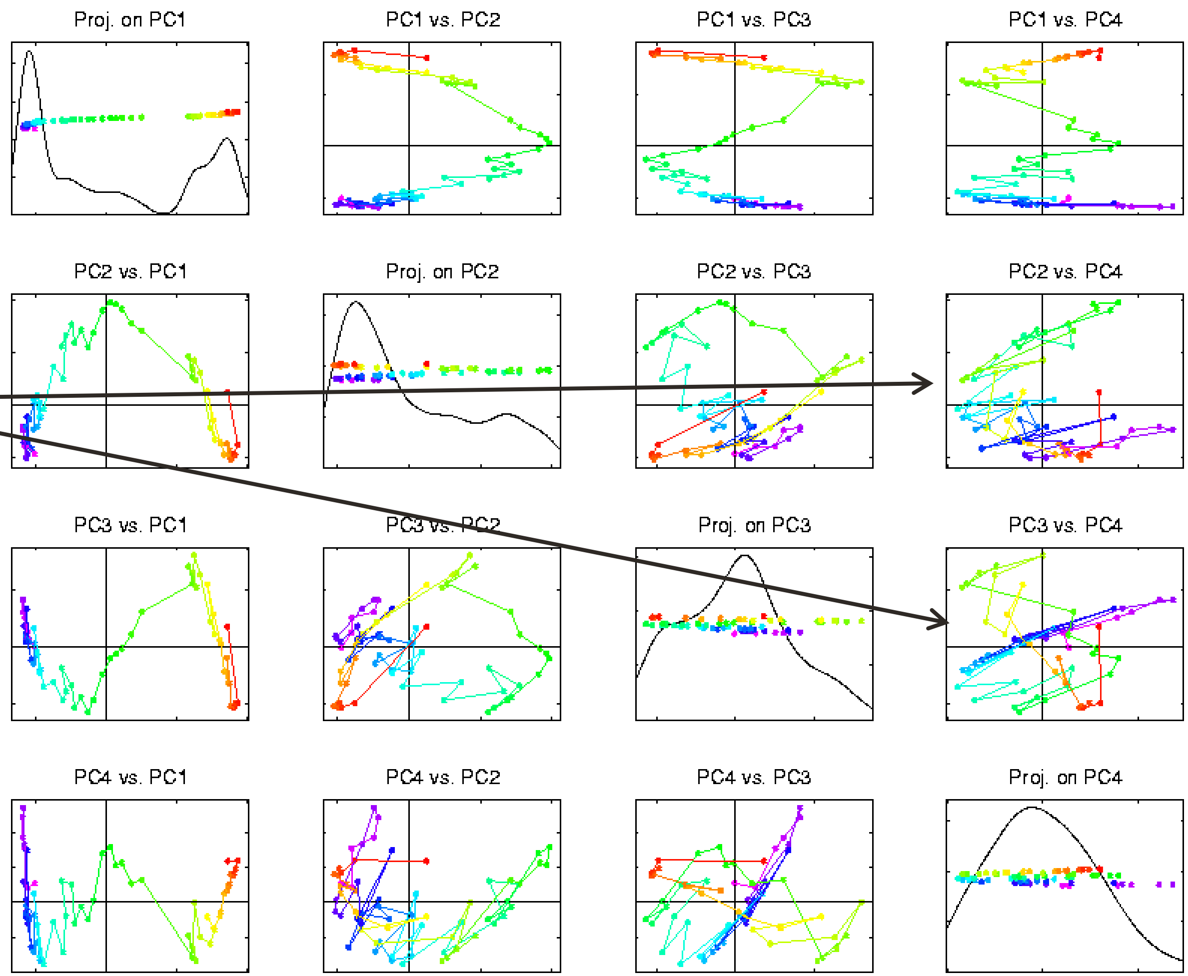
See these in real data???

Time Series of Chemical Spectra

Revisit
Real Data

Same
Patterns

Plus Noise



HDLSS Hypothesis Testing

Context: 2 – sample means

H_0: $\mu_{+1} = \mu_{-1}$ vs. H_1: $\mu_{+1} \neq \mu_{-1}$

HDLSS Hypothesis Testing

Context: 2 – sample means

H_0: $\mu_{+1} = \mu_{-1}$ vs. H_1: $\mu_{+1} \neq \mu_{-1}$

Challenges:

- Distributional Assumptions
- Parameter Estimation

HDLSS Hypothesis Testing

Context: 2 – sample means

$H_0: \mu_{+1} = \mu_{-1}$ vs. $H_1: \mu_{+1} \neq \mu_{-1}$

Challenges:

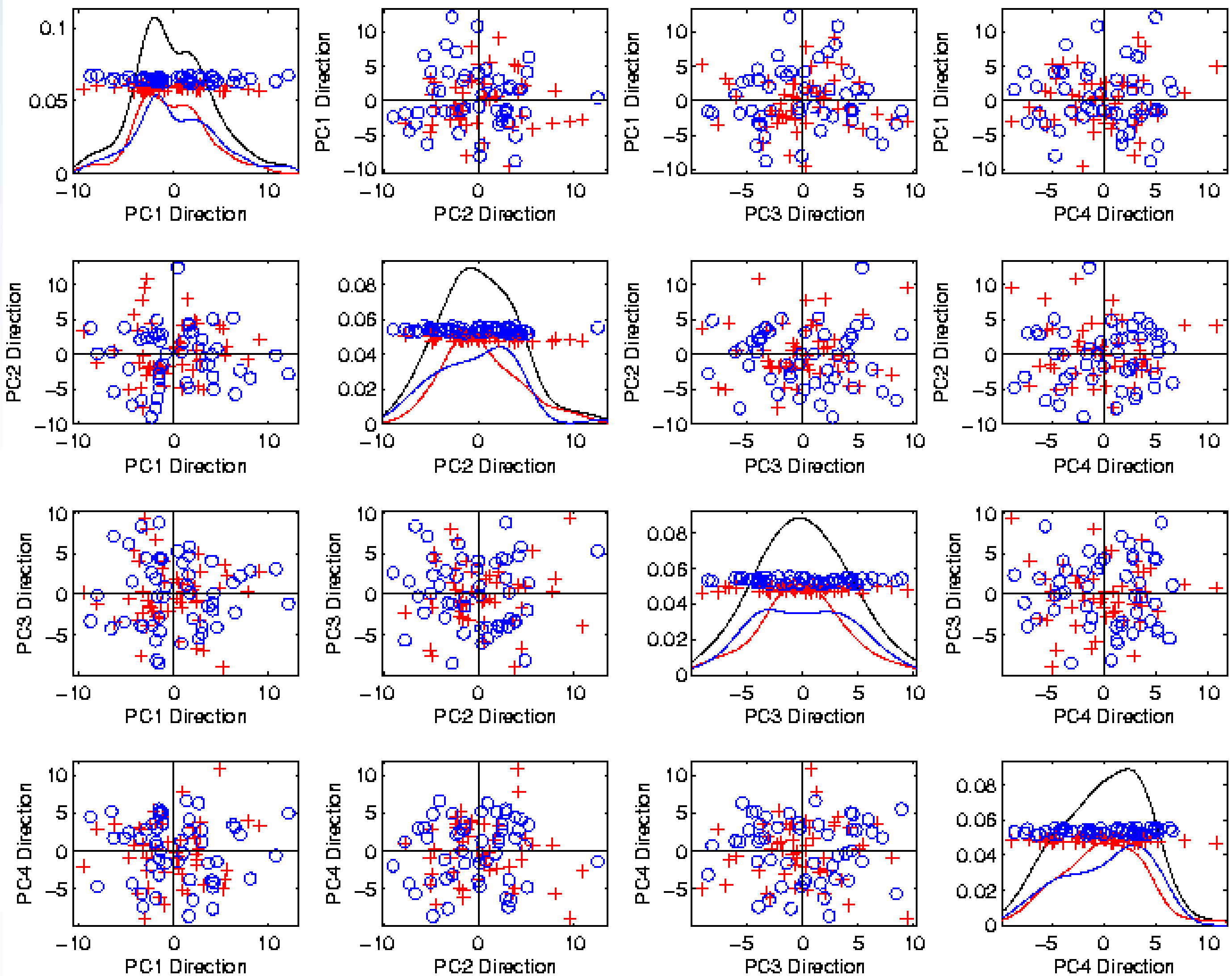
- Distributional Assumptions
- Parameter Estimation

- HDLSS space is slippery

HDLSS Hypothesis Testing

Toy 2-Class Example

See Structure?

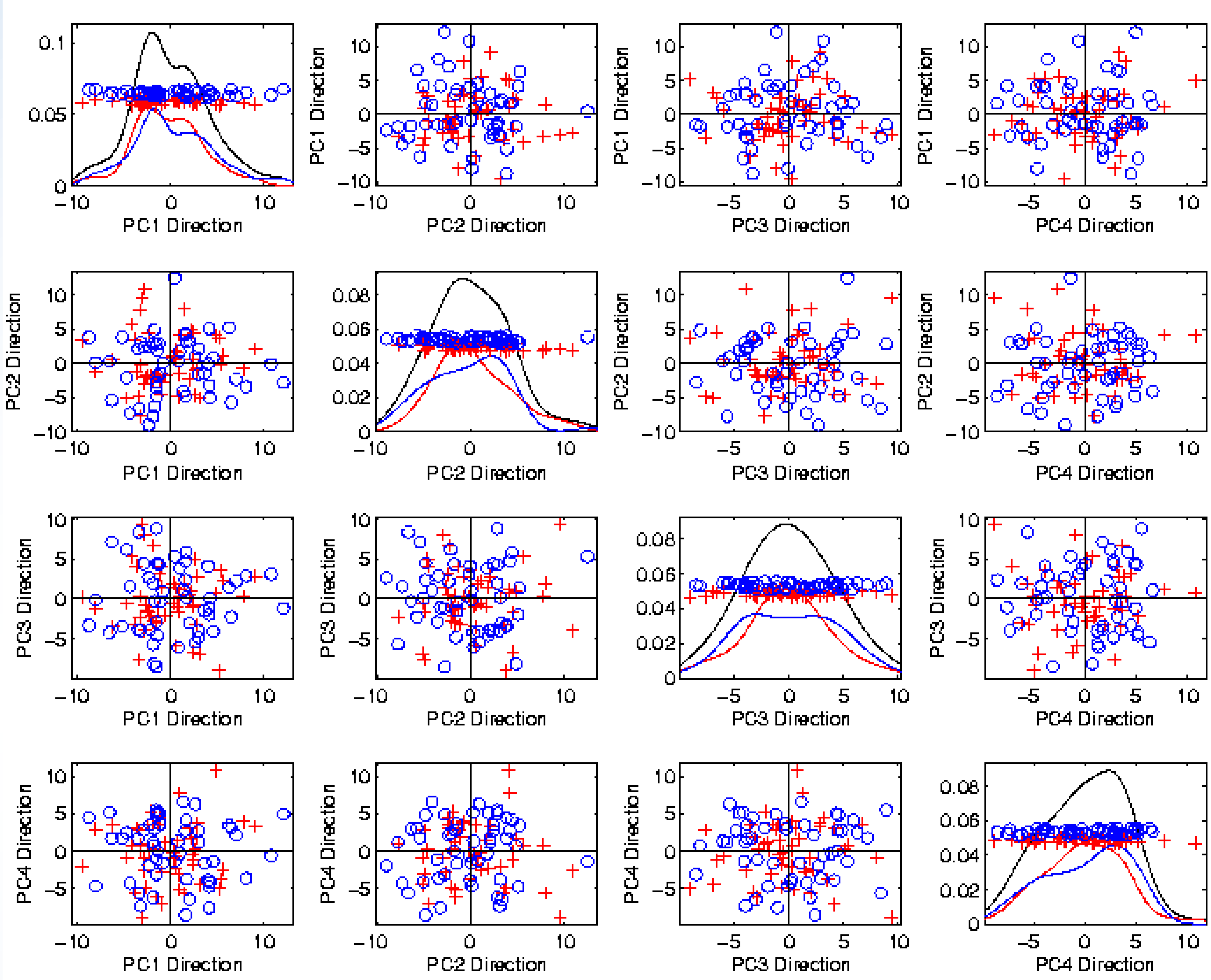


HDLSS Hypothesis Testing

Toy 2-Class Example

See Structure?

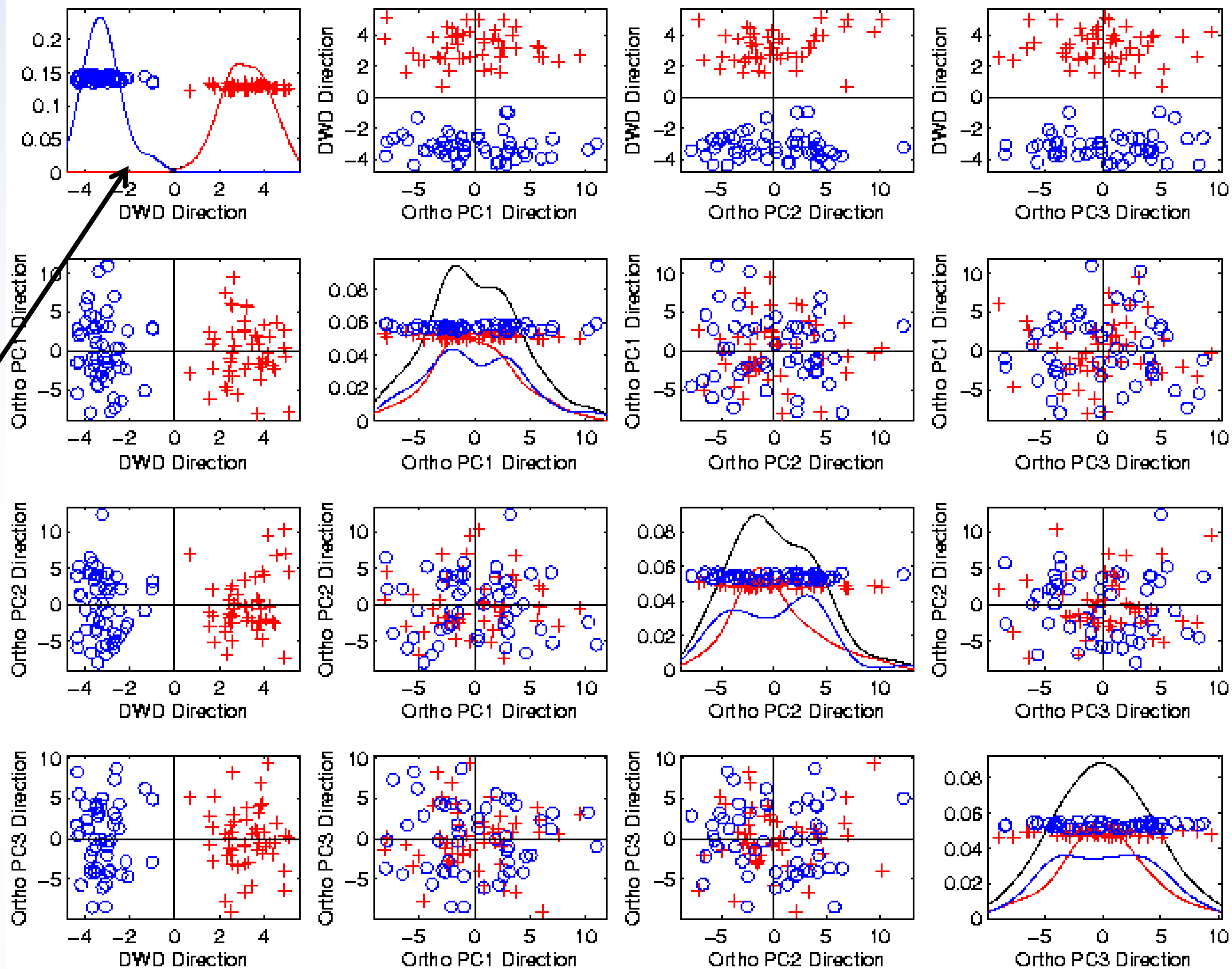
Careful, Only PC1-4



HDLSS Hypothesis Testing

Toy 2-Class Example

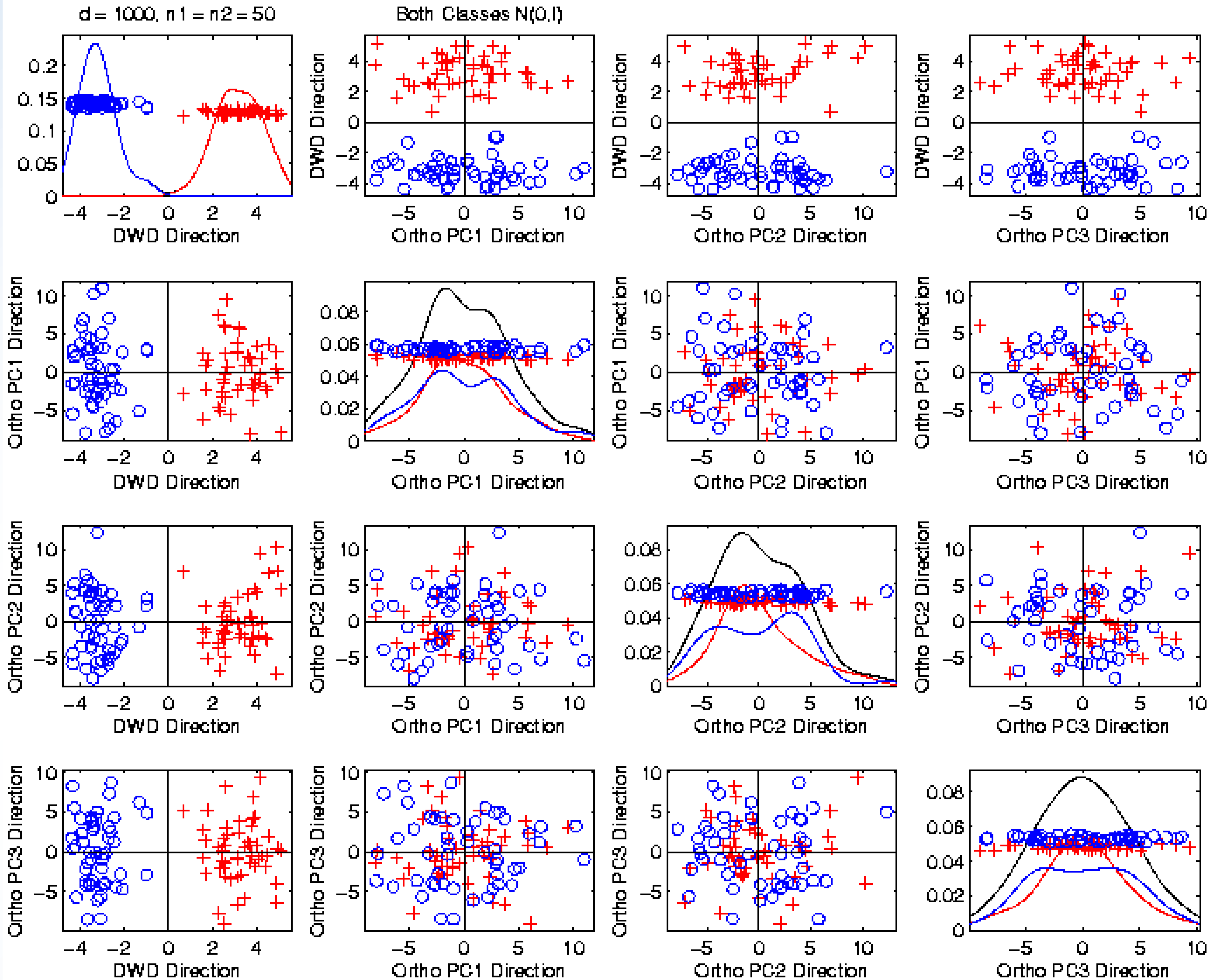
Structure Looks Real???



HDLSS Hypothesis Testing

Toy 2-Class Example

Actually Both Classes Are N(0,I), d = 1000

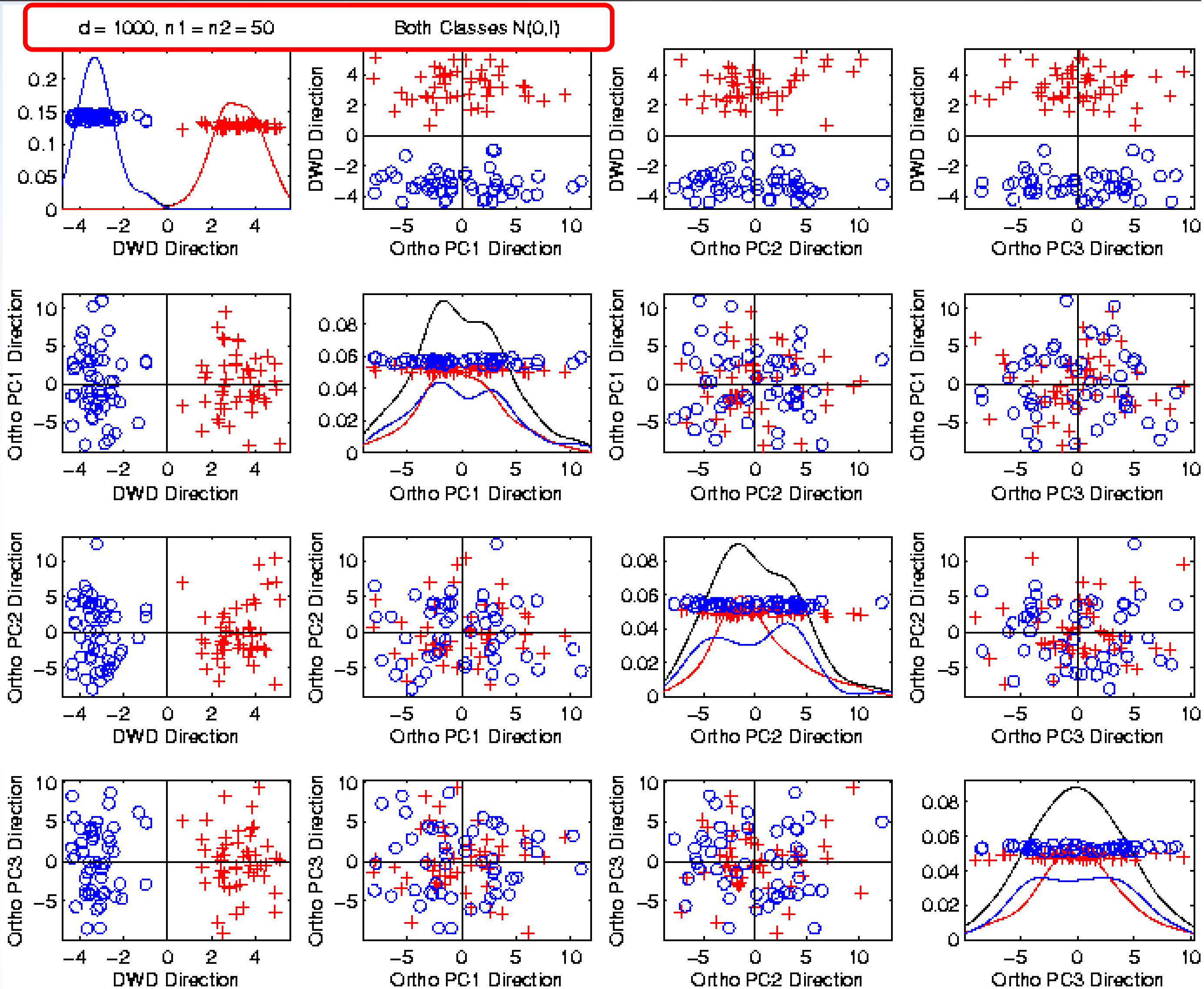


HDLSS Hypothesis Testing

UNC, Stat & OR

Toy 2-Class Example

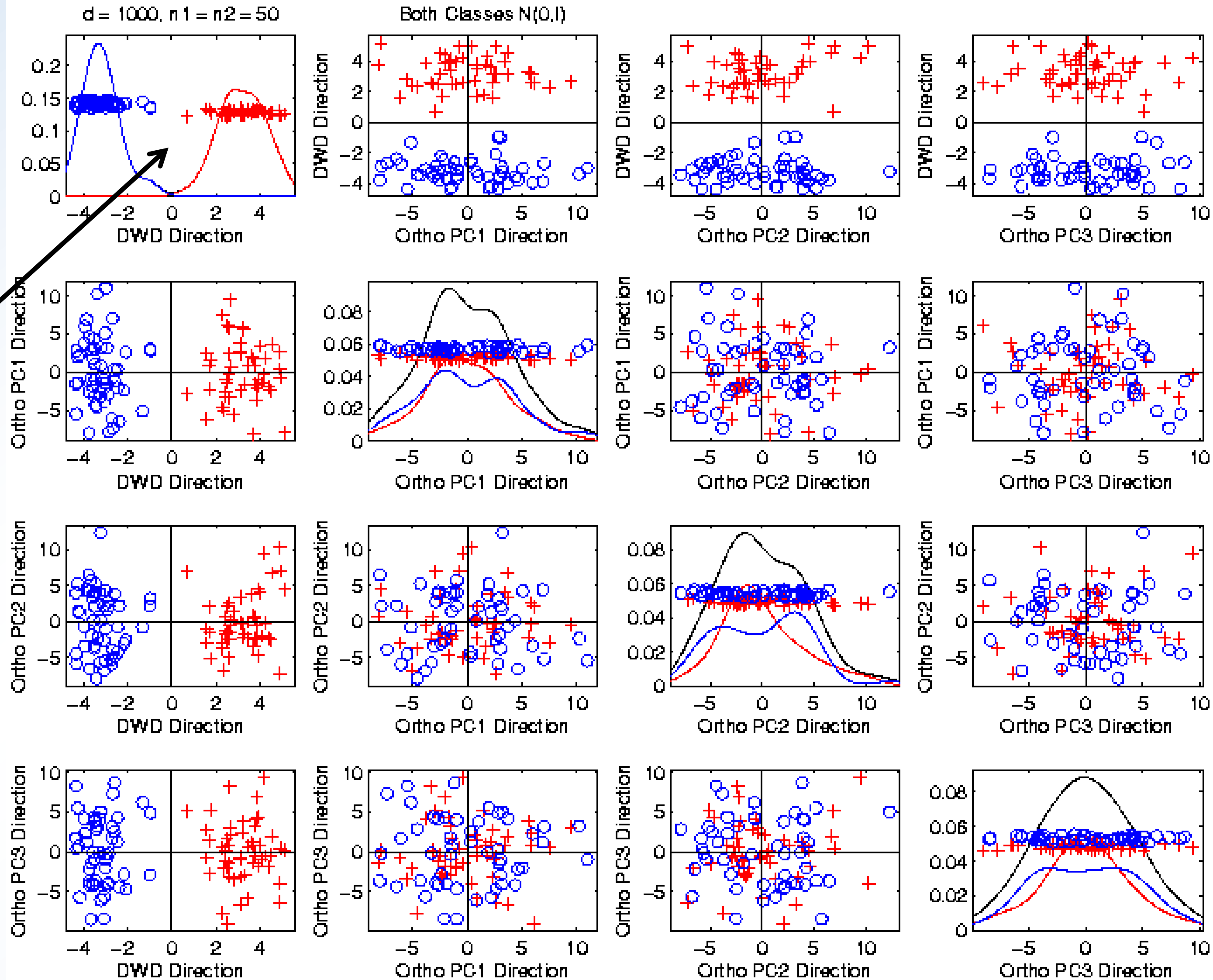
Actually <u>Both</u> <u>Classes</u> Are N(0,I), d = 1000



HDLSS Hypothesis Testing

Toy 2-Class Example

Separation Is *Natural Sampling Variation*



HDLSS Hypothesis Testing

Context: 2 – sample means

H_0: $\mu_{+1} = \mu_{-1}$ vs. H_1: $\mu_{+1} \neq \mu_{-1}$

Challenges:

- Distributional Assumptions
- Parameter Estimation

- HDLSS space is slippery

HDLSS Hypothesis Testing

Context: 2 – sample means

H_0: $\mu_{+1} = \mu_{-1}$ vs. H_1: $\mu_{+1} \neq \mu_{-1}$

Challenges:

- Distributional Assumptions
- Parameter Estimation

Suggested Approach:

Permutation test

HDLSS Hypothesis Testing - DiProPerm

Suggested Approach:

✓ Find a DIrection

(separating classes)

HDLSS Hypothesis Testing - DiProPerm

Suggested Approach:

- ✓ Find a DIrection

  (separating classes)

- ✓ PROject the data

  (reduces to 1 dim)

HDLSS Hypothesis Testing - DiProPerm

Suggested Approach:

- ✓ Find a DIrection

  (separating classes)

- ✓ PROject the data

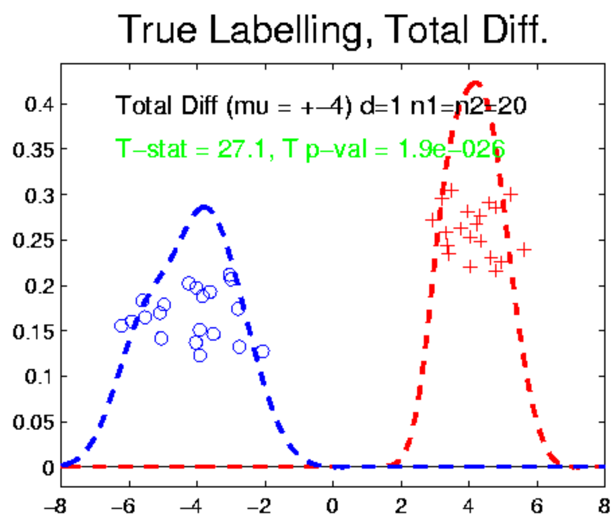
  (reduces to 1 dim)

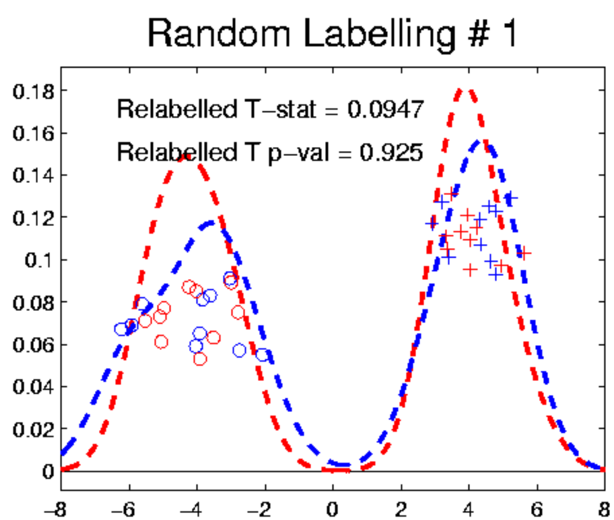
- ✓ PERMute

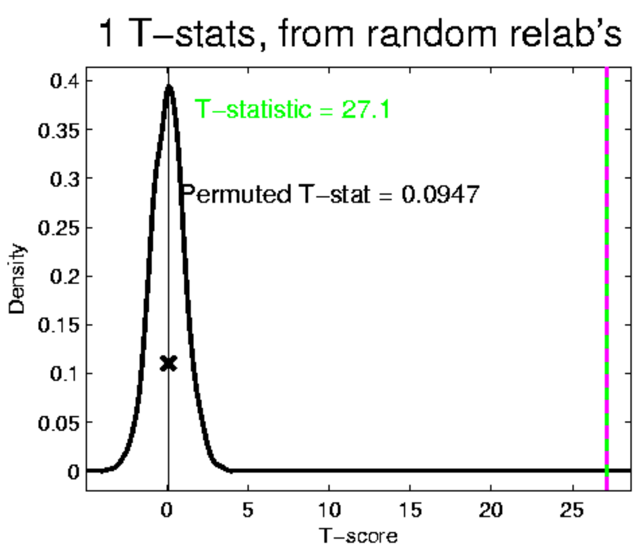
  (class labels, to assess significance,

  with recomputed direction)

HDLSS Hypothesis Testing - DiProPerm

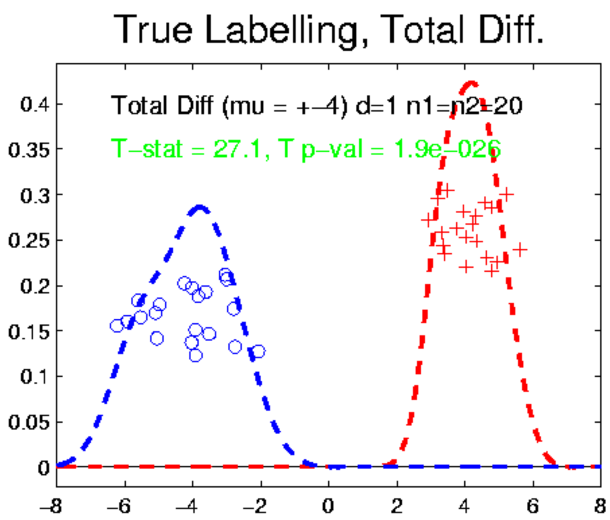


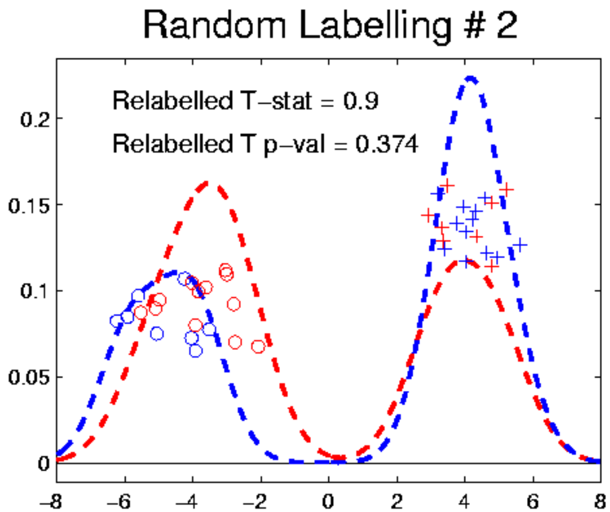


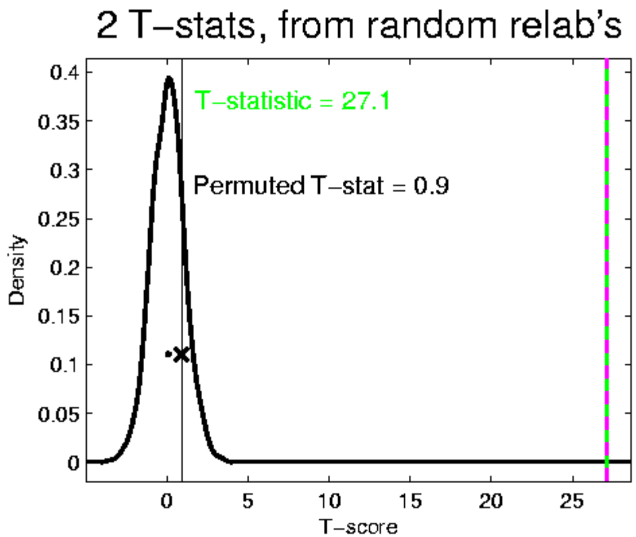


HDLSS Hypothesis Testing - DiProPerm

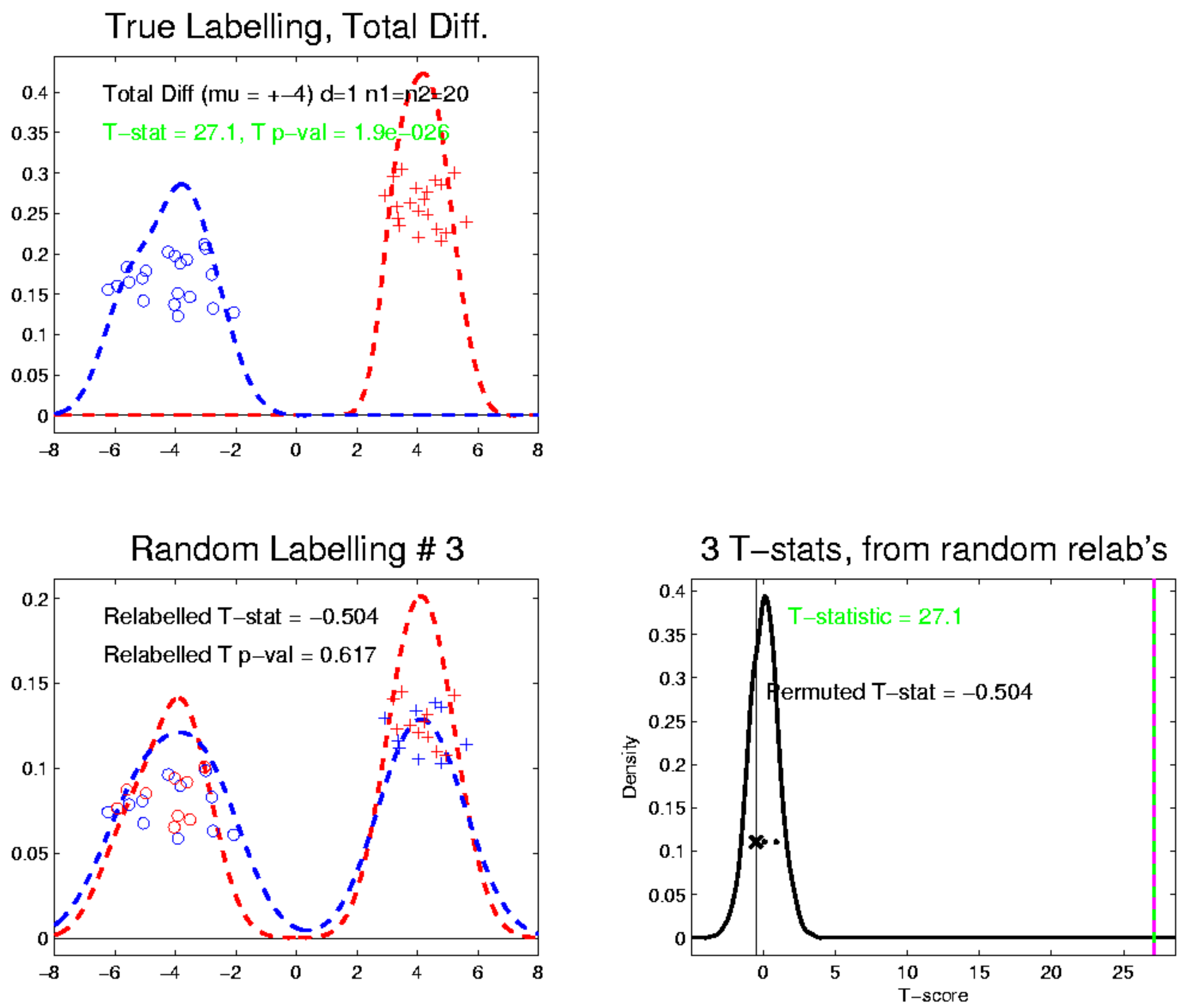
UNC, Stat & OR





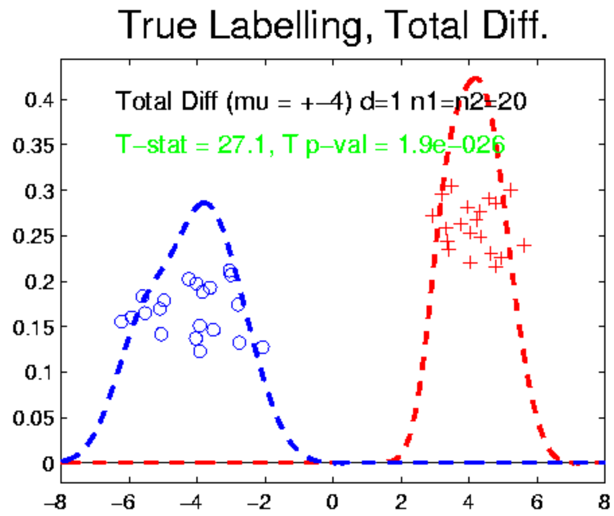


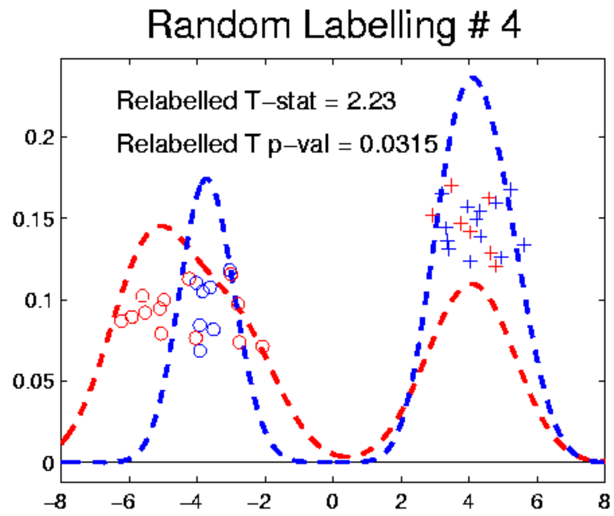
HDLSS Hypothesis Testing - DiProPerm

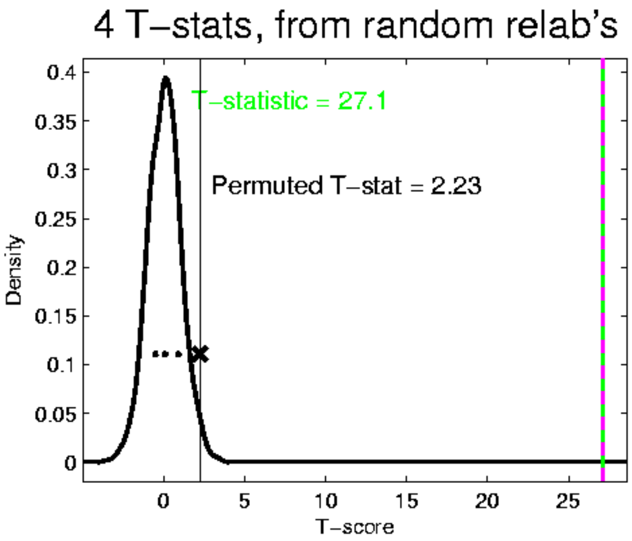


HDLSS Hypothesis Testing - DiProPerm

UNC, Stat & OR







HDLSS Hypothesis Testing - DiProPerm

UNC, Stat & OR

.

.

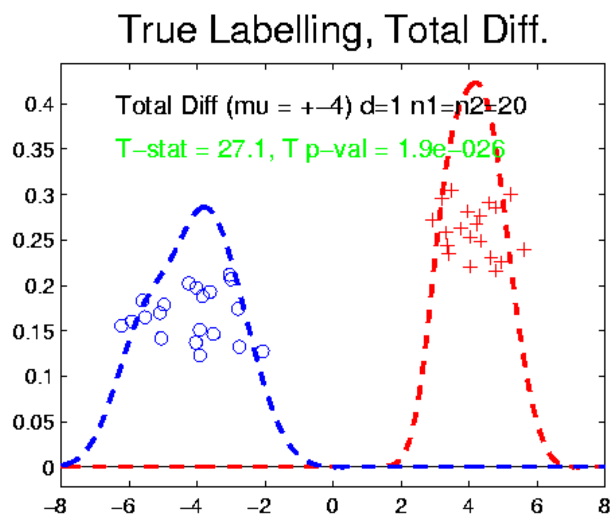
.

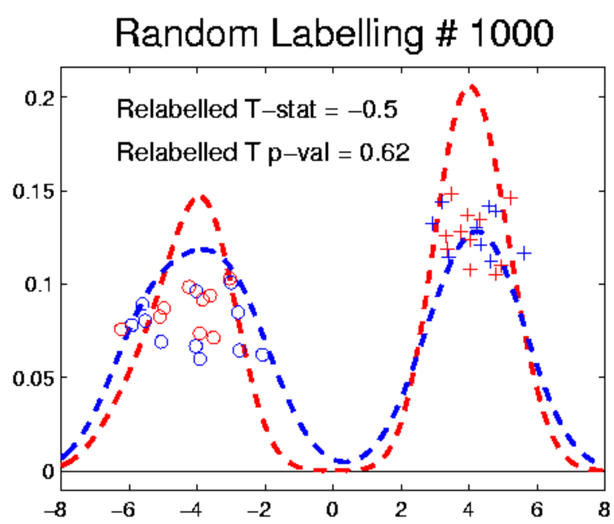
Repeat this 1,000 times

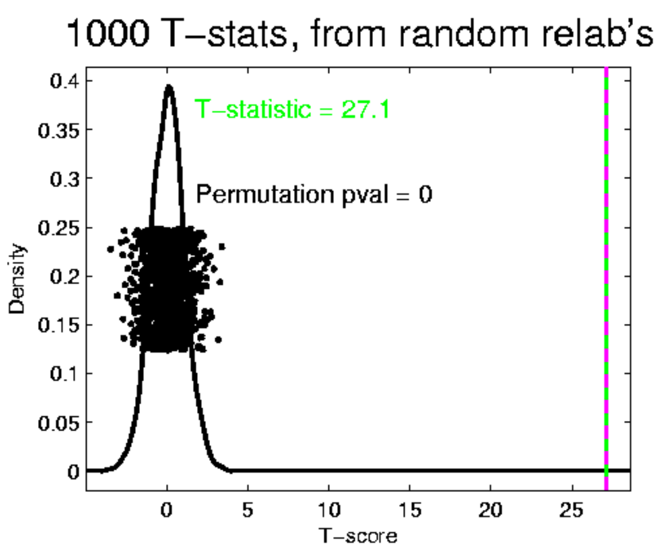
To get:

HDLSS Hypothesis Testing - DiProPerm

UNC, Stat & OR







UNC, Stat & OR

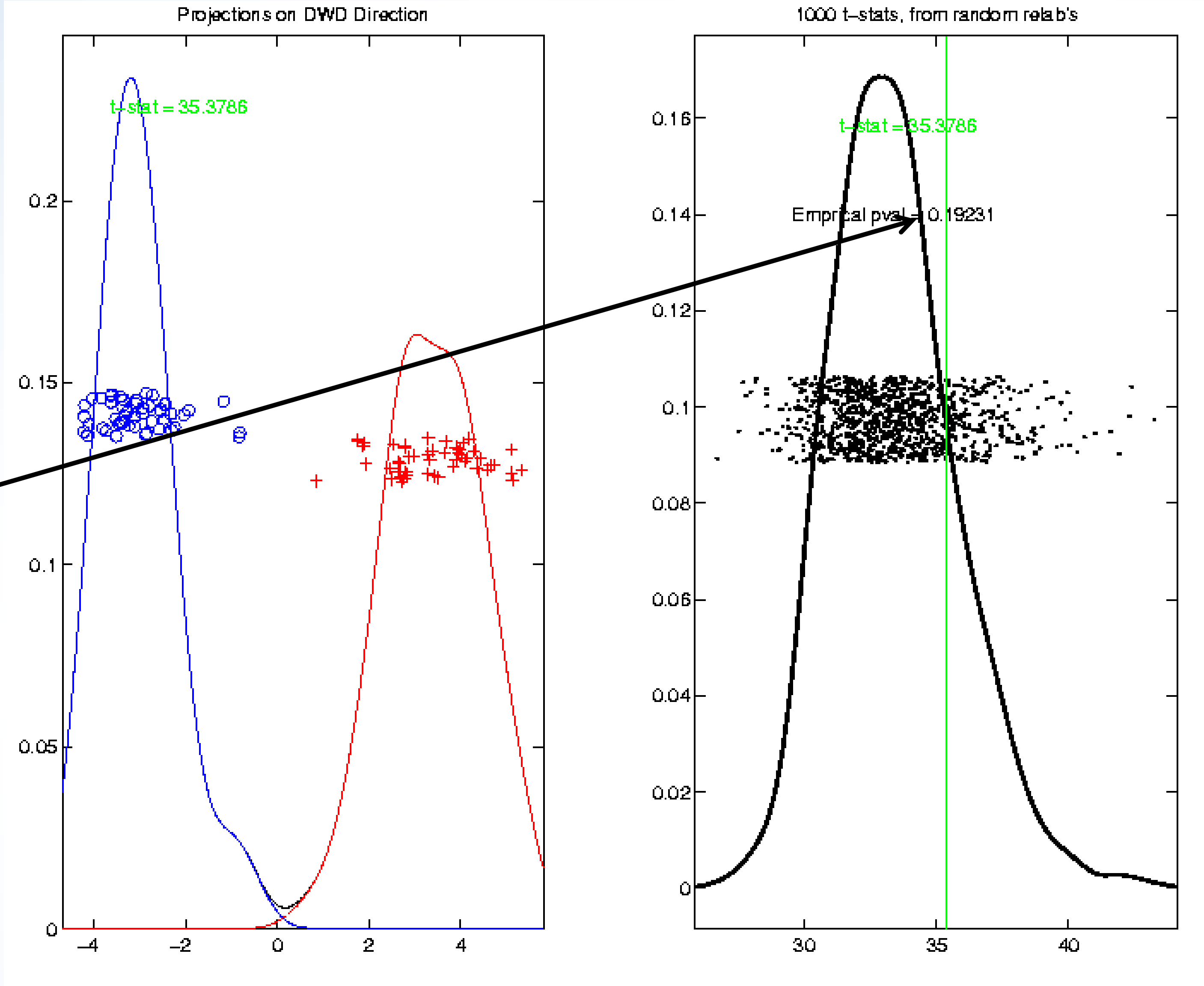
HDLSS Hypothesis Testing

Toy 2-Class Example

p-value

<u>Not</u>

Significant



HDLSS Hypothesis Testing - DiProPerm

Real Data Example: Autism

Caudate Shape

(sub-cortical brain structure)

Shape summarized by 3-d locations of 1032 corresponding points

Autistic vs. Typically Developing

Autism Data - DiProPerm

UNC, Stat & OR

Finds

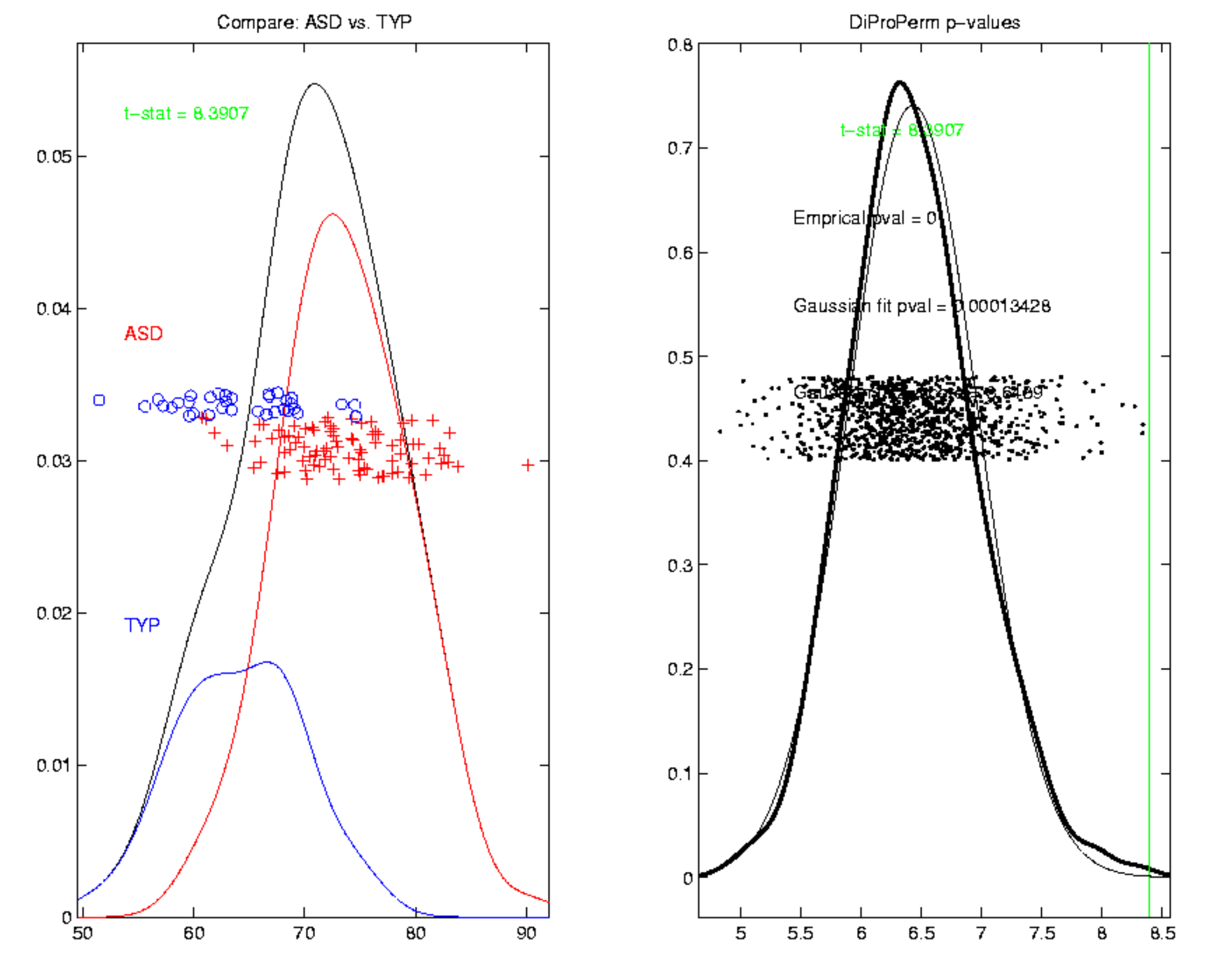
Significant

Difference

Despite Weak

Visual

Impression



Thanks to Josh Cates

Autism Data - DiProPerm

UNC, Stat & OR

Also Compare: Developmentally Delayed

<u>No</u>

Significant

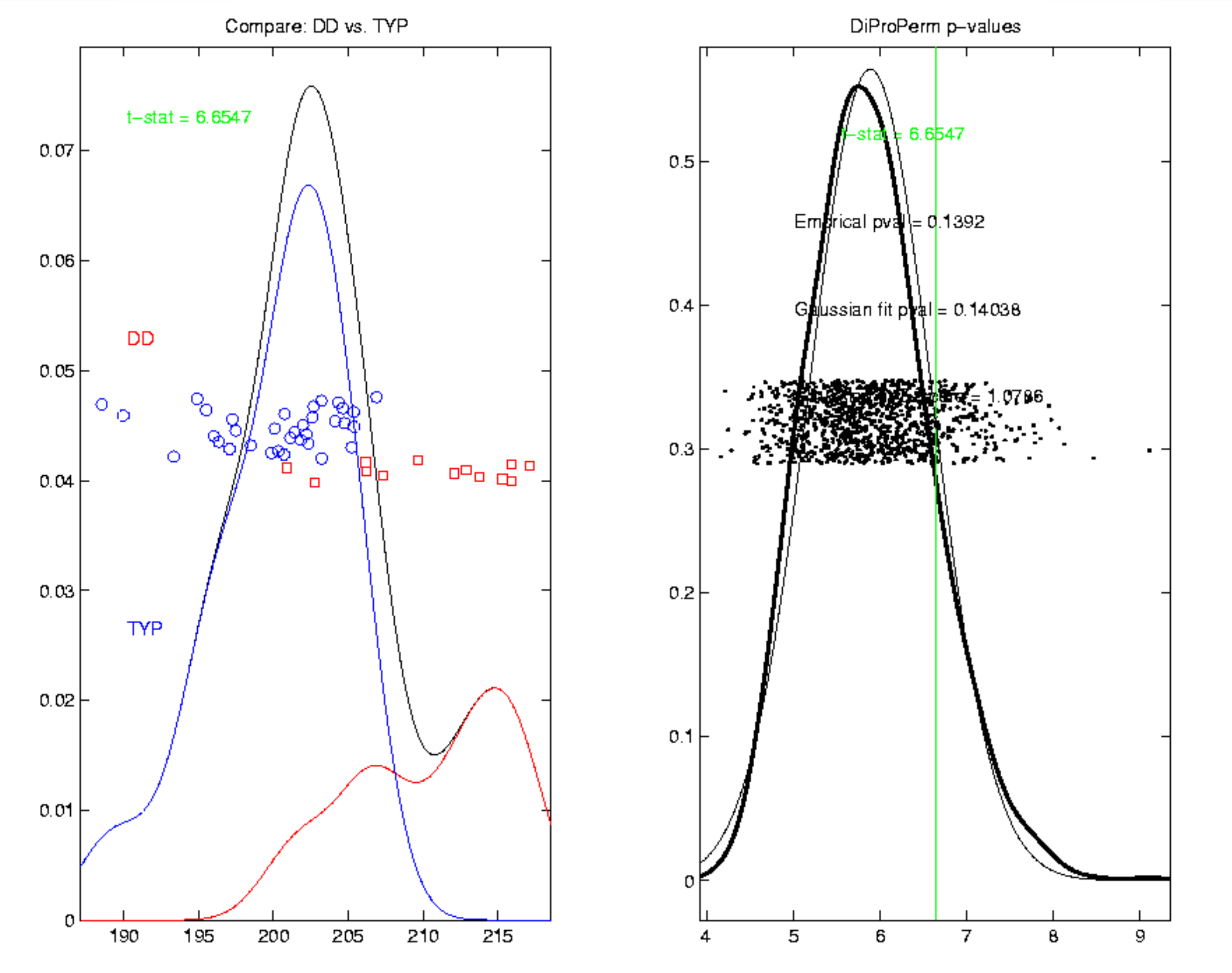
Difference

But Strong

Visual

Impression

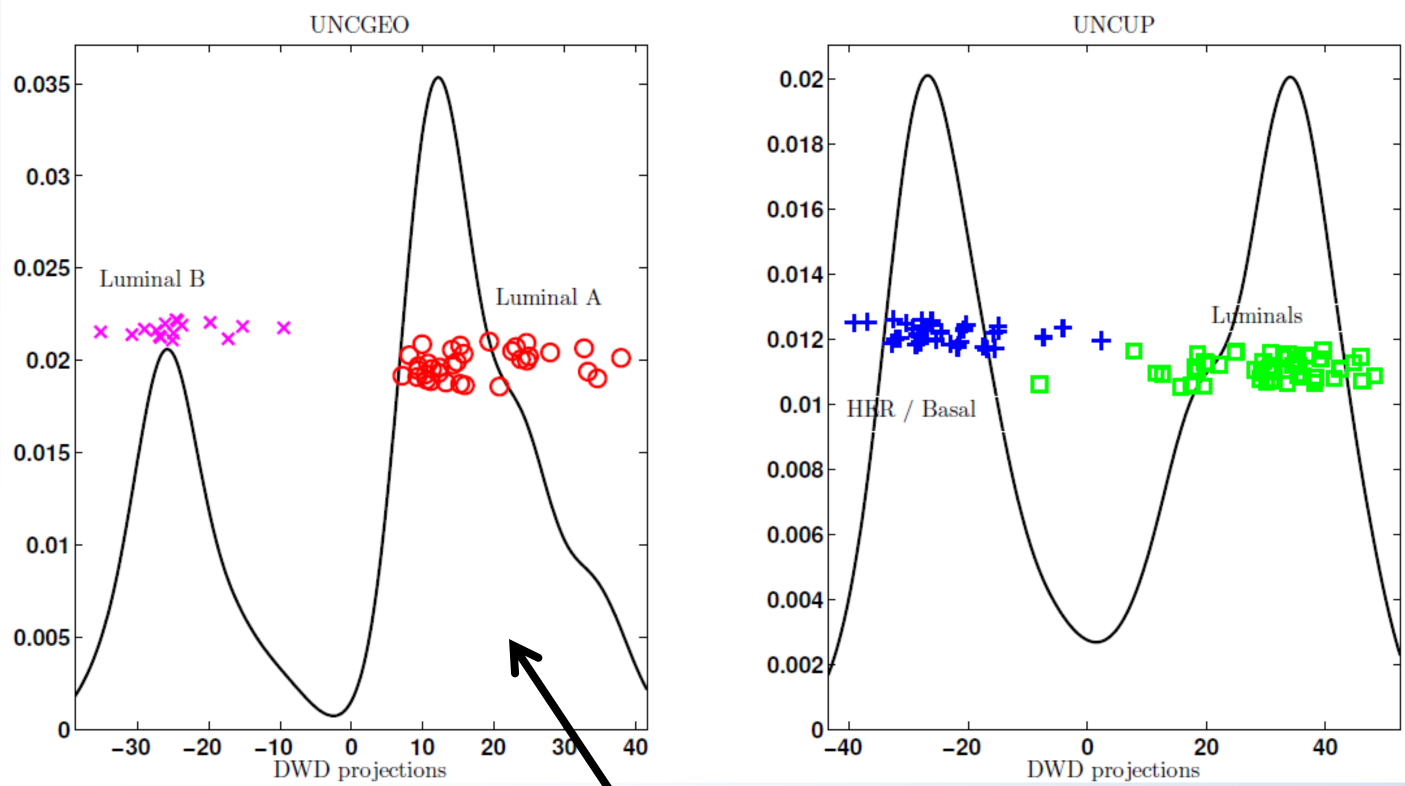
Thanks to Josh Cates



Breast Cancer Microarray Data - DiProPerm

Two

Examples



Which Is

"More

Distinct"?

Visually *Better Separation*?

Thanks to Katie Hoadley

Breast Cancer Microarray Data - DiProPerm

UNC, Stat & OR

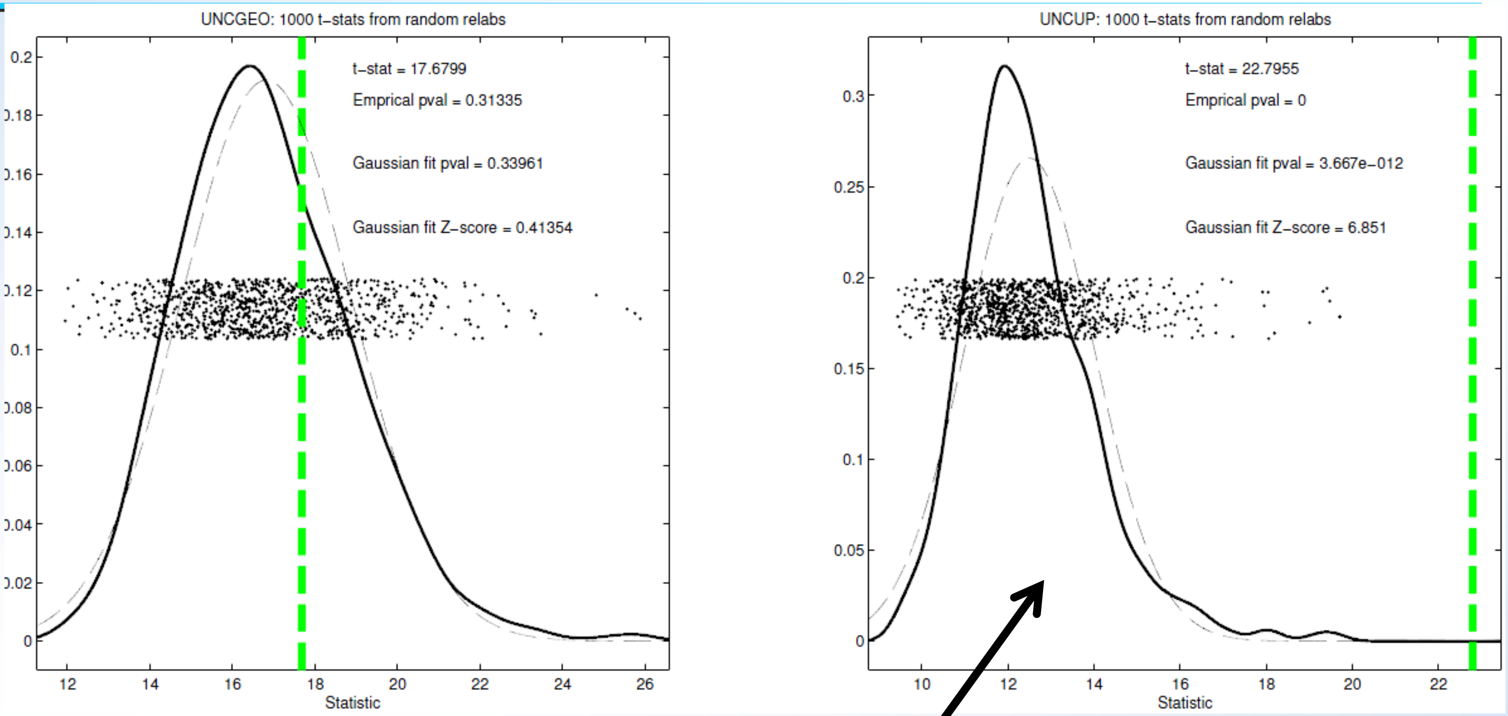
Two

Examples

Which Is

"More

Distinct"?



Stronger *Statistical Significance*

Thanks to Katie Hoadley

HDLSS Hypothesis Testing - DiProPerm

Value of DiProPerm:

- Visual Impression is Easily Misleading
  (onto HDLSS projections,
  e.g. Maximal Data Piling)
- Really Need to Assess Significance
- DiProPerm used routinely
  (even for variable selection)

HDLSS Hypothesis Testing - DiProPerm

Choice of Direction:

- Distance Weighted Discrimination (DWD)
- Support Vector Machine (SVM)
- Mean Difference
- Maximal Data Piling

$\vdots$

HDLSS Hypothesis Testing - DiProPerm

Choice of 1-d Summary Statistic:

- 2-sample t-stat
- Mean difference
- Median difference
- Area Under ROC Curve

$\vdots$

Carry Away Concept

UNC, Stat & OR

OODA is more than a “framework”

It Provides a Focal Point

Highlights Pivotal Choices:

What should be the Data Objects?

How should they be Represented?