STAT3622 Data Visualization

PCA

Data Object Conceptualization

Object Space	←→	Descriptor Space
Curves		$\Re^d$
Images		Manifolds
Shapes		Tree Space
Trees		

Curves As Data

Object Space: Set of curves

Descriptor Space(s):

- Curves digitized to vectors
- Basis Representations:
  - Fourier (sin & cos)
  - B-splines
  - Wavelets

Curves As Data, I

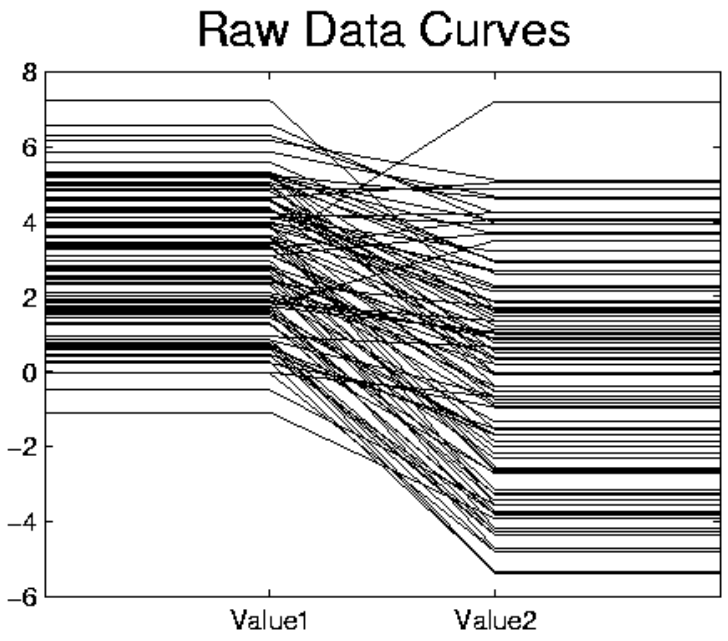
Very simple example

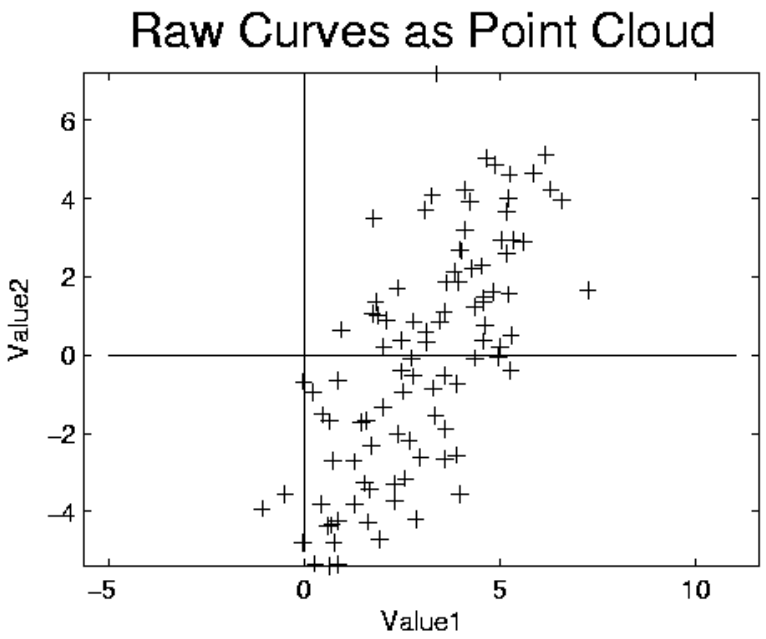
- “2 dimensional” family of (digitized) curves
- Object space: piece-wise linear f’ns
- Descriptor space $= \mathbb{R}^2$

PCA: reveals “population structure”

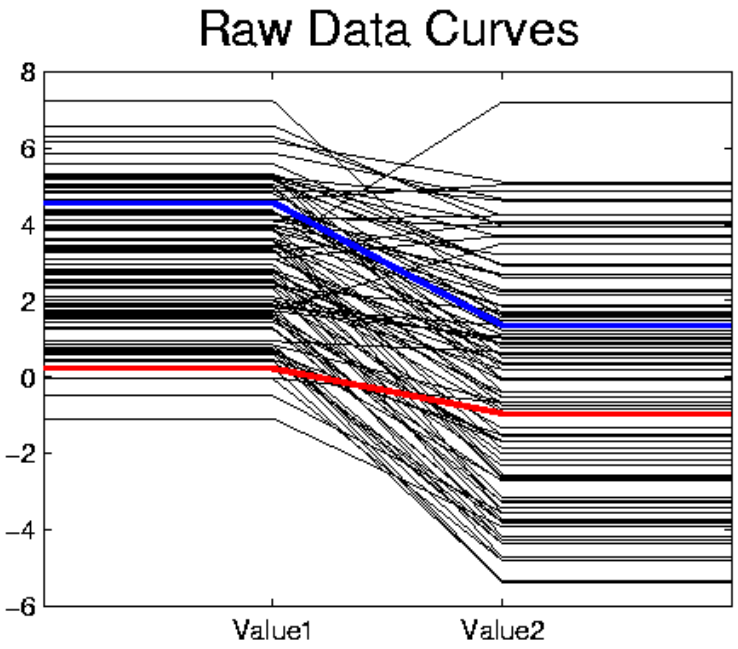
Decomposition into *modes of variation*

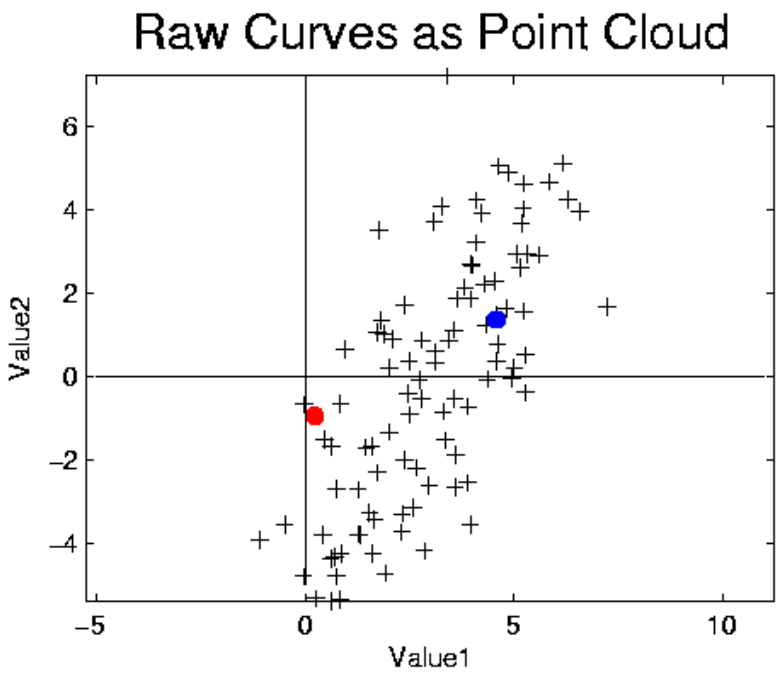
Functional Data Visualization



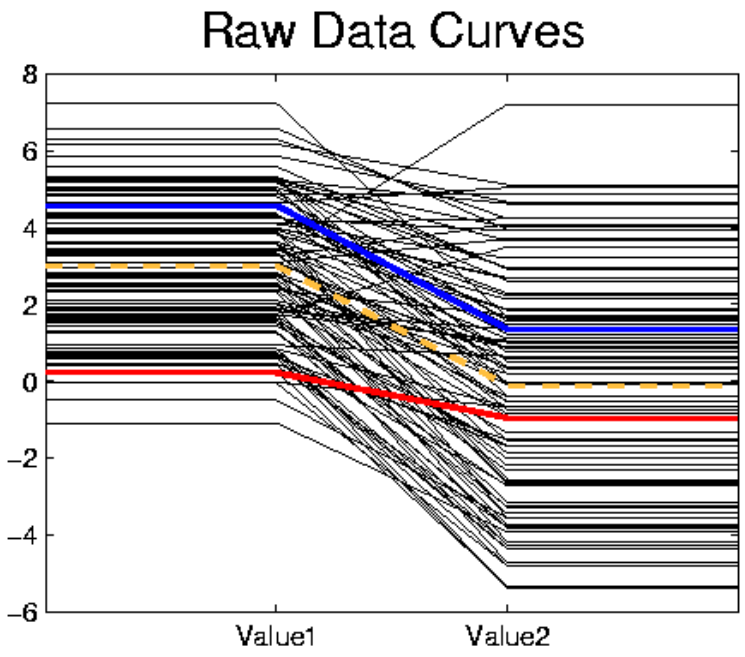


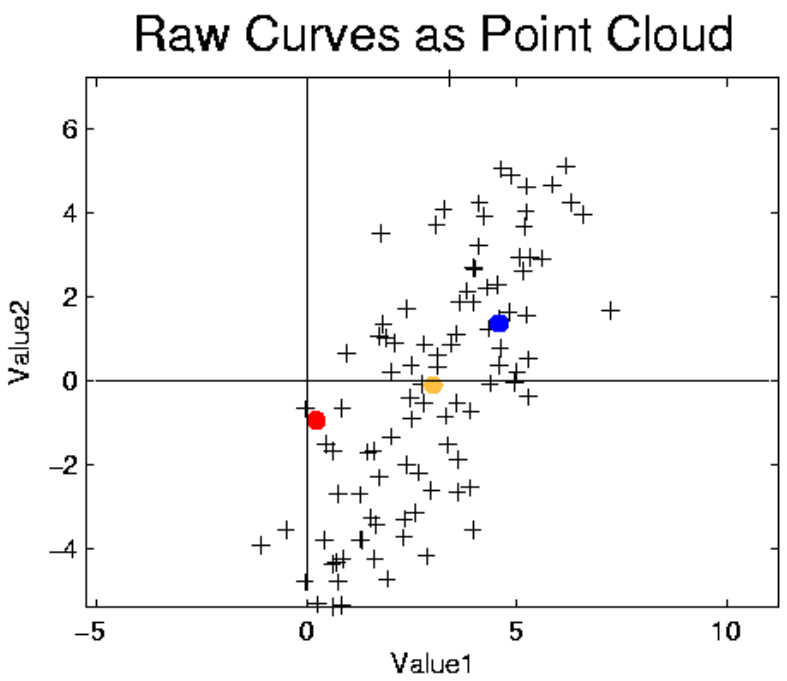
Functional Data Visualization



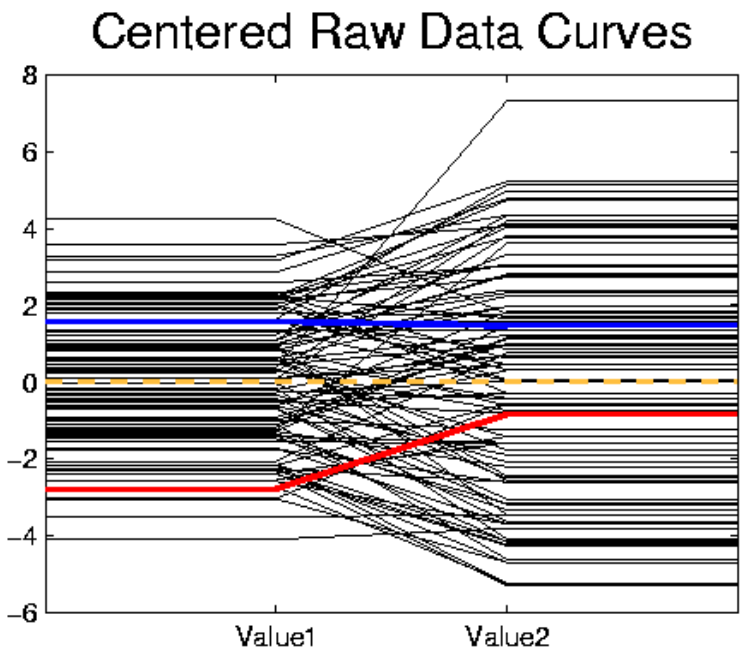


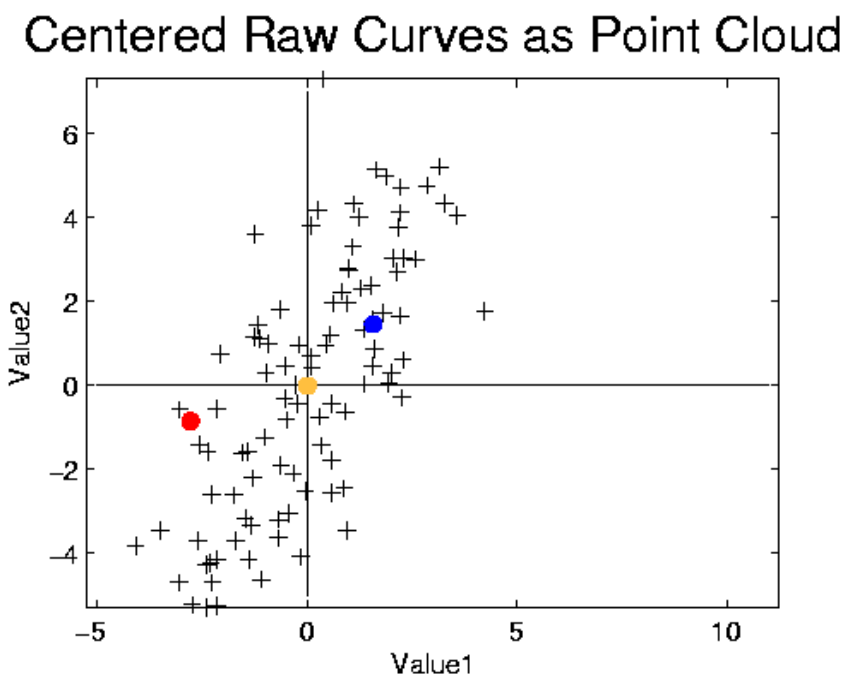
Functional Data Visualization



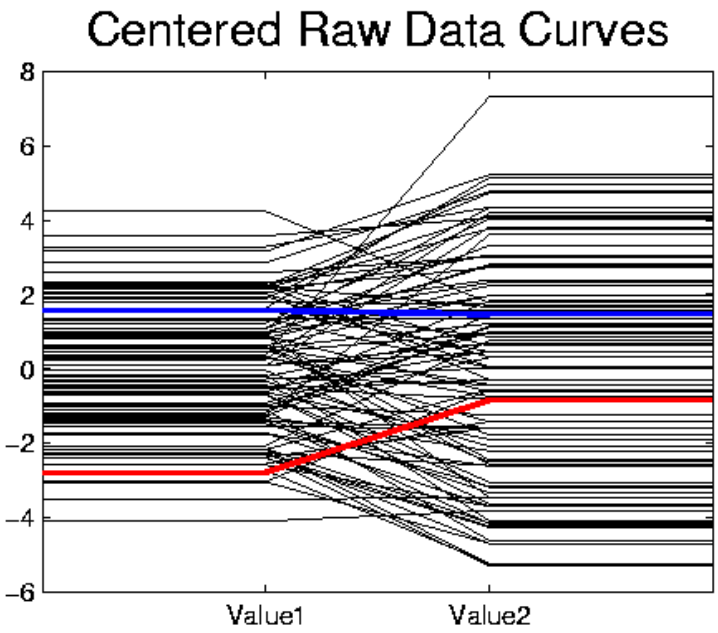


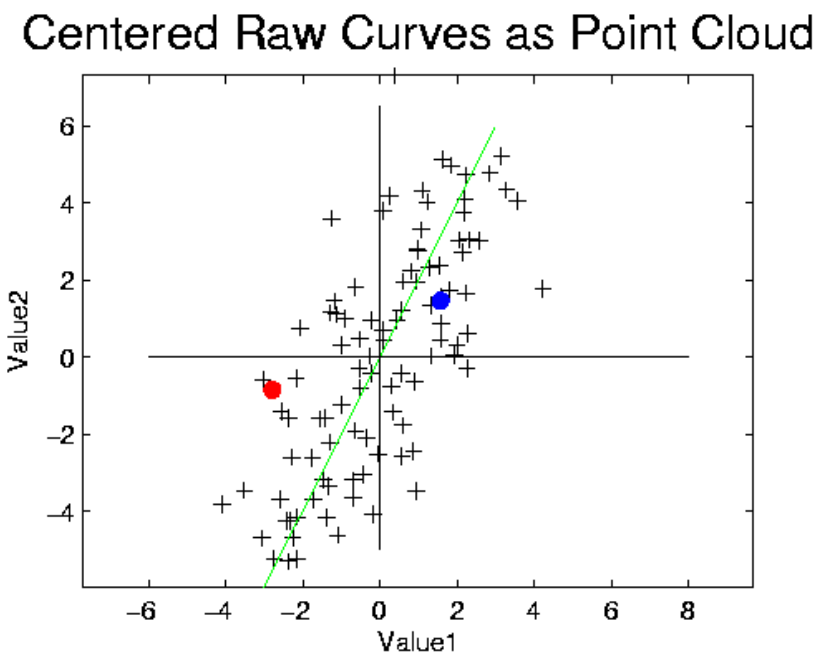
Functional Data Visualization



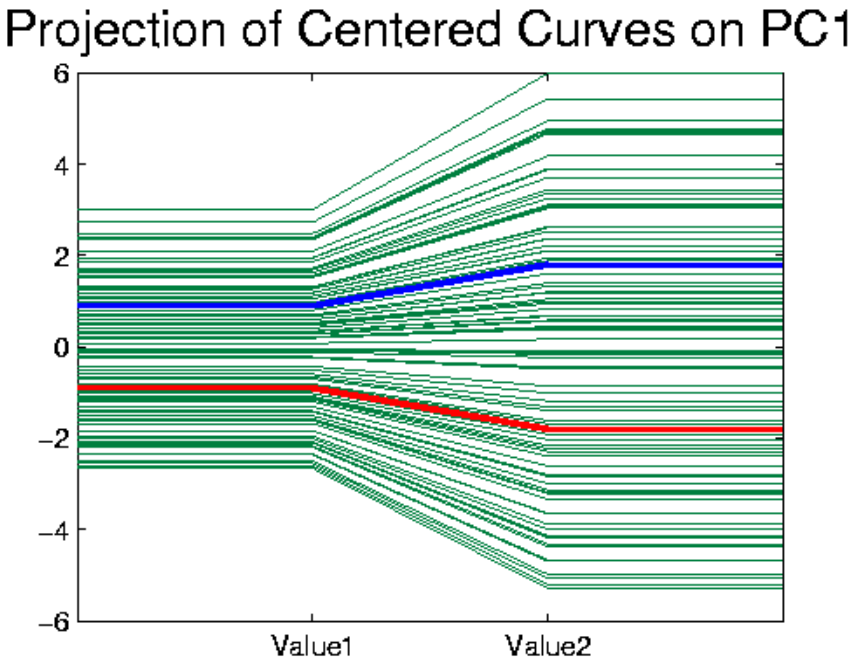


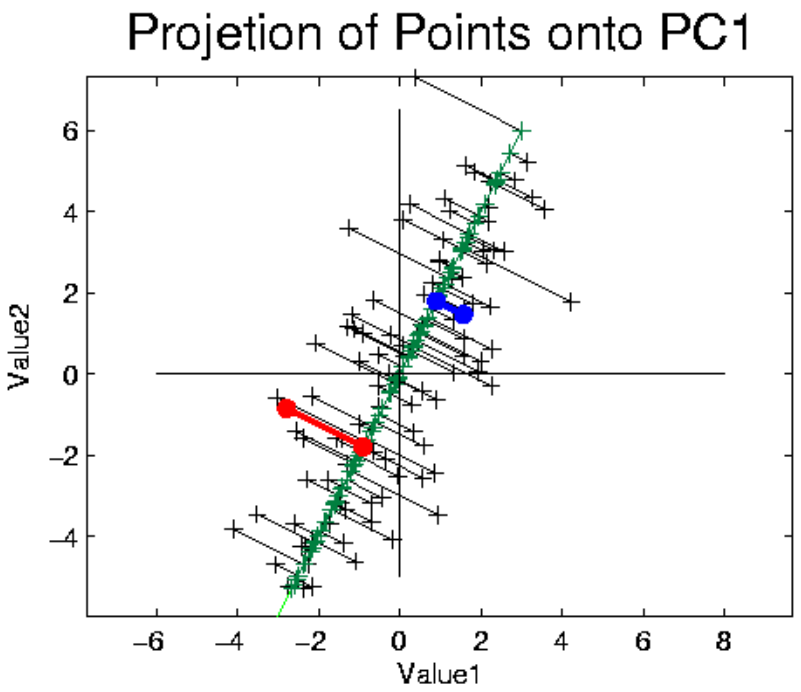
Functional Data Visualization



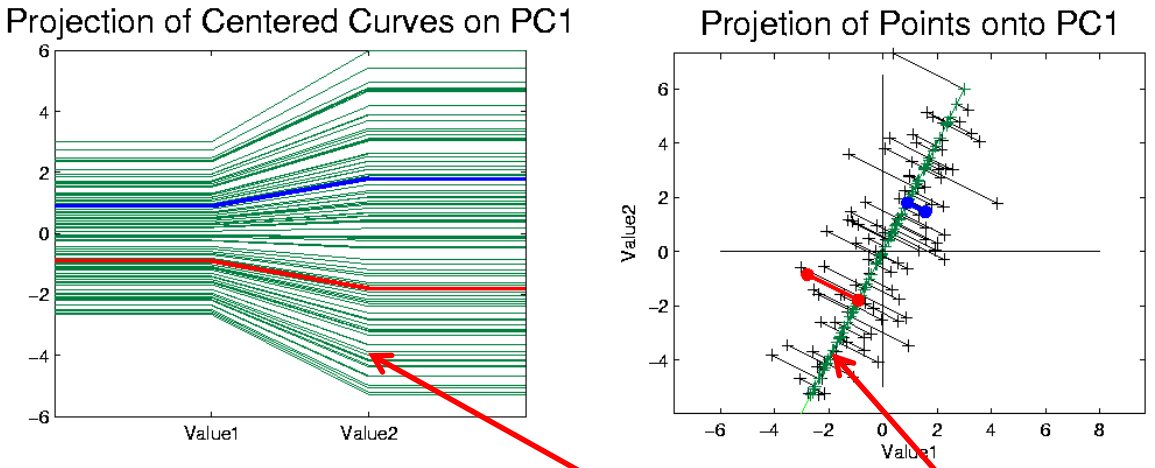


Functional Data Visualization





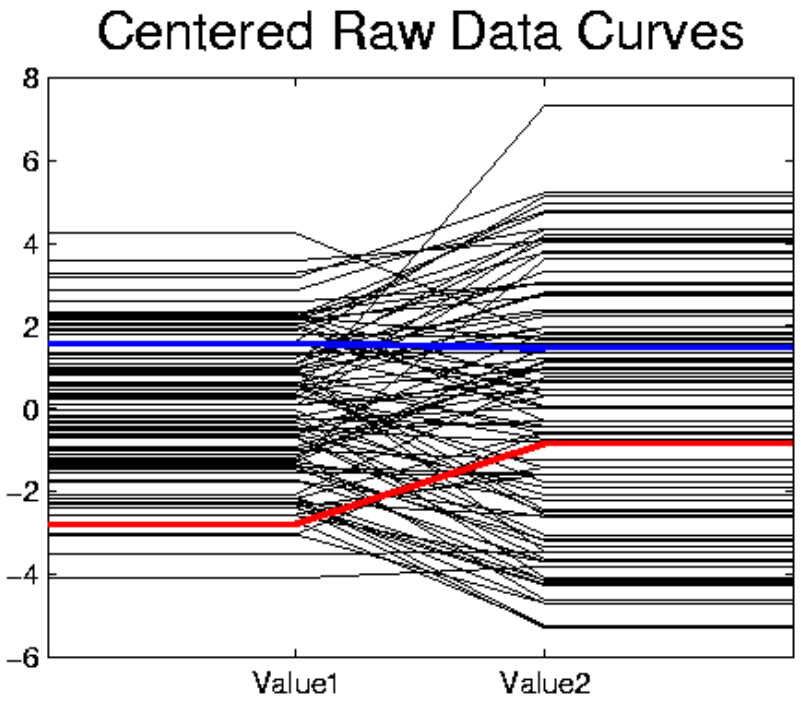
Functional Data Visualization

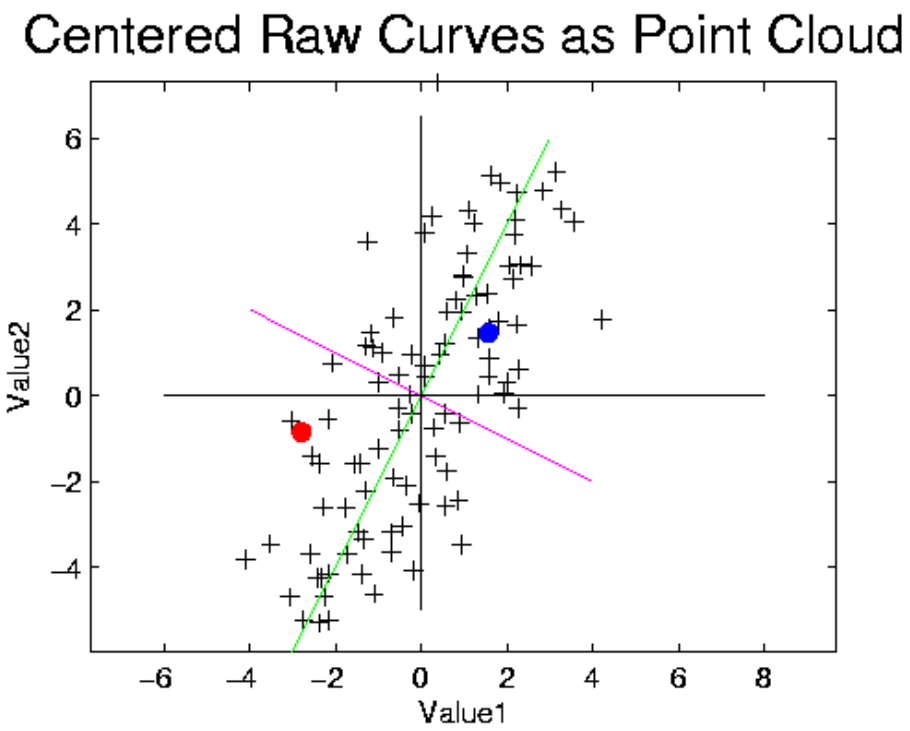


Classical Terminology:

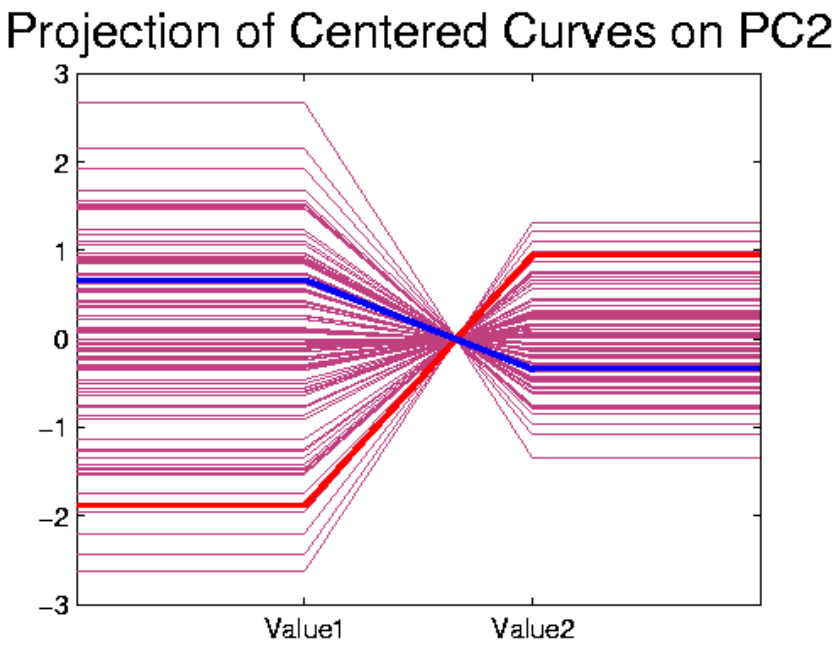
- Coefficients of Projections are “Scores”
- Entries of Direction Vector are “Loadings”

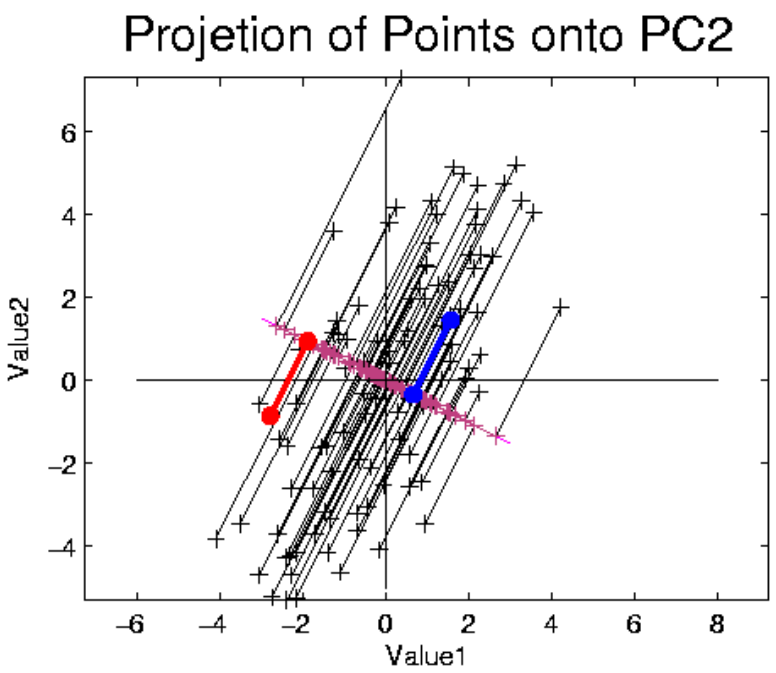
Functional Data Visualization



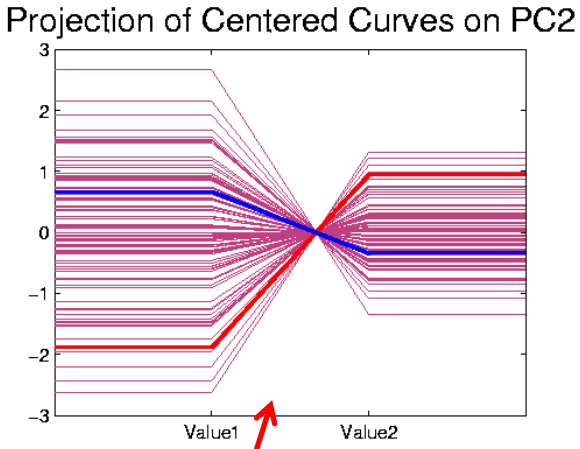


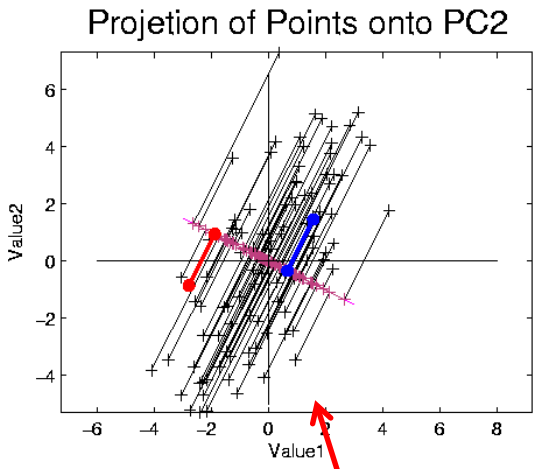
Functional Data Visualization





Functional Data Visualization



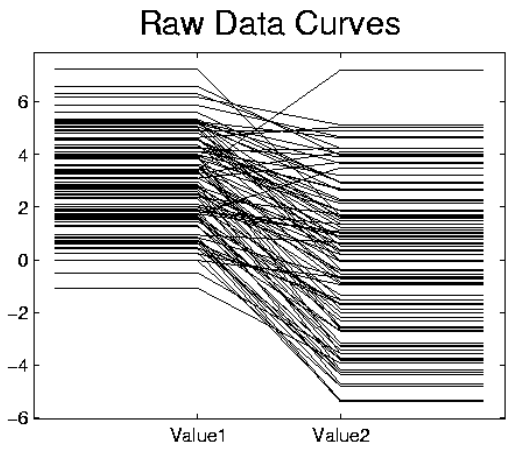


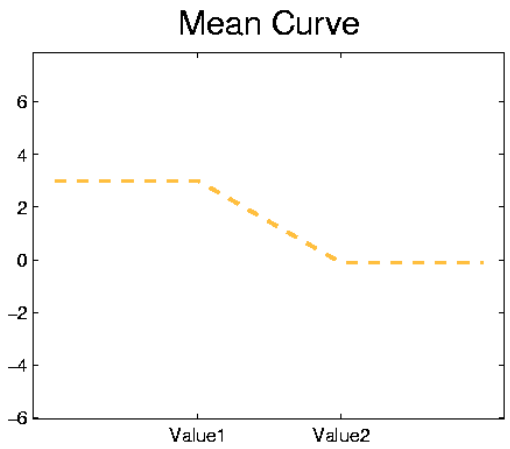
Terminology:

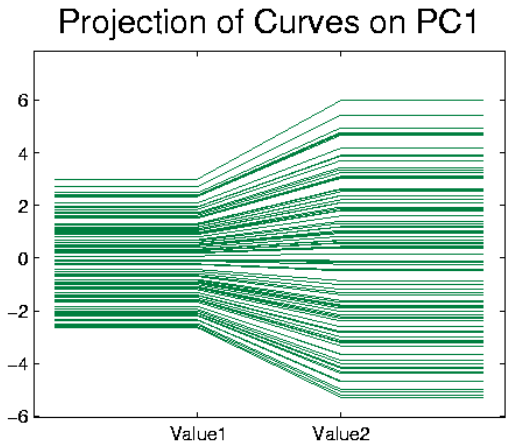
"Loadings Plot"

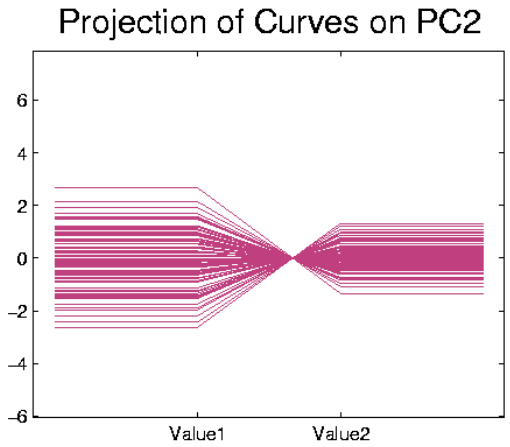
"Scores Plot"

Functional Data Visualization

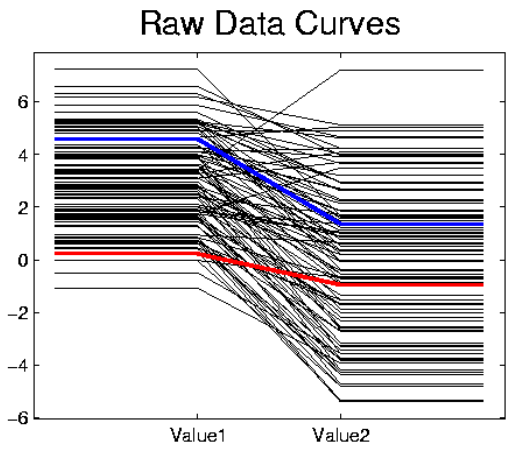


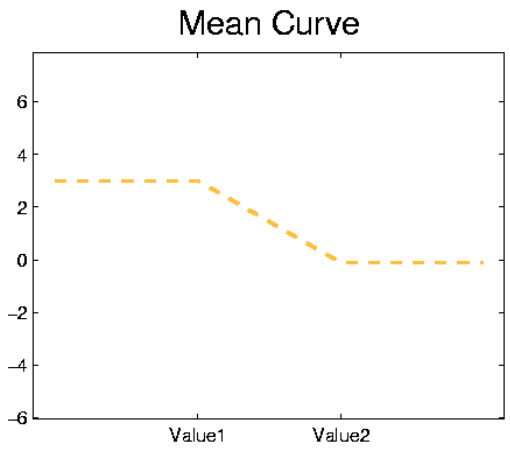


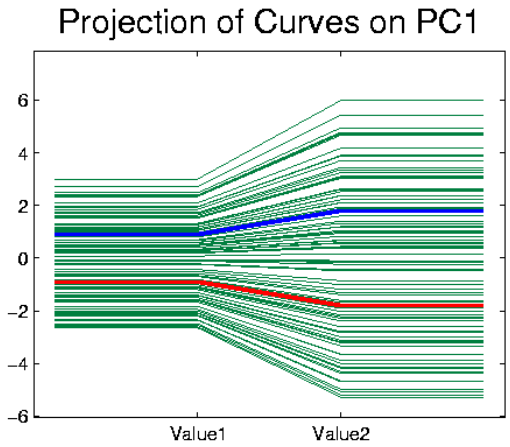


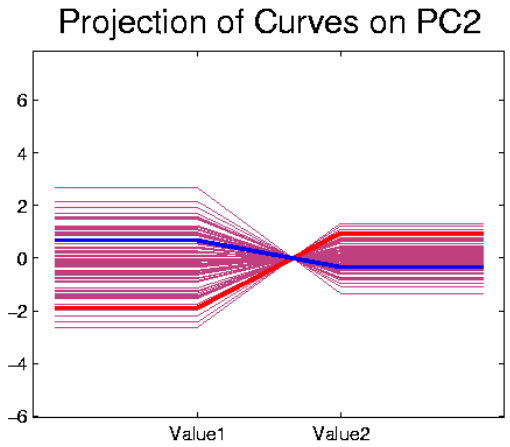


Functional Data Visualization









Curves As Data, II

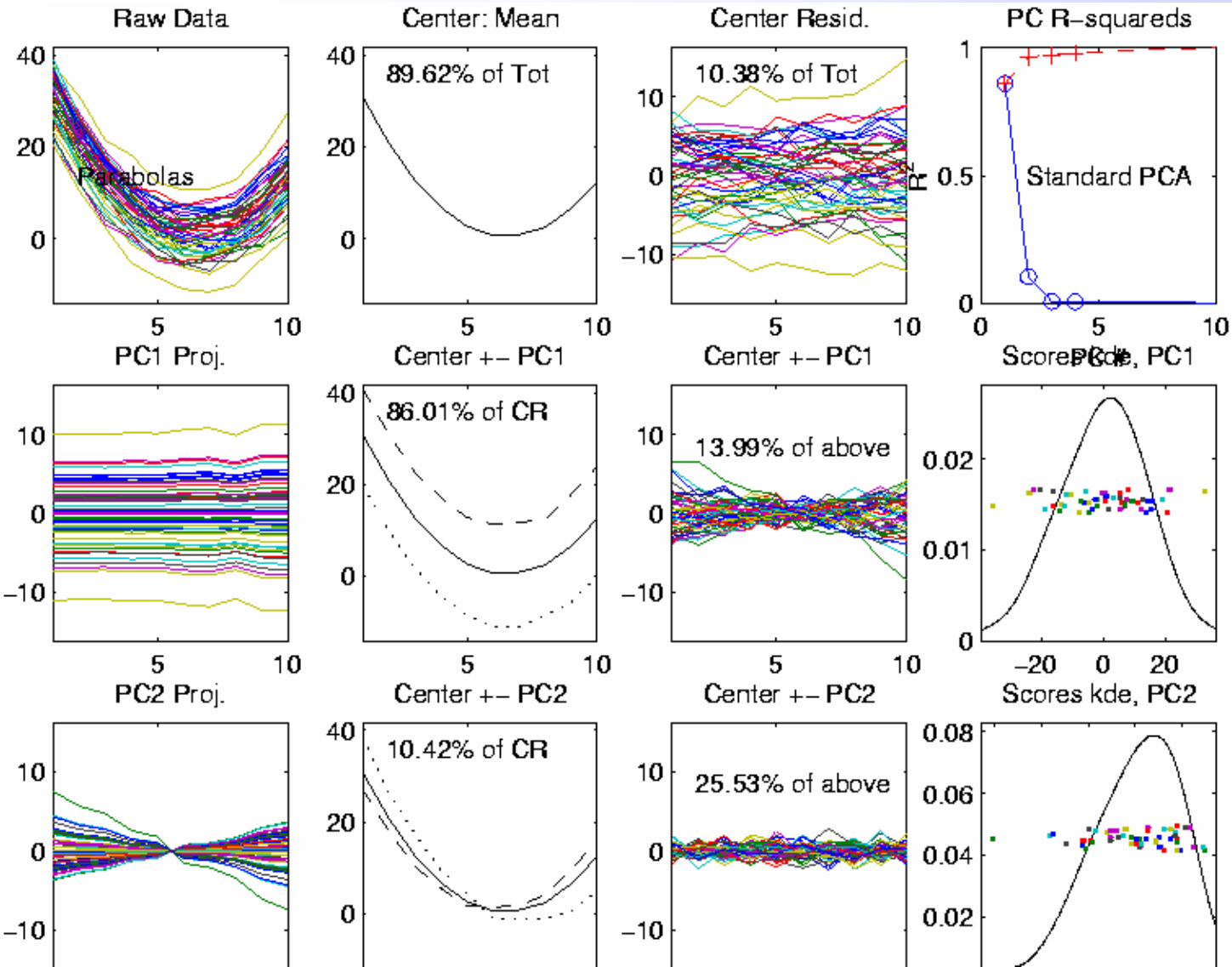
Deeper example

- 10-d family of (digitized) curves
- Object space: bundles of curves
- Descriptor space = $\mathbb{R}^{10}$

    (harder to visualize as point cloud)

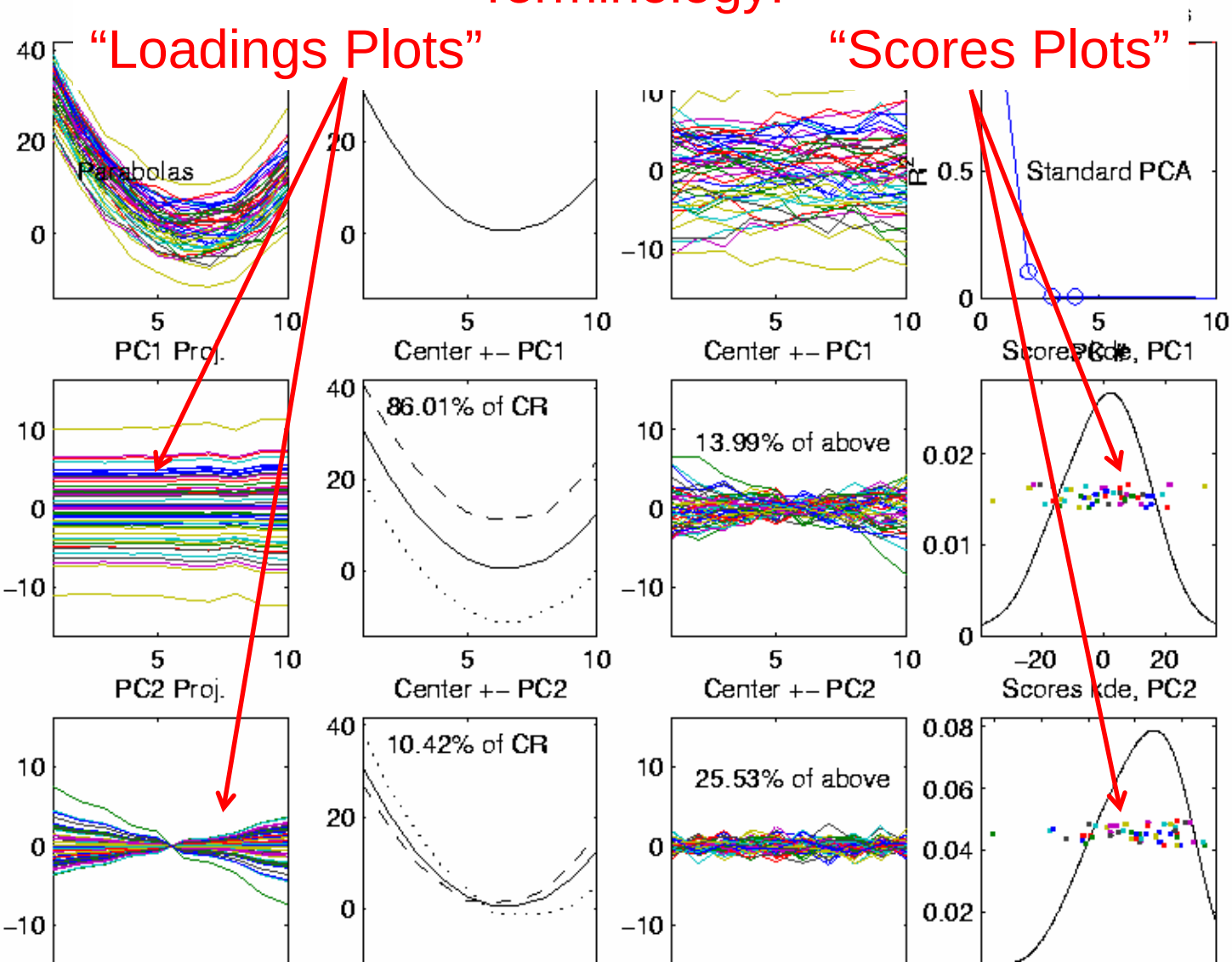
PCA: reveals “population structure”

Functional Data Analysis, 10-d

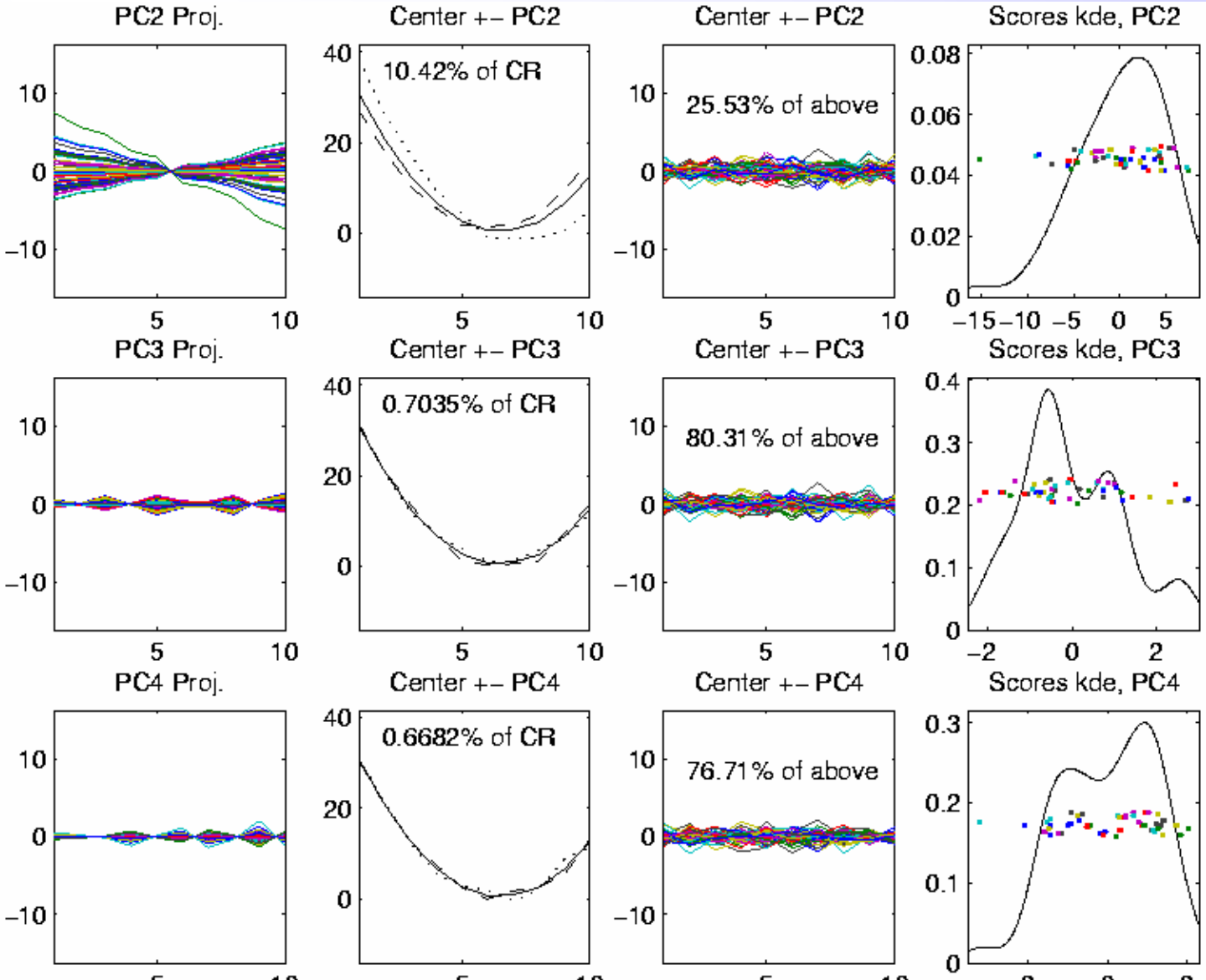


Functional Data Analysis, 10-d

Terminology:



Functional Data Analysis, 10-d



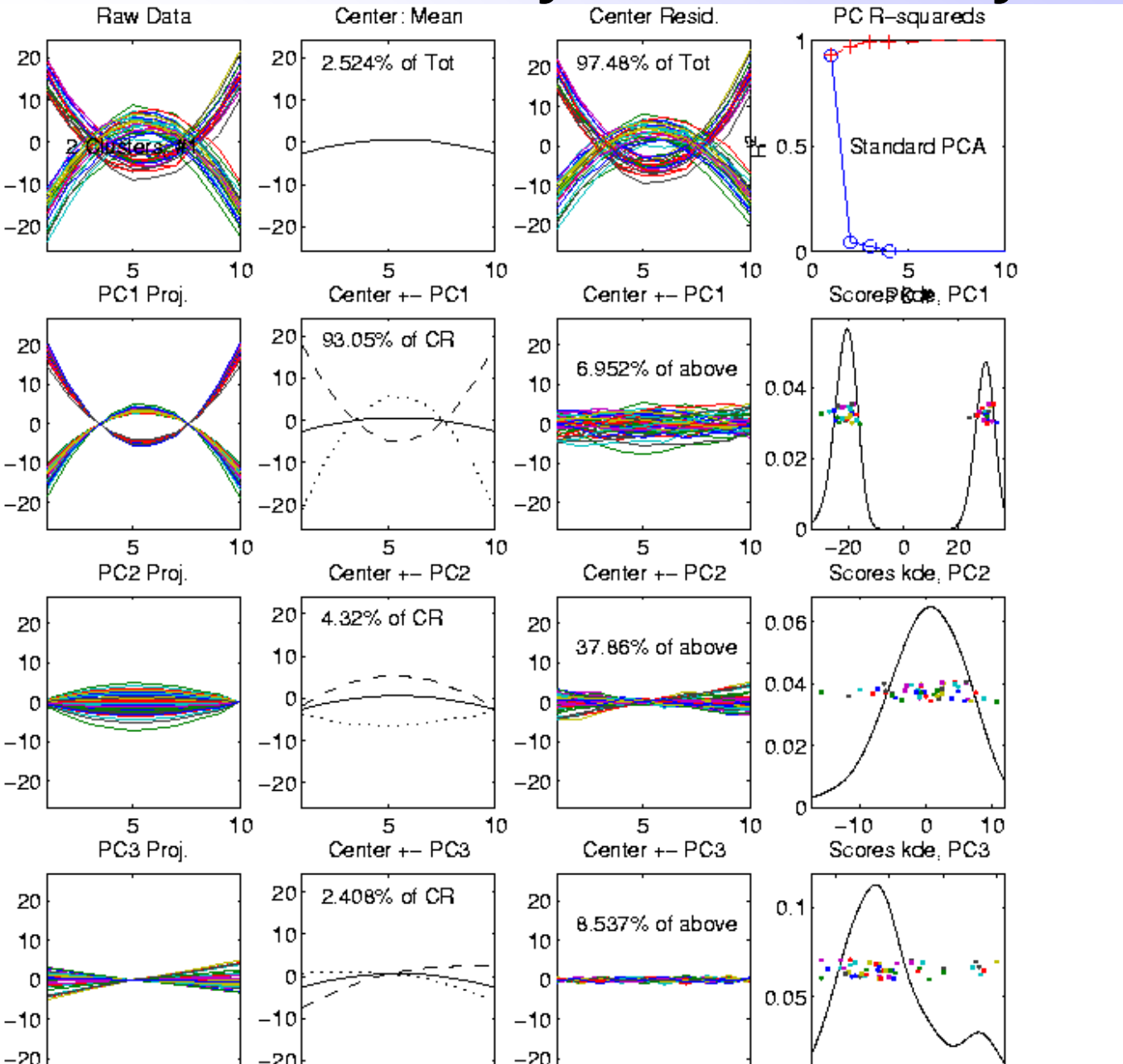
Curves As Data, II

PCA: reveals “population structure”

- Mean → Parabolic Structure
- PC1 → Vertical Shift
- PC2 → Tilt
- higher PCs → Gaussian (spherical)

Decomposition into *modes of variation*

Functional Data Analysis, 10-d Toy Ex 2



Curves As Data, III

Two Cluster Example

- 10-d curves again
- Two big clusters
- Revealed by 1-d projection plot (left side)
- Note: *Cluster Difference* is not orthogonal to *Vertical Shift*

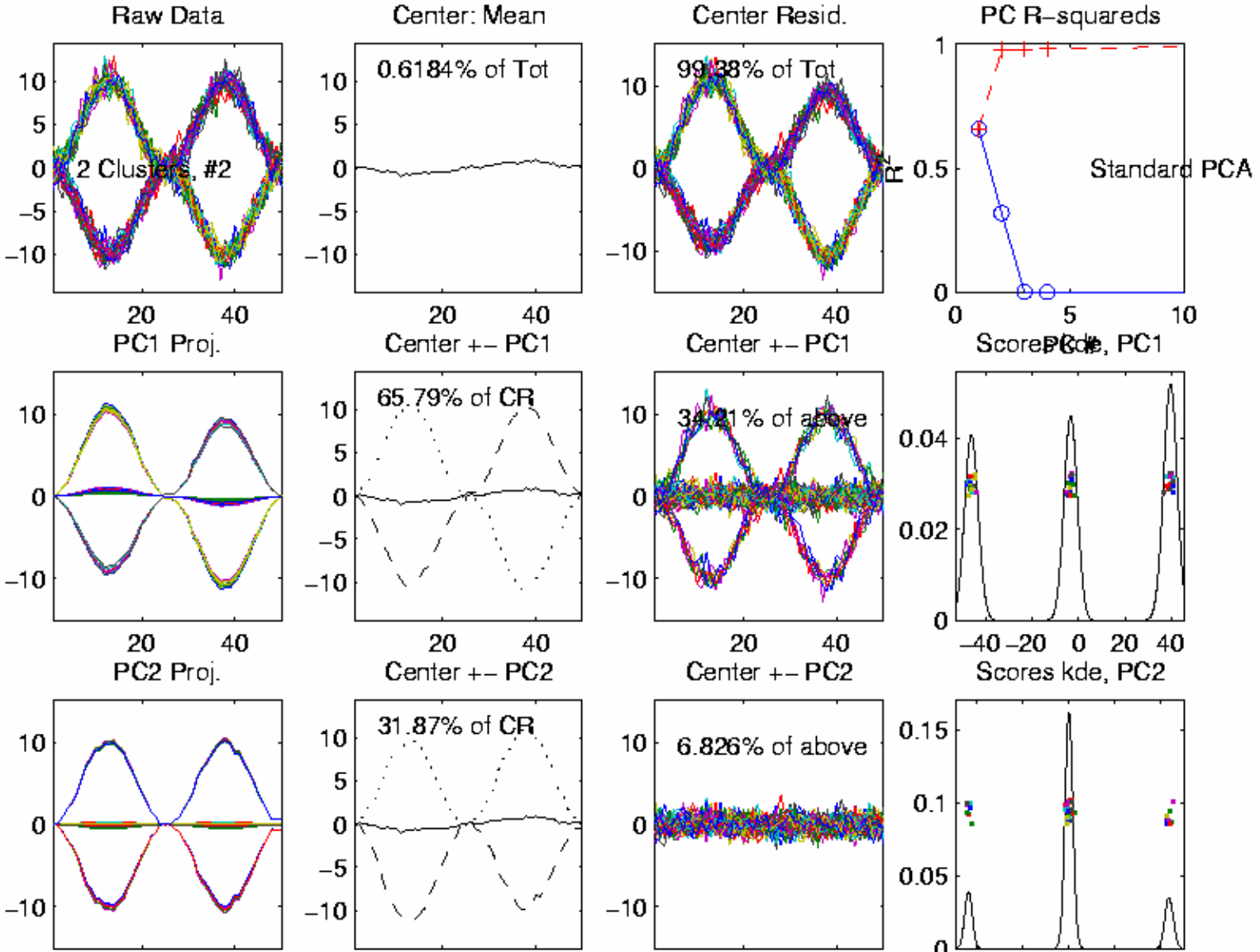
PCA: reveals “population structure”

Curves As Data, IV

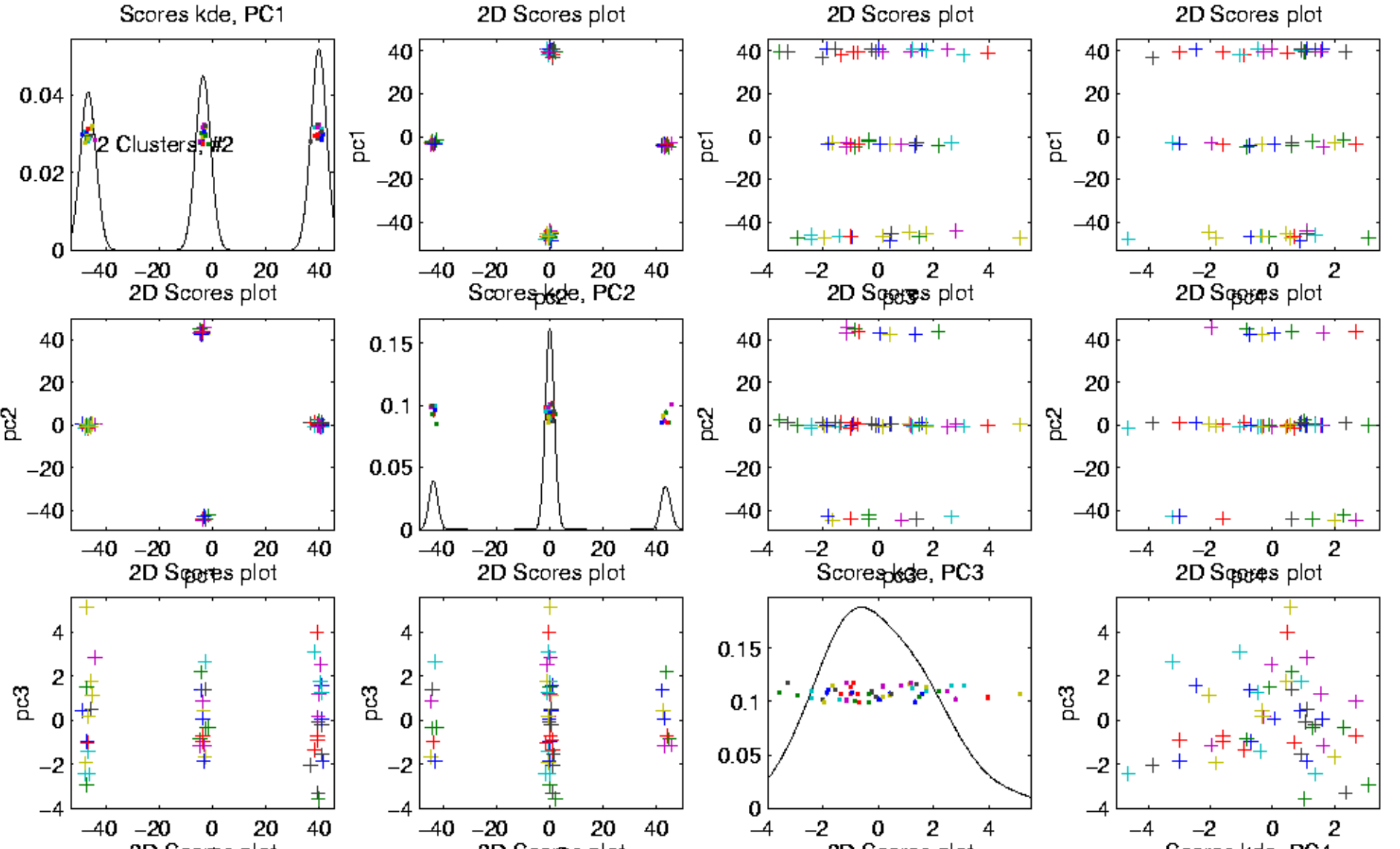
More Complicated Example

- 50-d curves

Functional Data Analysis, 50-d Toy Ex 3



Functional Data Analysis, 50-d Toy Ex 3



E.g. Curves As Data, IV

More Complicated Example

- 50-d curves
- Pop’n structure hard to see in 1-d
- 2-d projections make structure clear

PCA: reveals “population structure”

Functional Data Analysis

Interesting Data Set:

- Mortality Data
- For Spanish Males (thus can relate to history)
- Each curve is a single year
- x coordinate is age
- Mortality = # died / total # (for each age)
- Study on log scale
- Investigate *change* over years 1908 – 2002

Note: Choice made of *Data Object*
(could also study age as curves,
x coordinate = time)

Functional Data Analysis

Important Issue:

What are the Data Objects?

- Mortality vs. Age Curves (over years)
- Mortality vs. Year Curves (over ages)

Note: Rows vs. Columns of Data Matrix

Functional Data Analysis

Important Issue:

What are the Data Objects?

- Mortality vs. Age Curves (over years)
- Mortality vs. Year Curves (over ages)

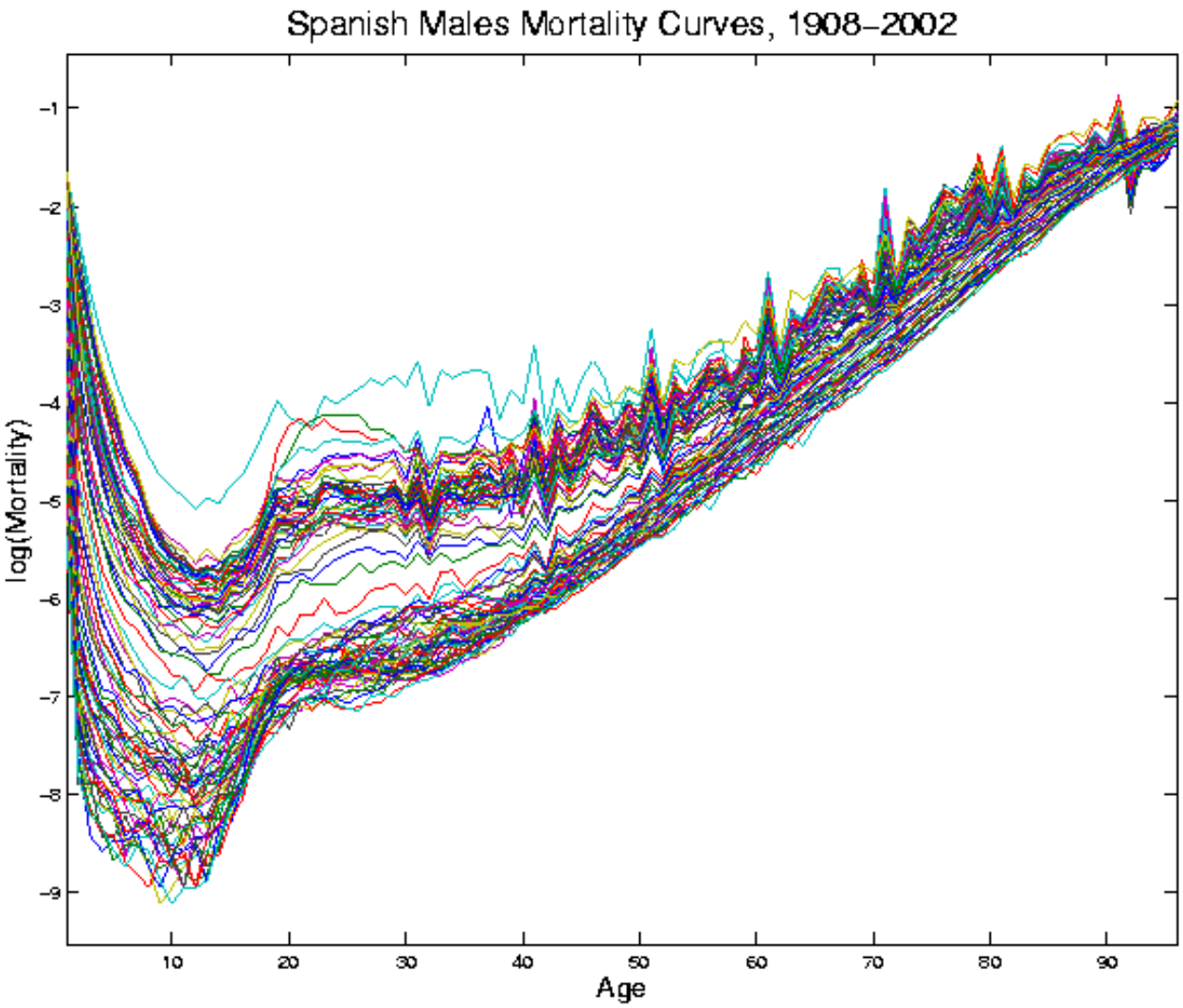
Note: Rows vs. Columns of Data Matrix

Mortality Time Series

Conventional Coloring:

Rotate Through (7) Colors

Hard to See Time Structure



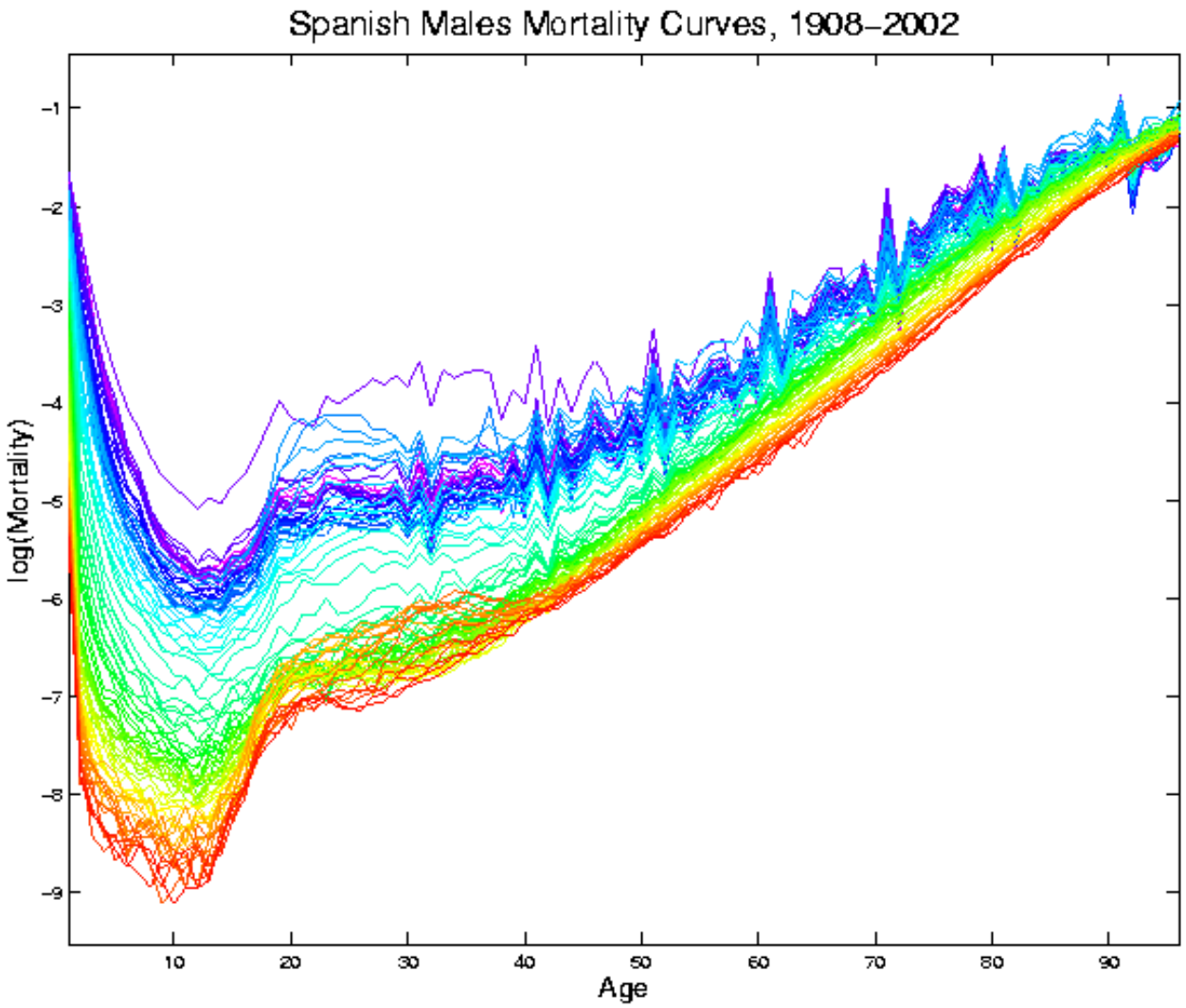
Mortality Time Series

Improved Coloring:

Rainbow Representing Year:

Magenta = 1908

Red = 2002



Mortality Time Series

Color Code (Years)

1908	1927	1946	1965	1984
1909	1928	1947	1966	1985
1910	1929	1948	1967	1986
1911	1930	1949	1968	1987
1912	1931	1950	1969	1988
1913	1932	1951	1970	1989
1914	1933	1952	1971	1990
1915	1934	1953	1972	1991
1916	1935	1954	1973	1992
1917	1936	1955	1974	1993
1918	1937	1956	1975	1994
1919	1938	1957	1976	1995
1920	1939	1958	1977	1996
1921	1940	1959	1978	1997
1922	1941	1960	1979	1998
1923	1942	1961	1980	1999
1924	1943	1962	1981	2000
1925	1944	1963	1982	2001
1926	1945	1964	1983	2002

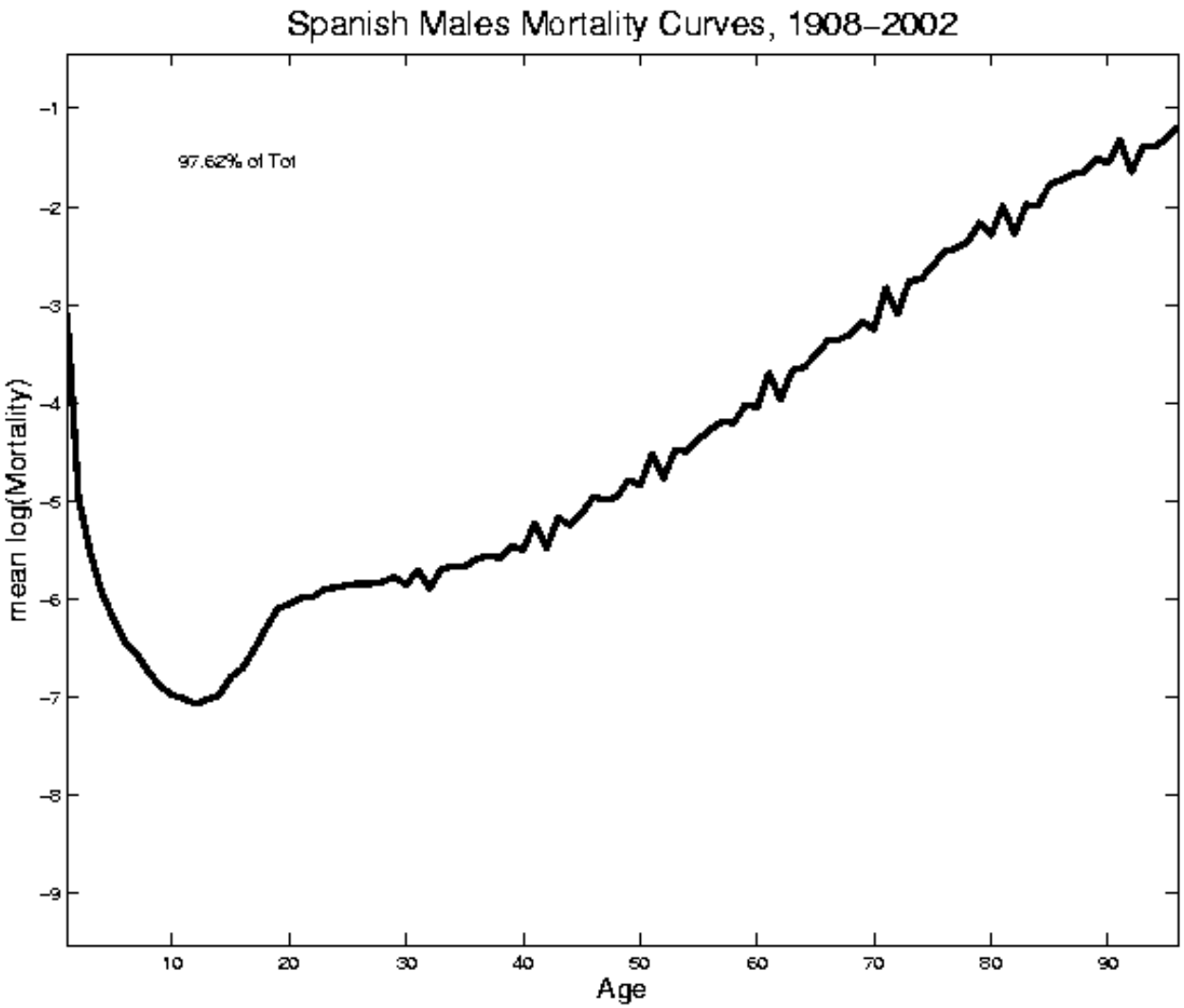
Mortality Time Series

Find Population Center

(Mean Vector)

Compute in Descriptor Space

Show in Object Space

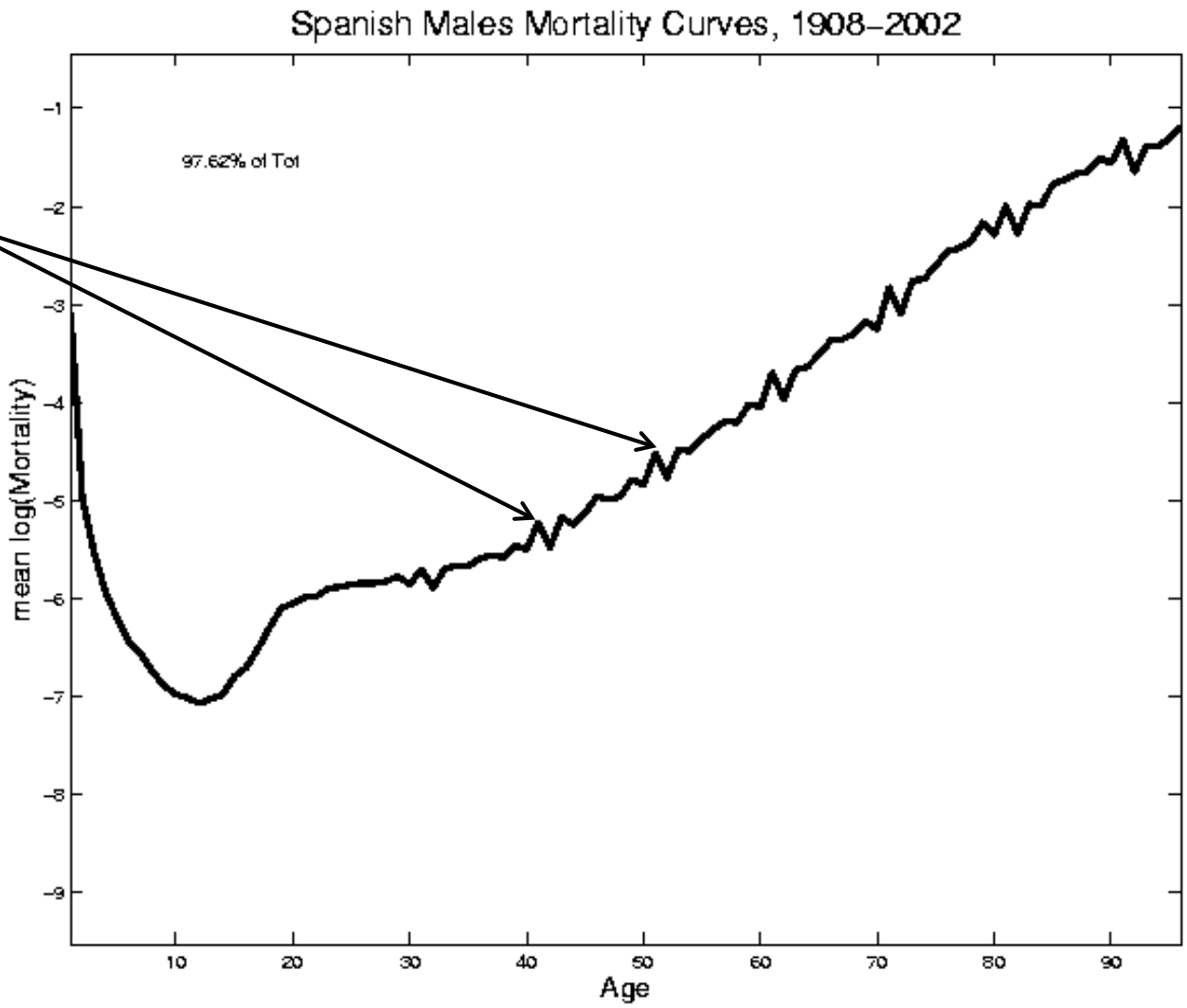


Mortality Time Series

Blips Appear At Decades

Since Ages Not Precise (in Spain)

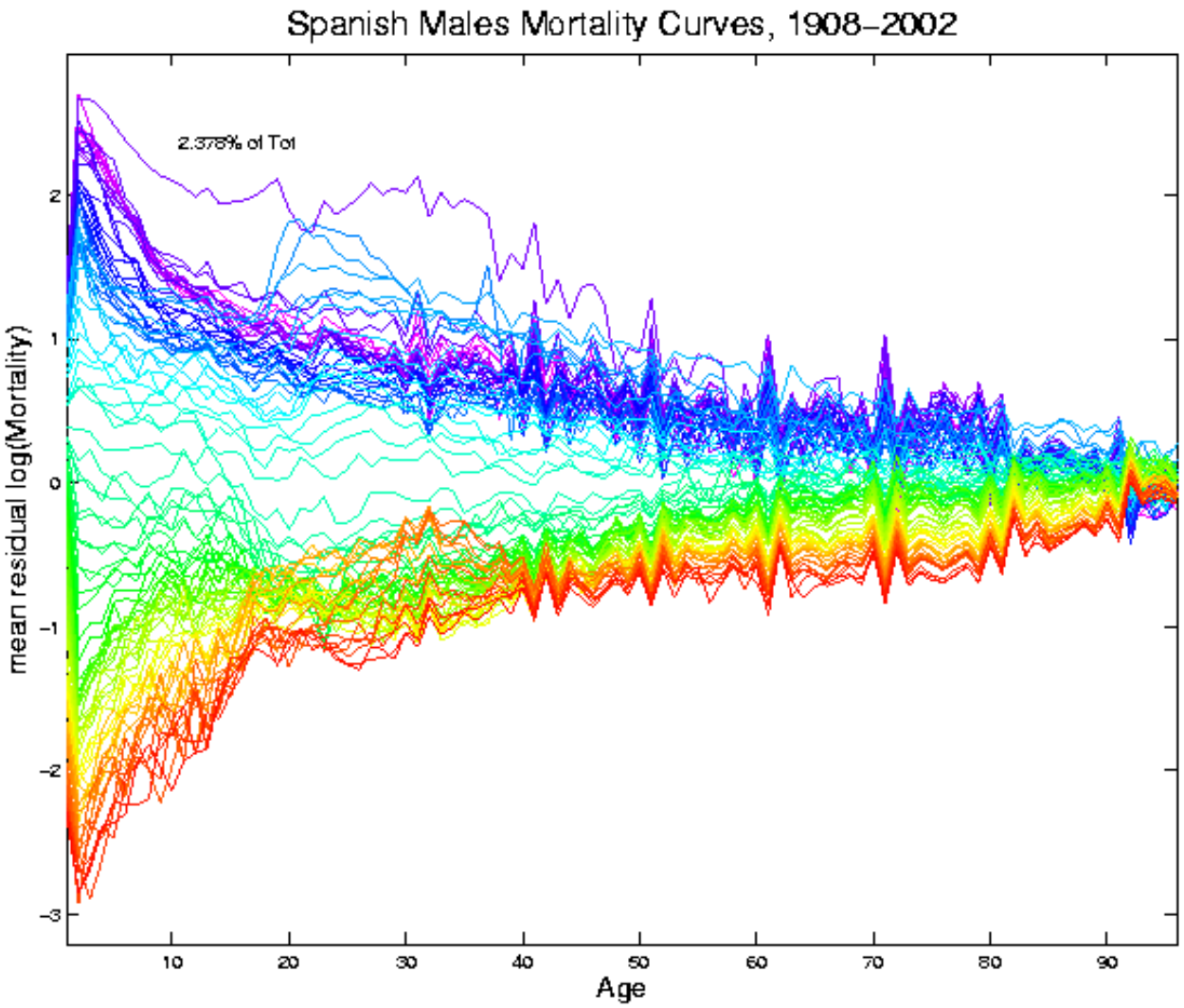
Reported as "about 50", Etc.



Mortality Time Series

Mean Residual

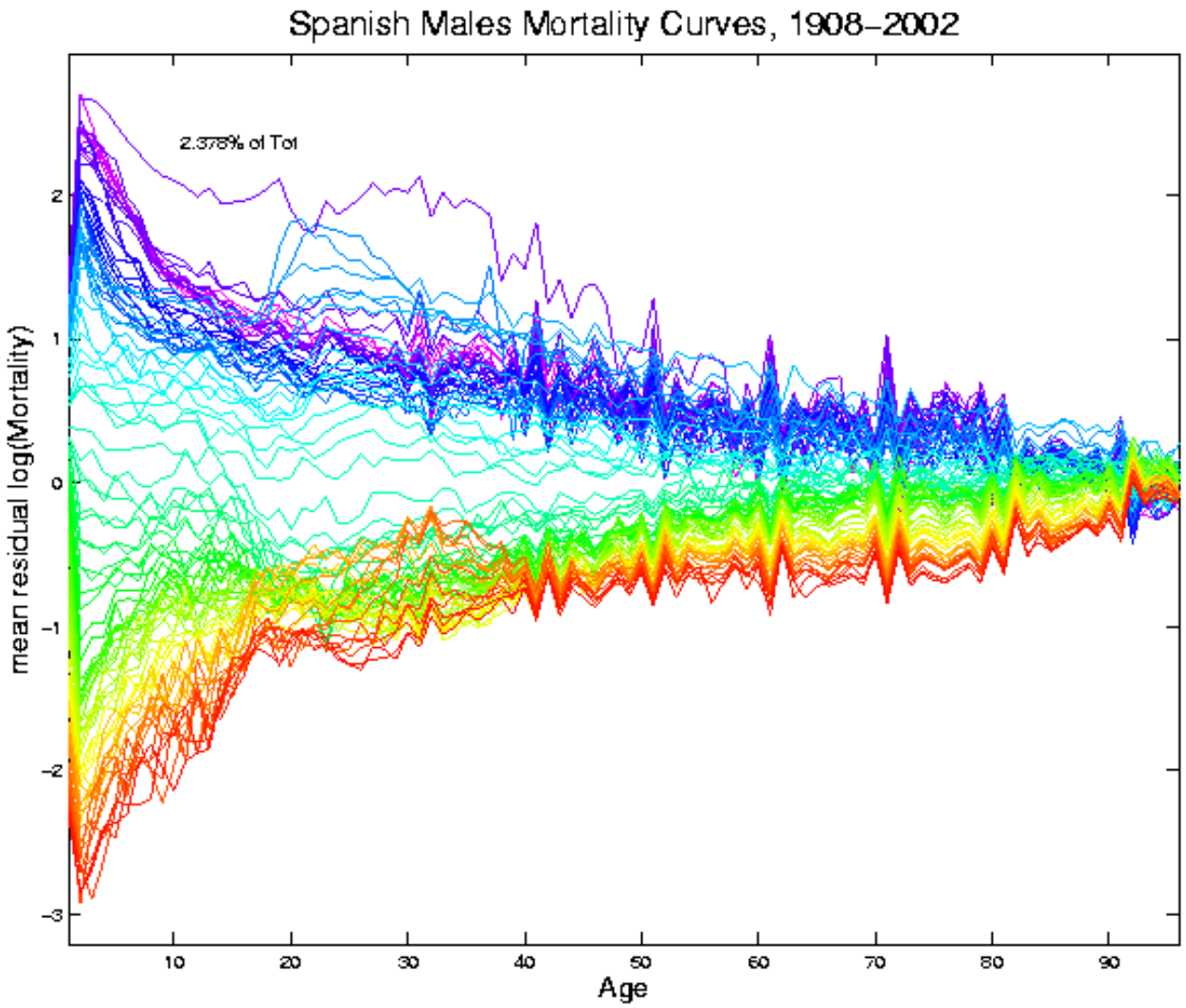
Descriptor Space View of Shifting Data To Origin In Feature Space



Mortality Time Series

Shows:

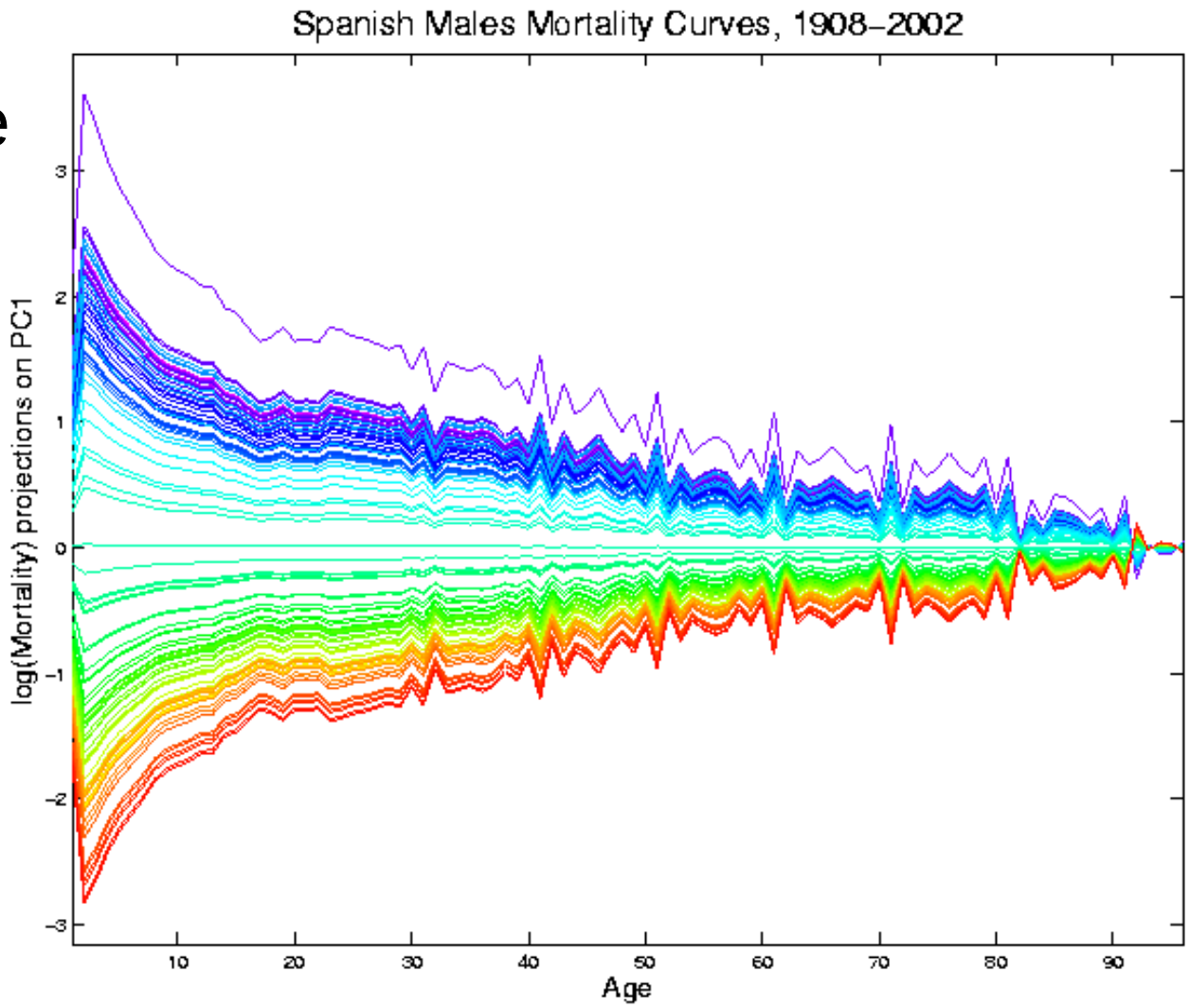
Main Age Effects in Mean, Not Variation About Mean



Mortality Time Series

Object Space View of Projections Onto PC1 Direction

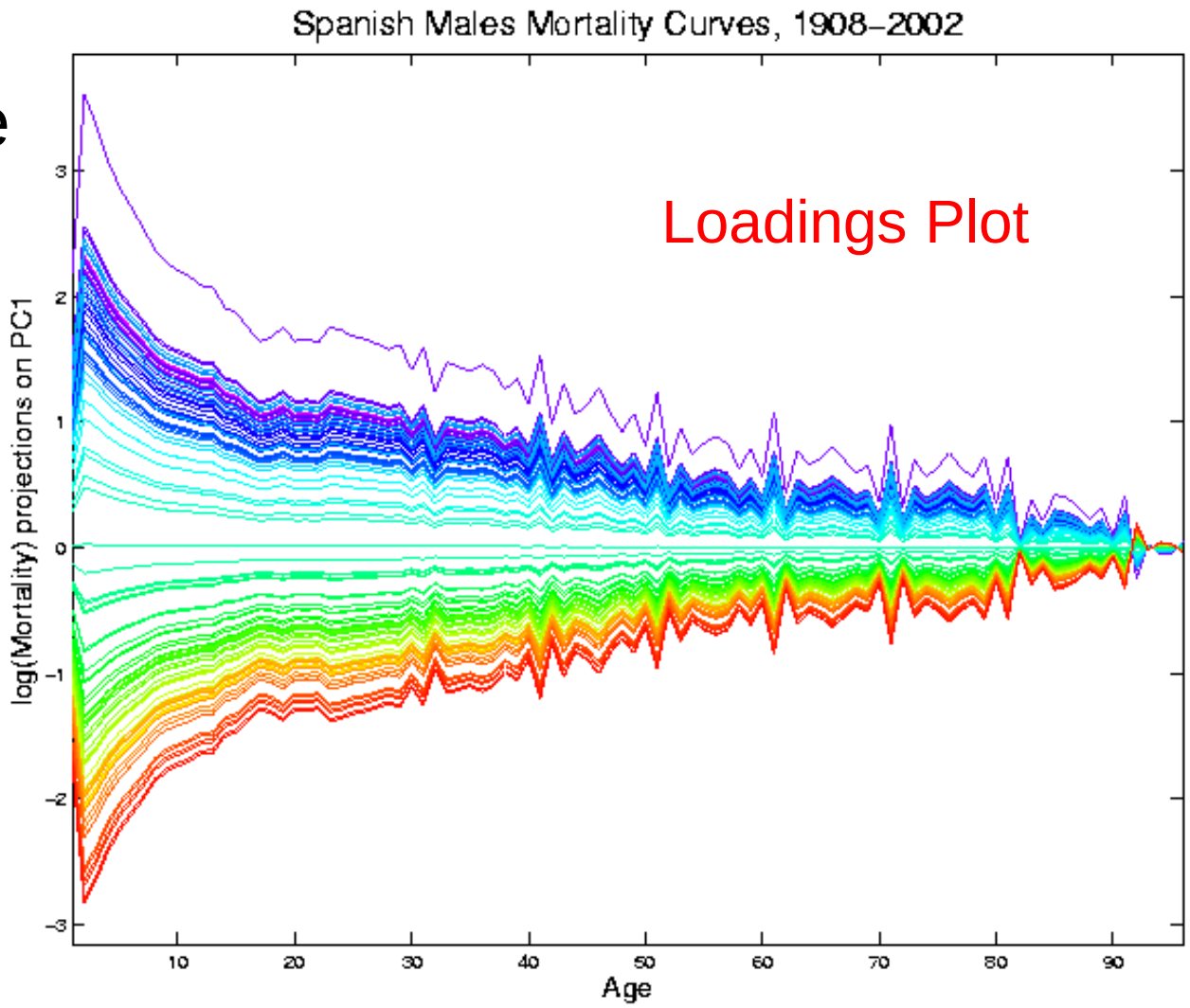
Main Mode Of Variation: Constant Across Ages



Mortality Time Series

Object Space View of Projections Onto PC1 Direction

Main Mode Of Variation: Constant Across Ages

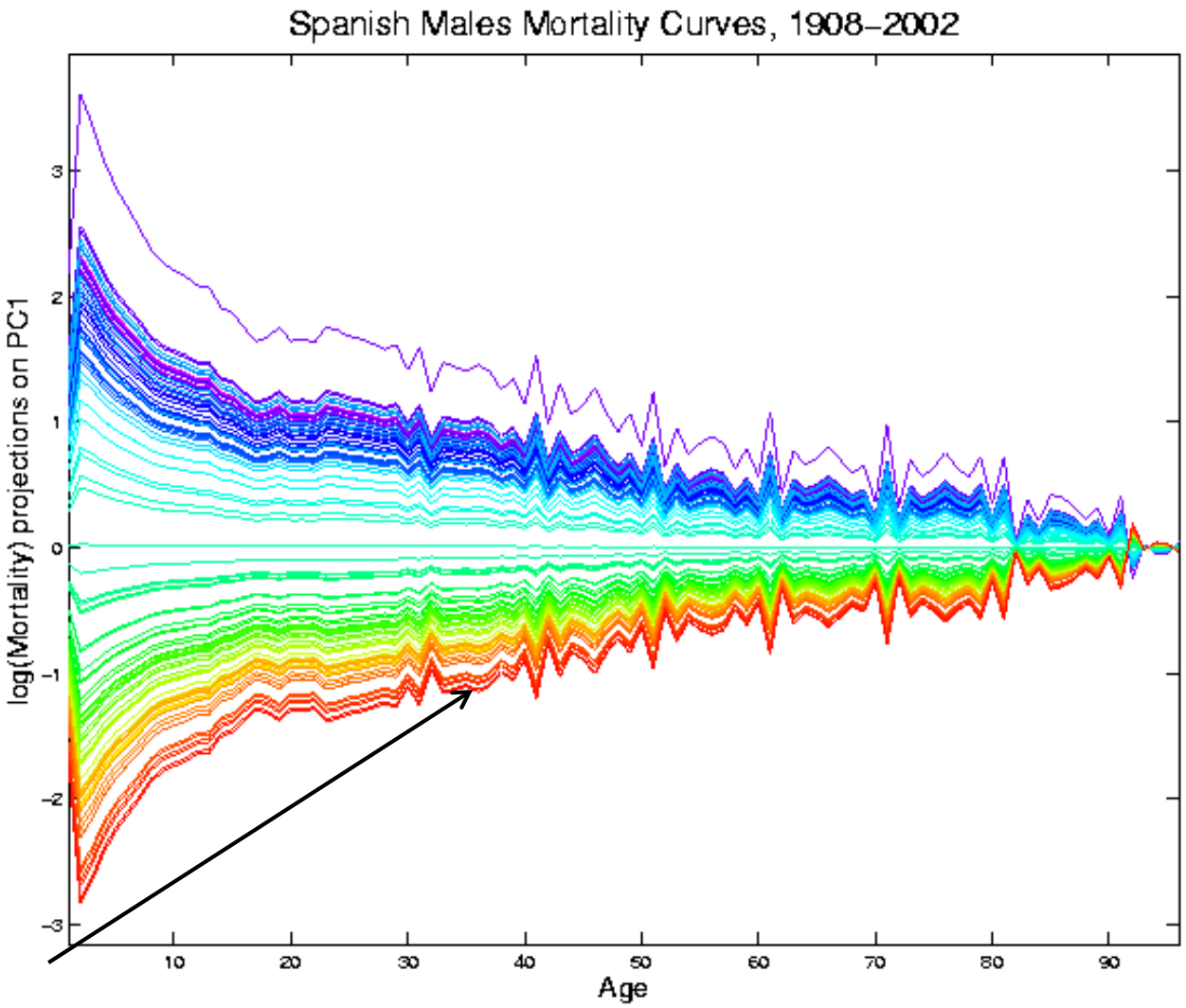


Mortality Time Series

Shows *Major* Improvement Over Time

(medical technology, etc.)

And Change In Age Rounding Blips

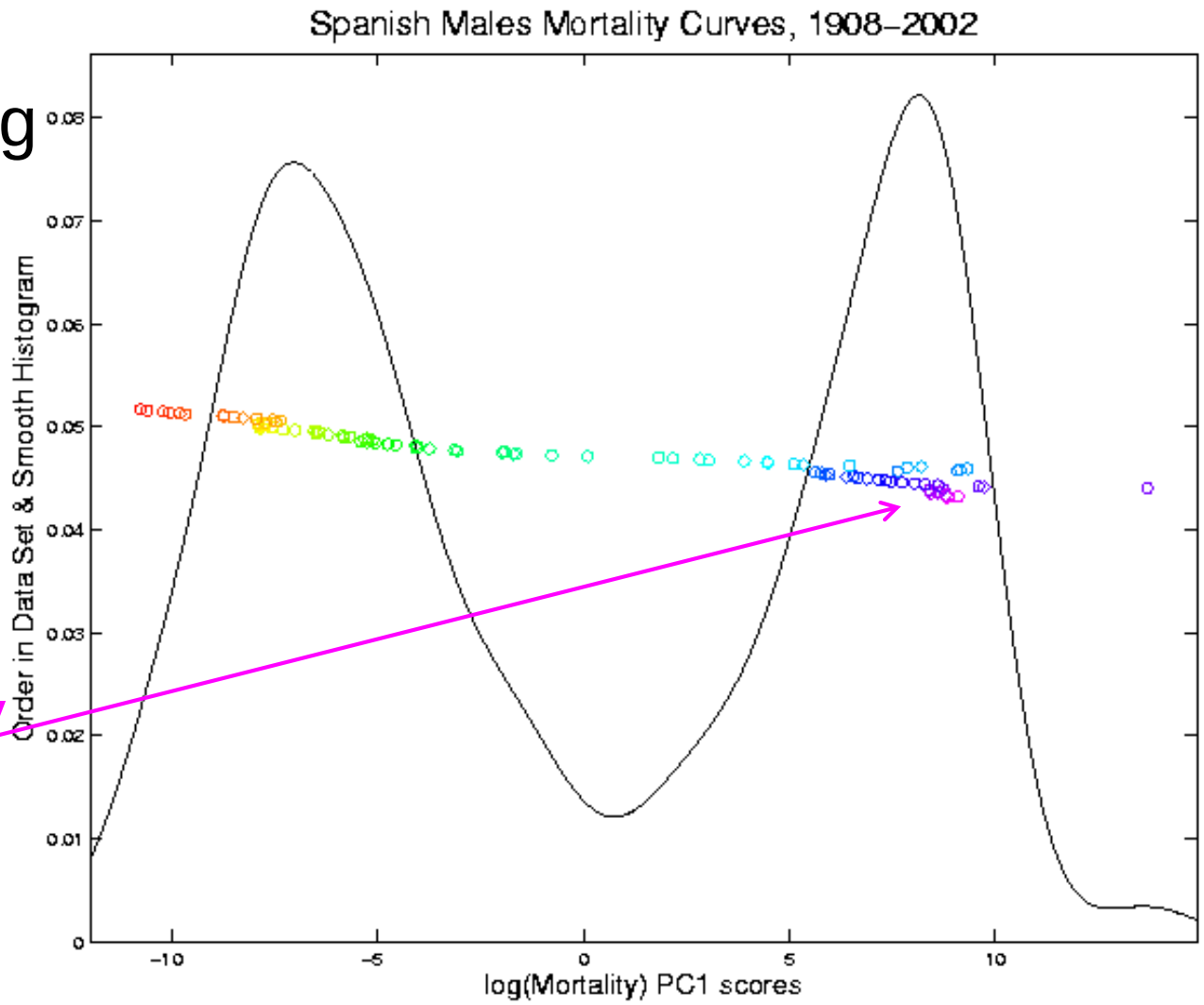


Mortality Time Series

Corresponding PC 1 Scores

Again Shows Overall Improvement

High Mortality Early



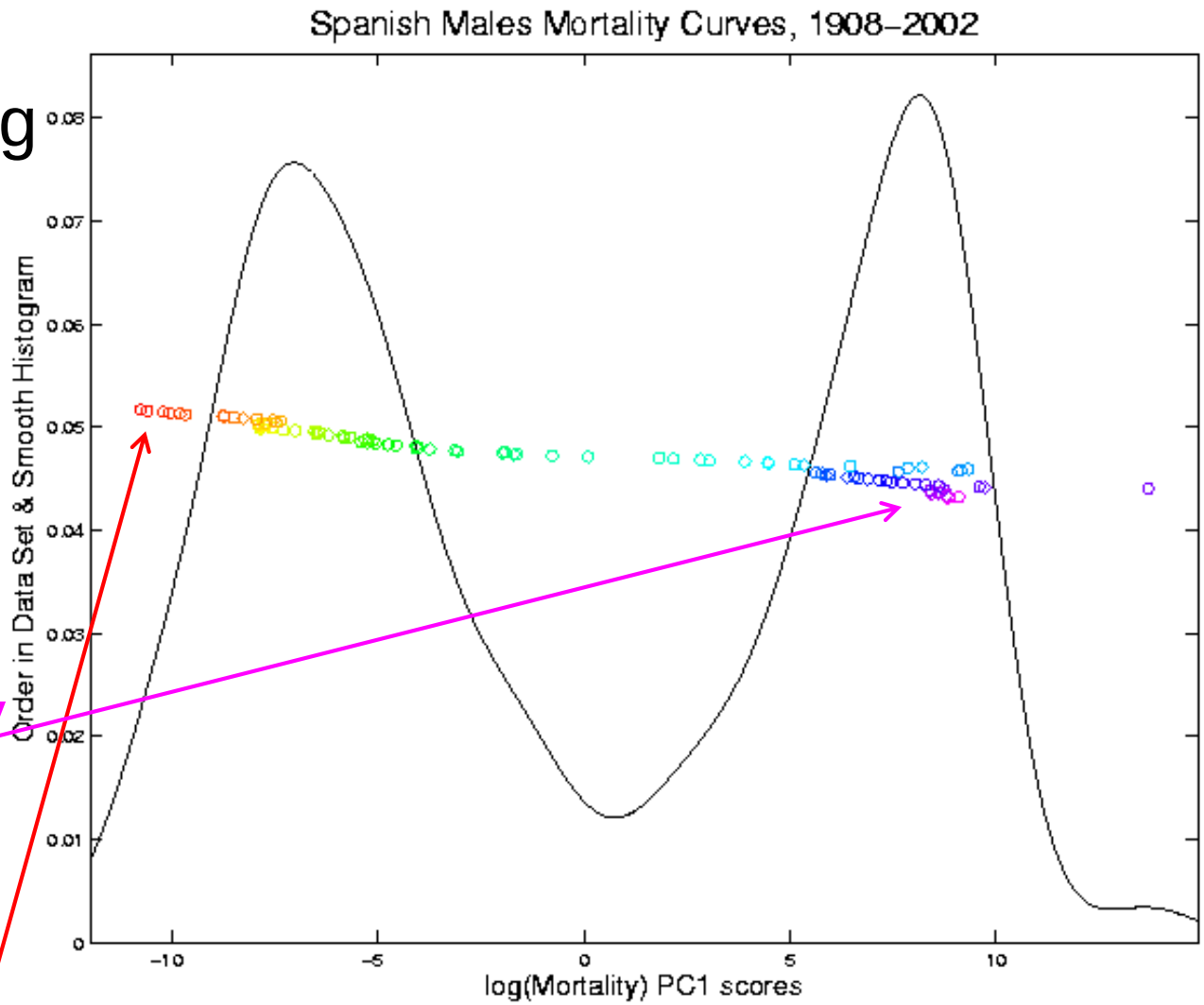
Mortality Time Series

Corresponding PC 1 Scores

Again Shows Overall Improvement

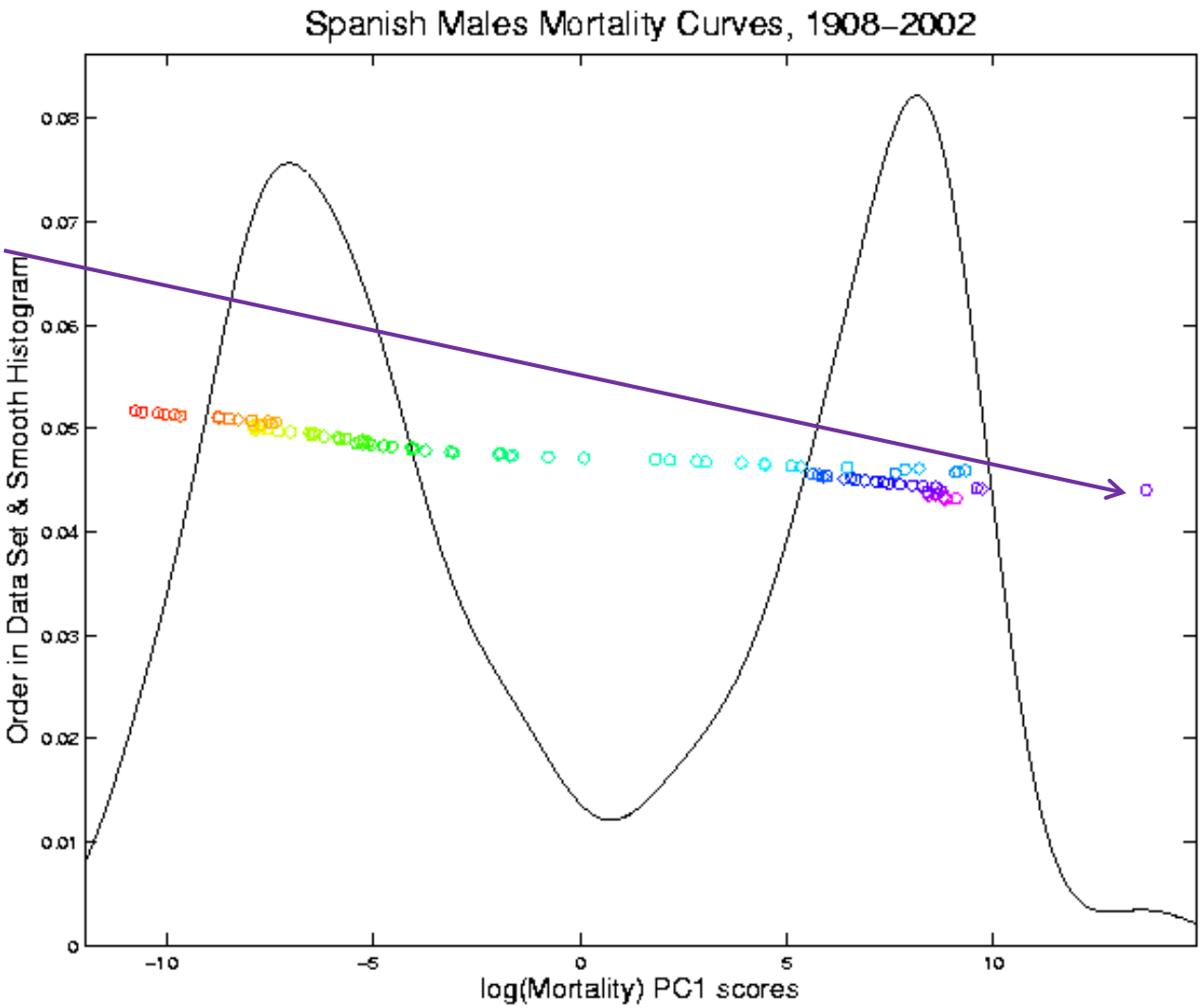
High Mortality Early

Lower Later



Mortality Time Series

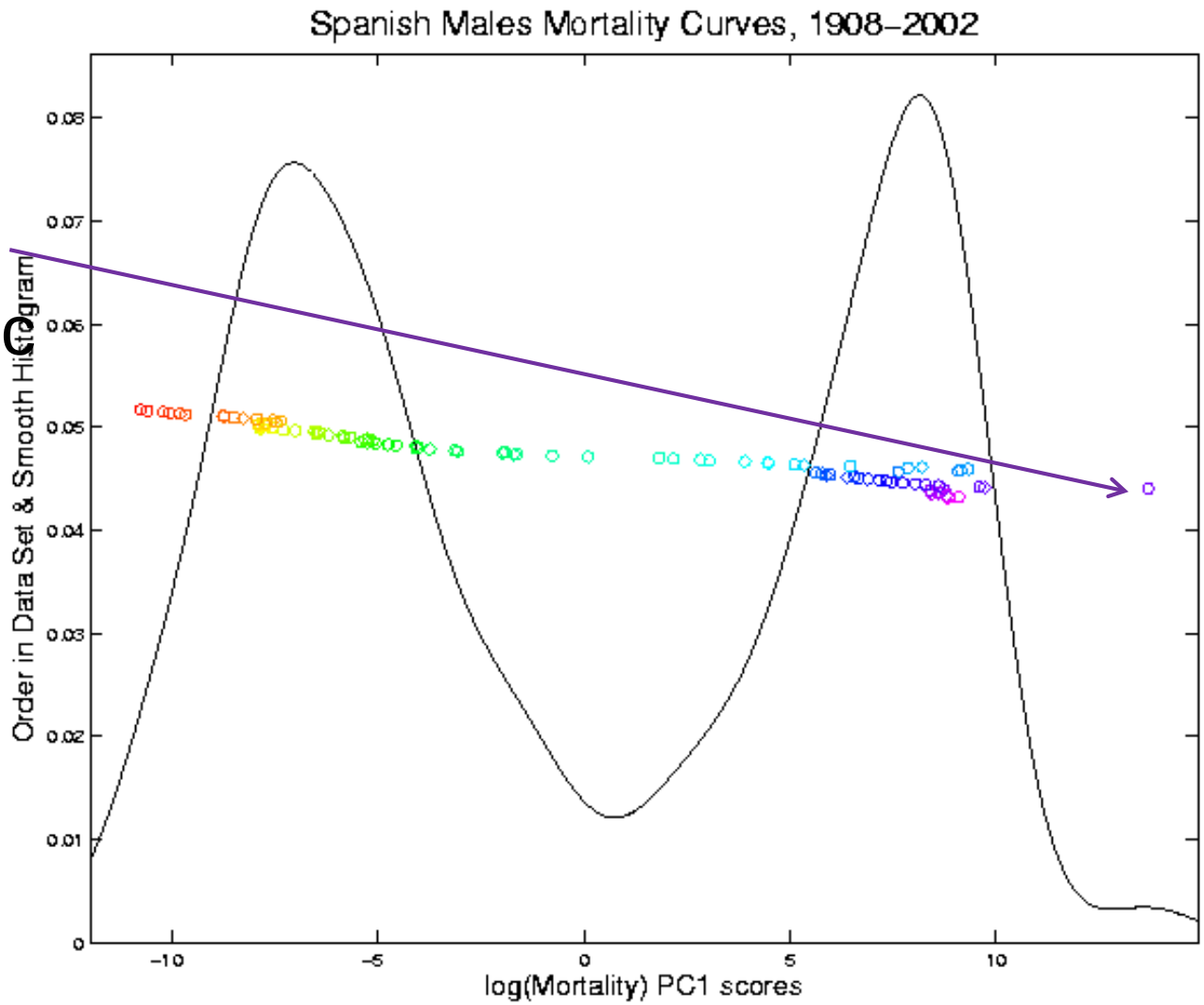
Outliers



Mortality Time Series

Outliers

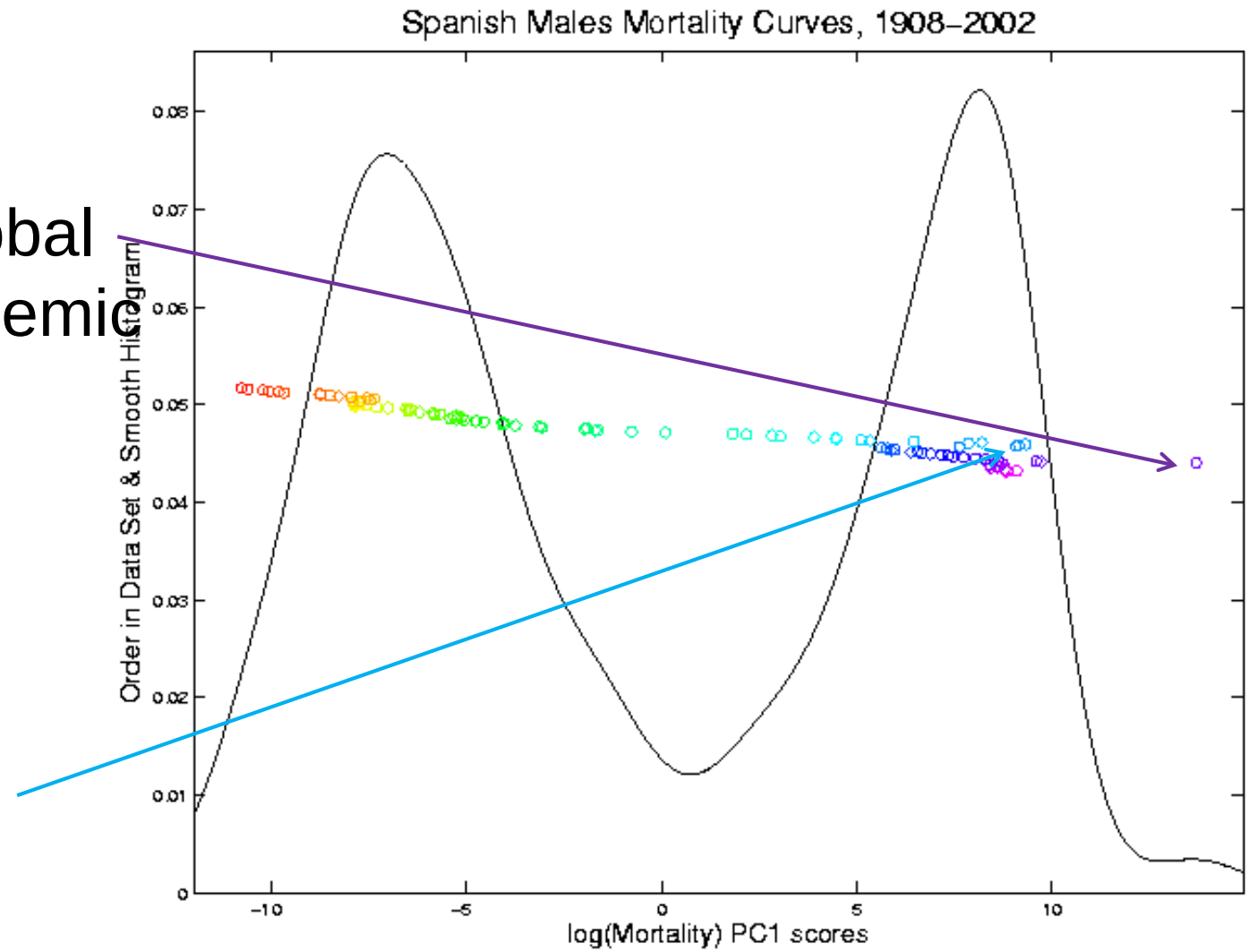
1918 Global Flu Pandemic



Mortality Time Series

Outliers

1918 Global Flu Pandemic

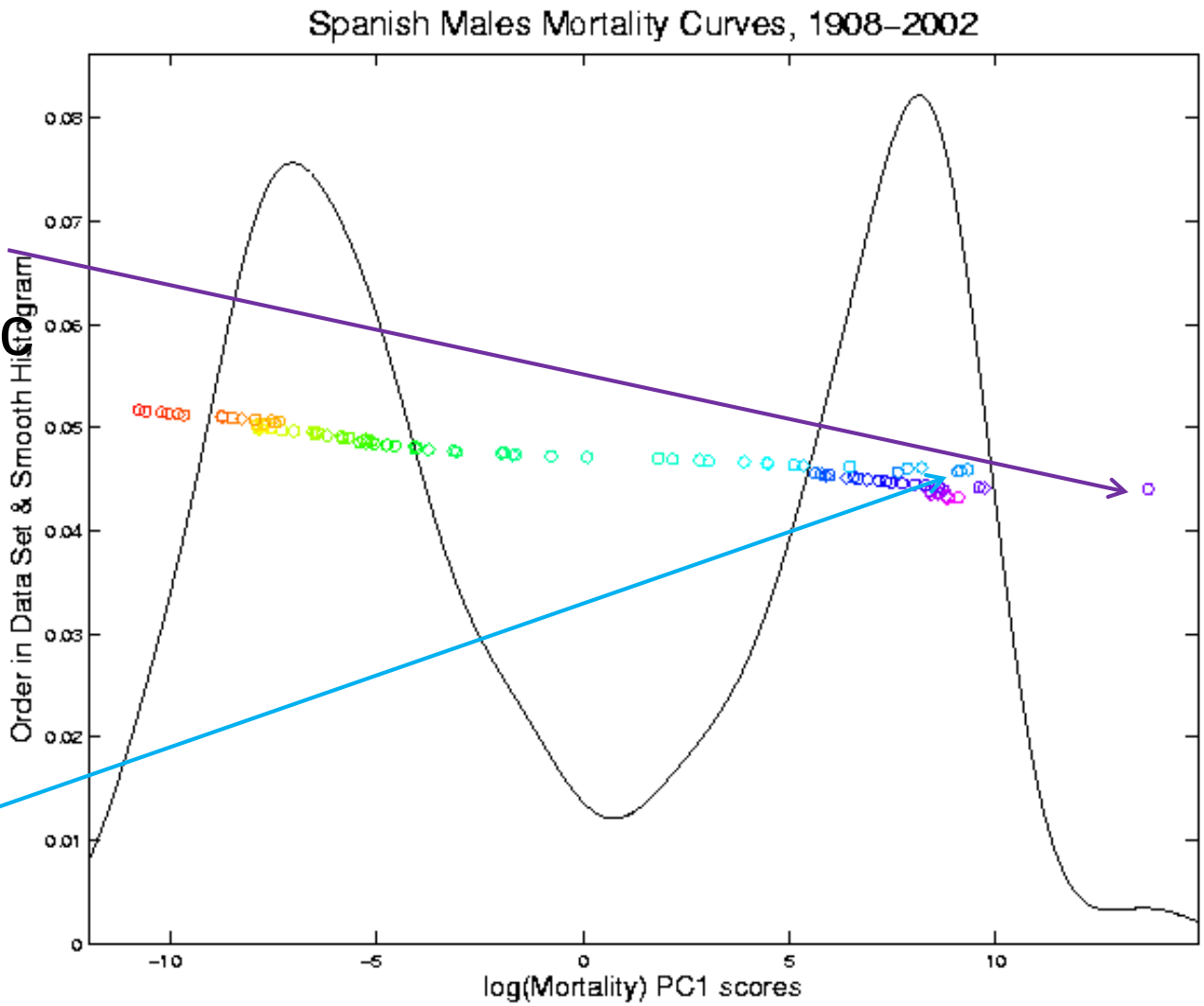


Mortality Time Series

Outliers

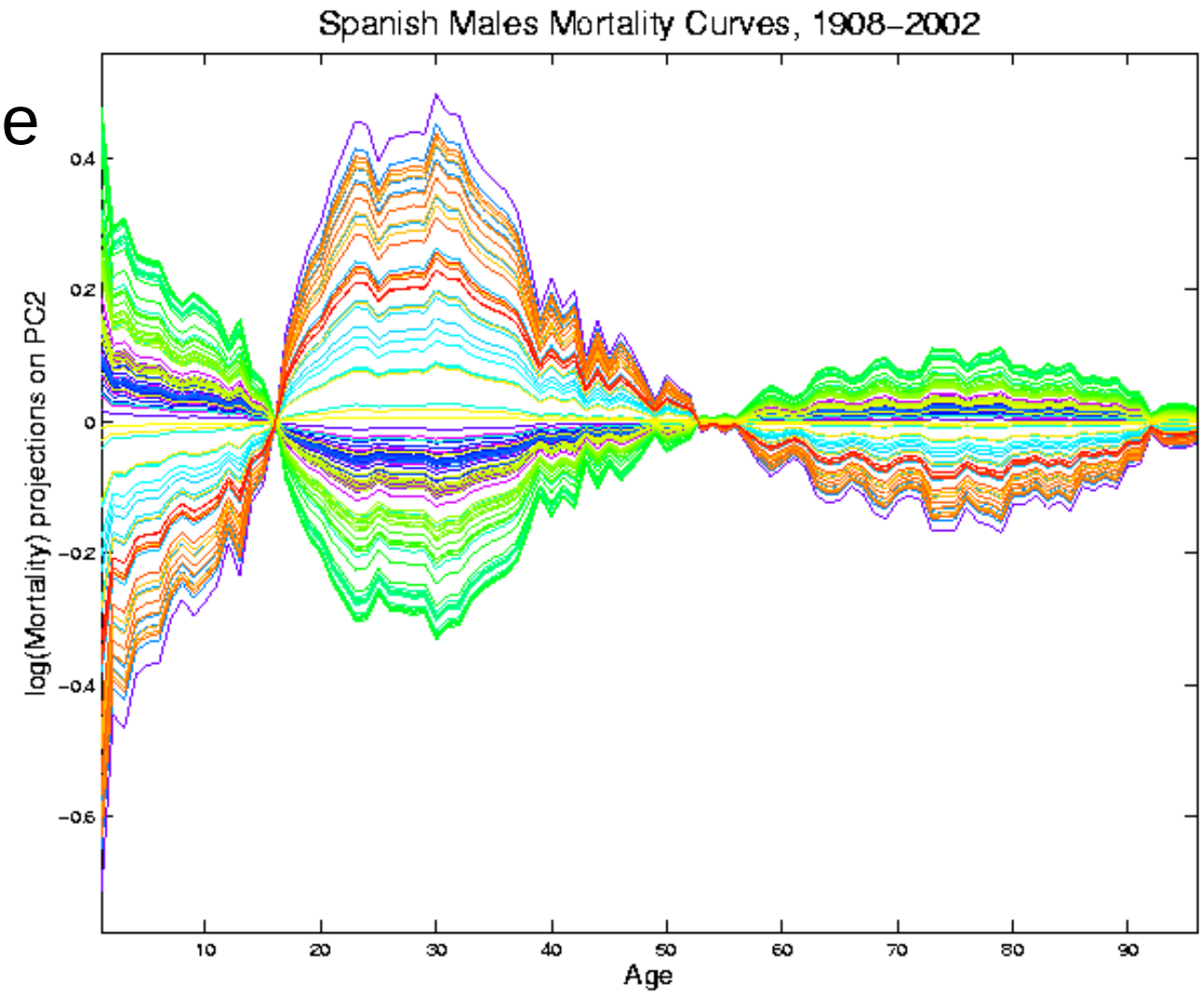
1918 Global Flu Pandemic

1936-1939 Spanish Civil War



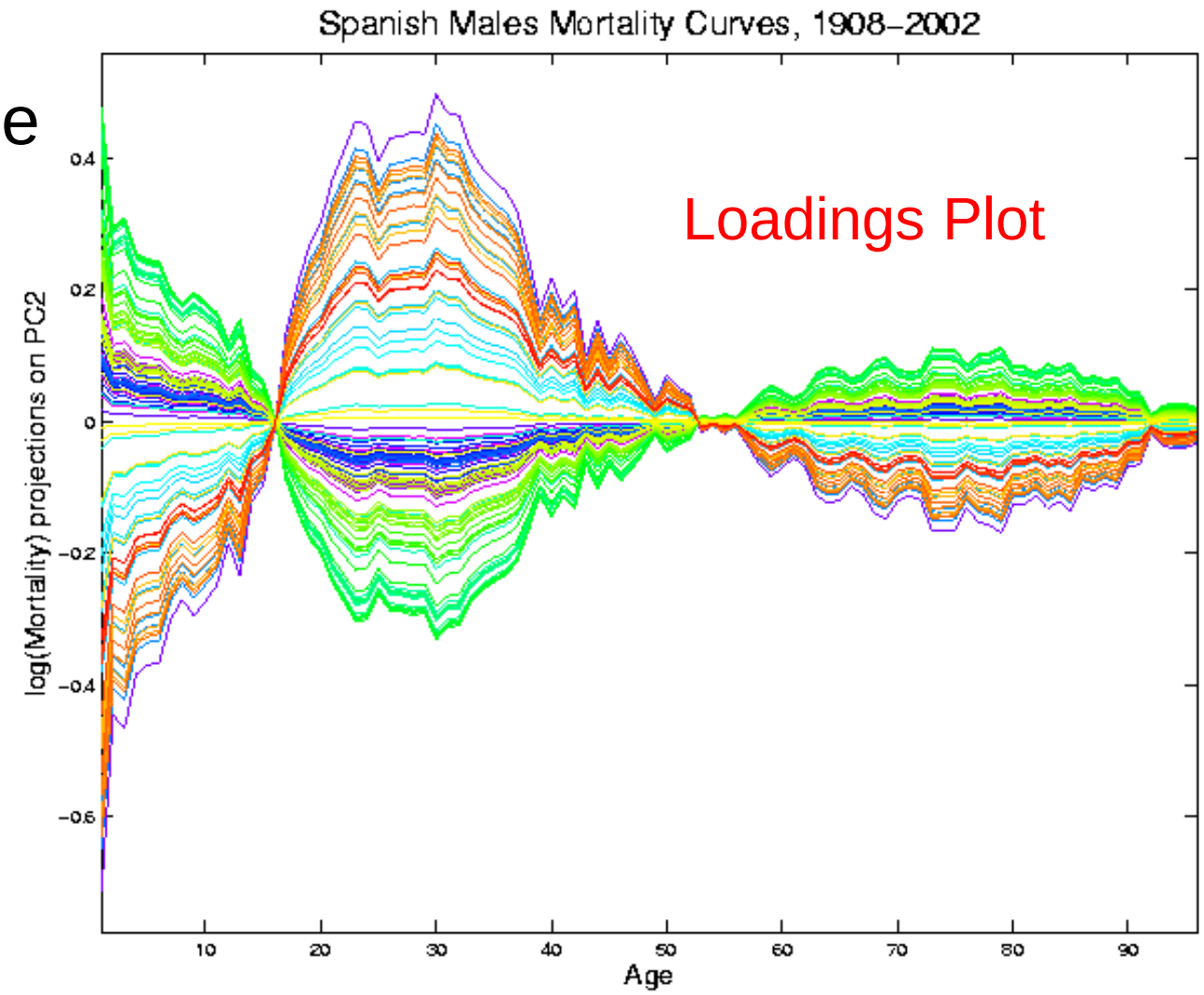
Mortality Time Series

Object Space View of Projections Onto PC2 Direction



Mortality Time Series

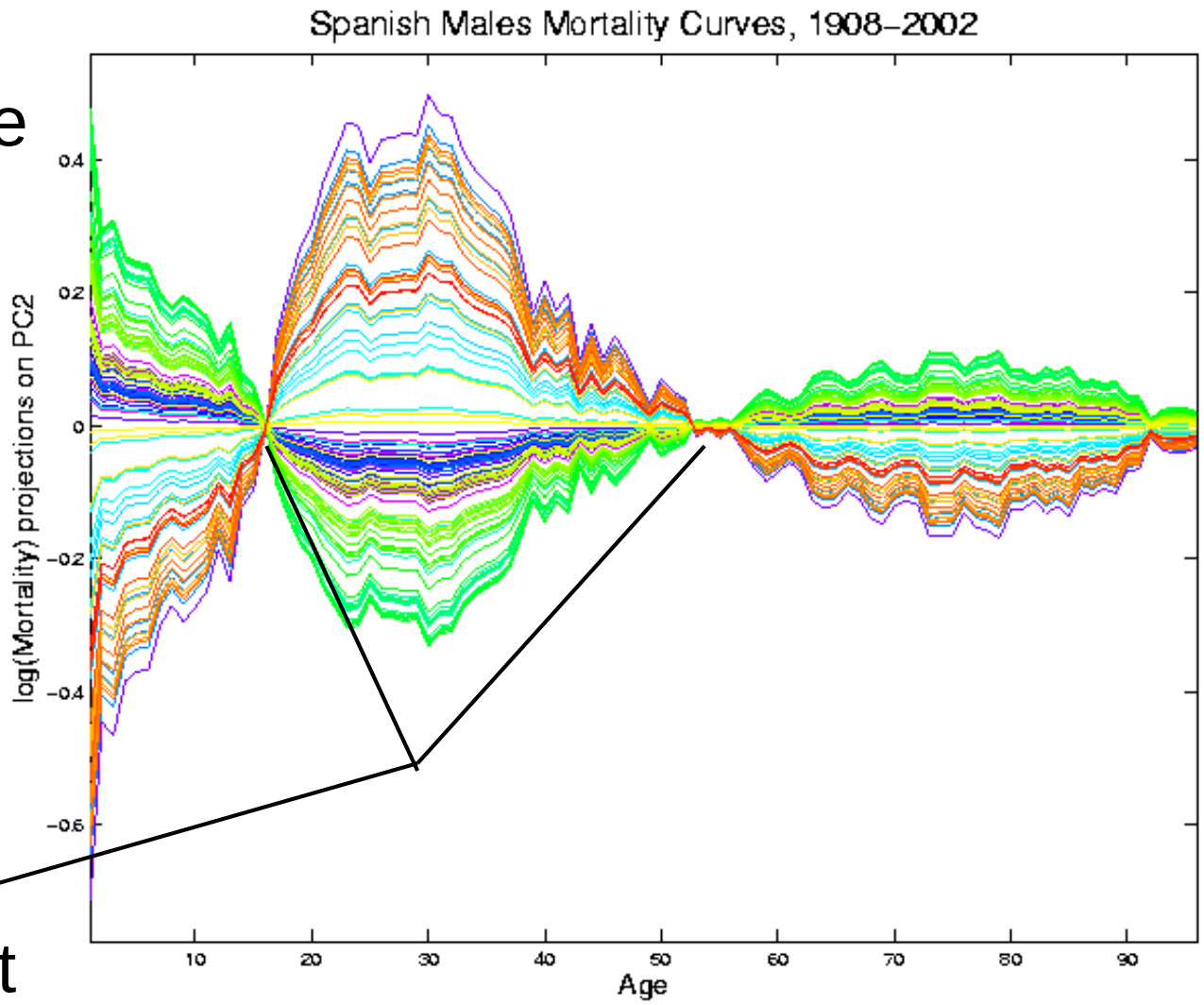
Object Space View of Projections Onto PC2 Direction



Mortality Time Series

Object Space View of Projections Onto PC2 Direction

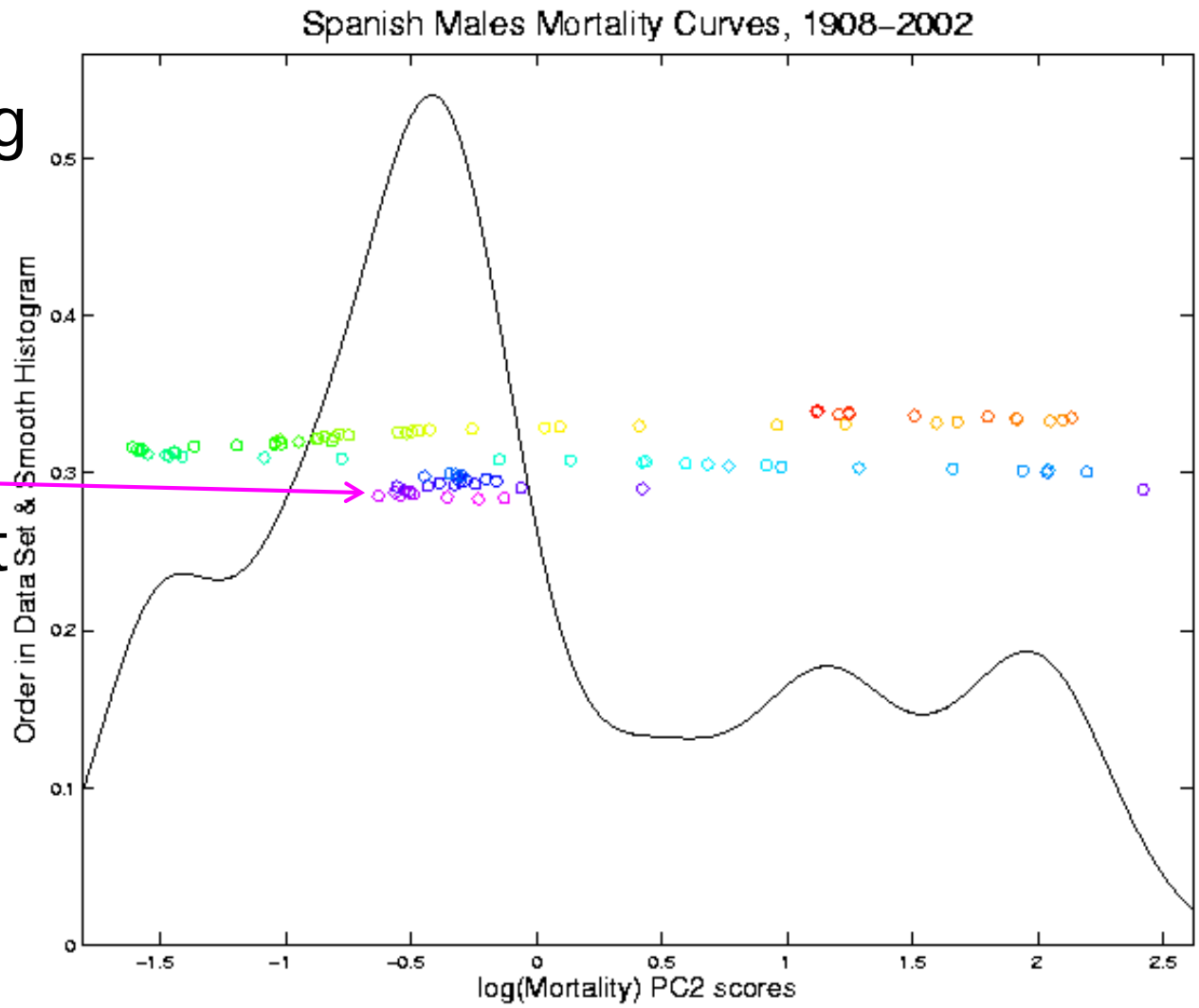
2nd Mode Of Variation: Difference Between 20-45 & Rest



Mortality Time Series

Explain Using PC 2 Scores

Early Improvement

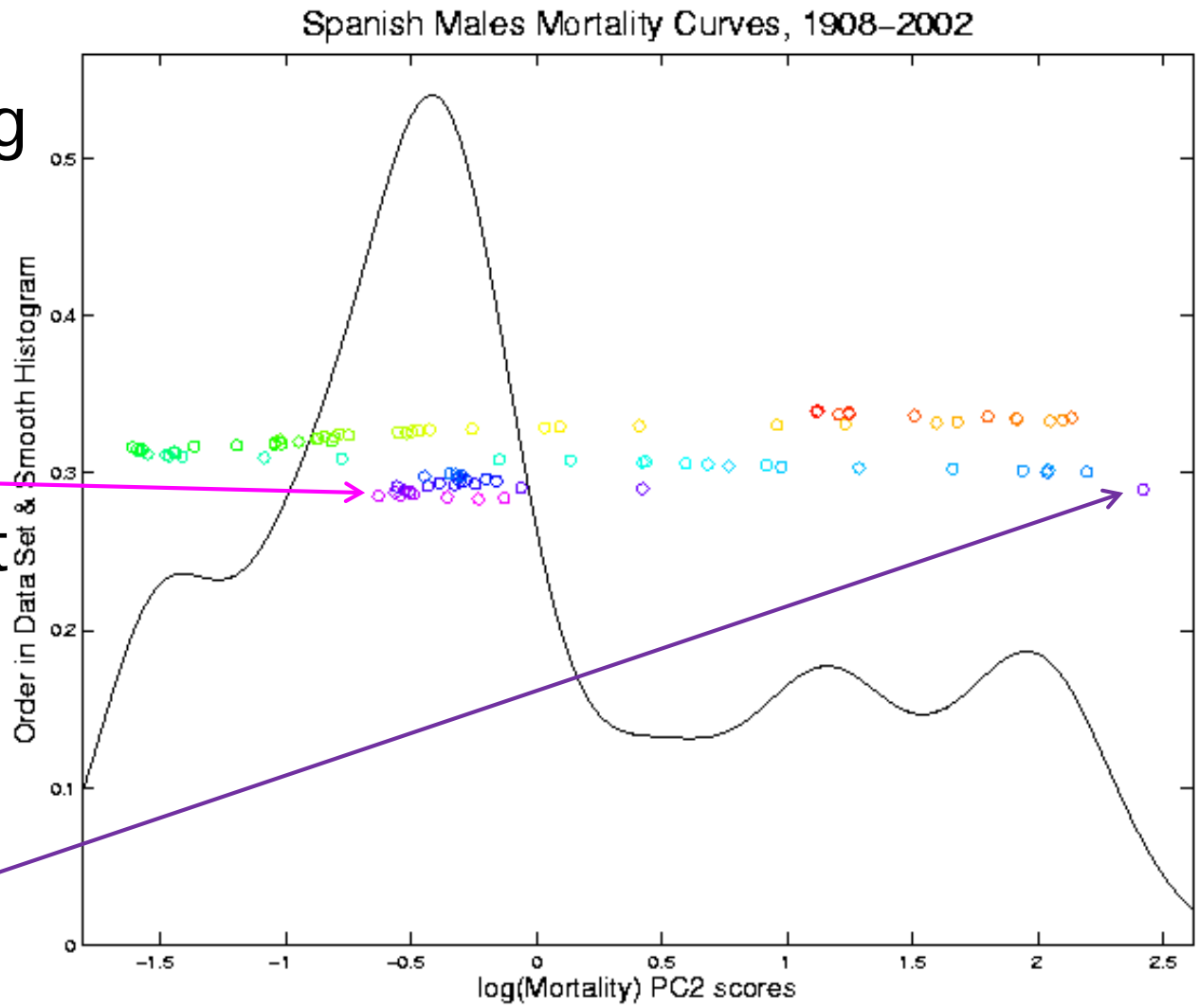


Mortality Time Series

Explain Using PC 2 Scores

Early Improvement

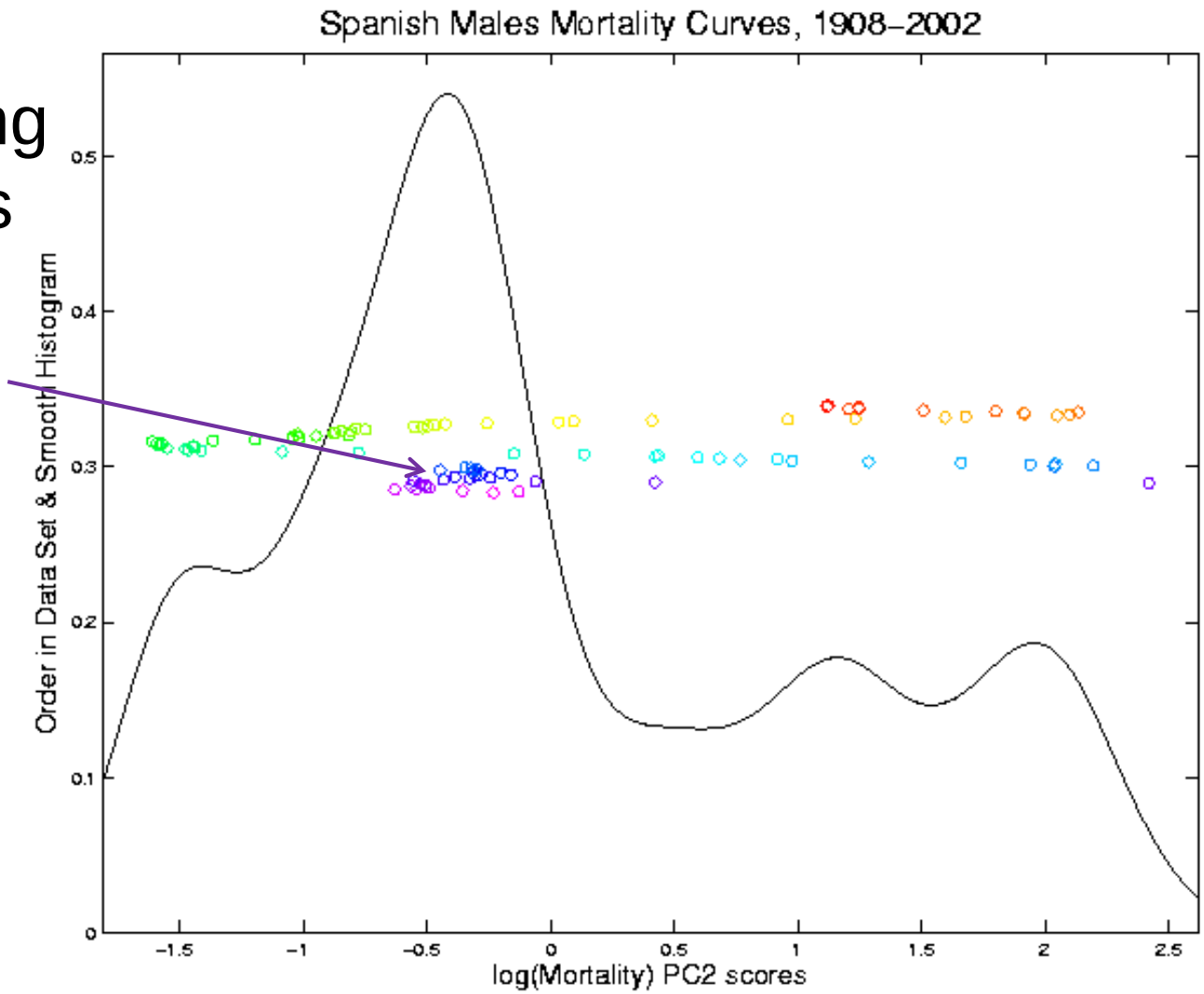
Pandemic Hit Hardest



Mortality Time Series

Explain Using PC 2 Scores

Then better

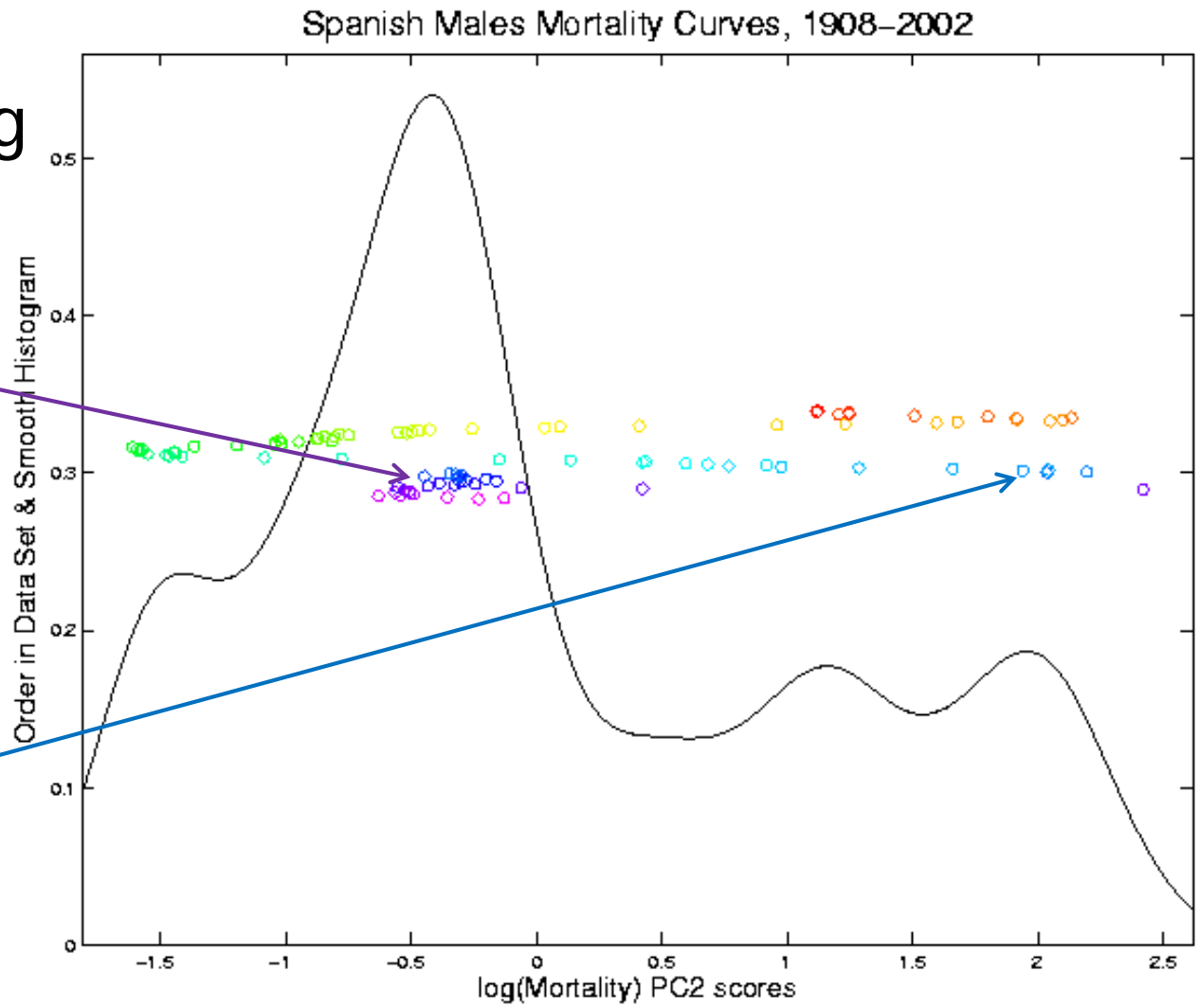


Mortality Time Series

Explain Using PC 2 Scores

Then better

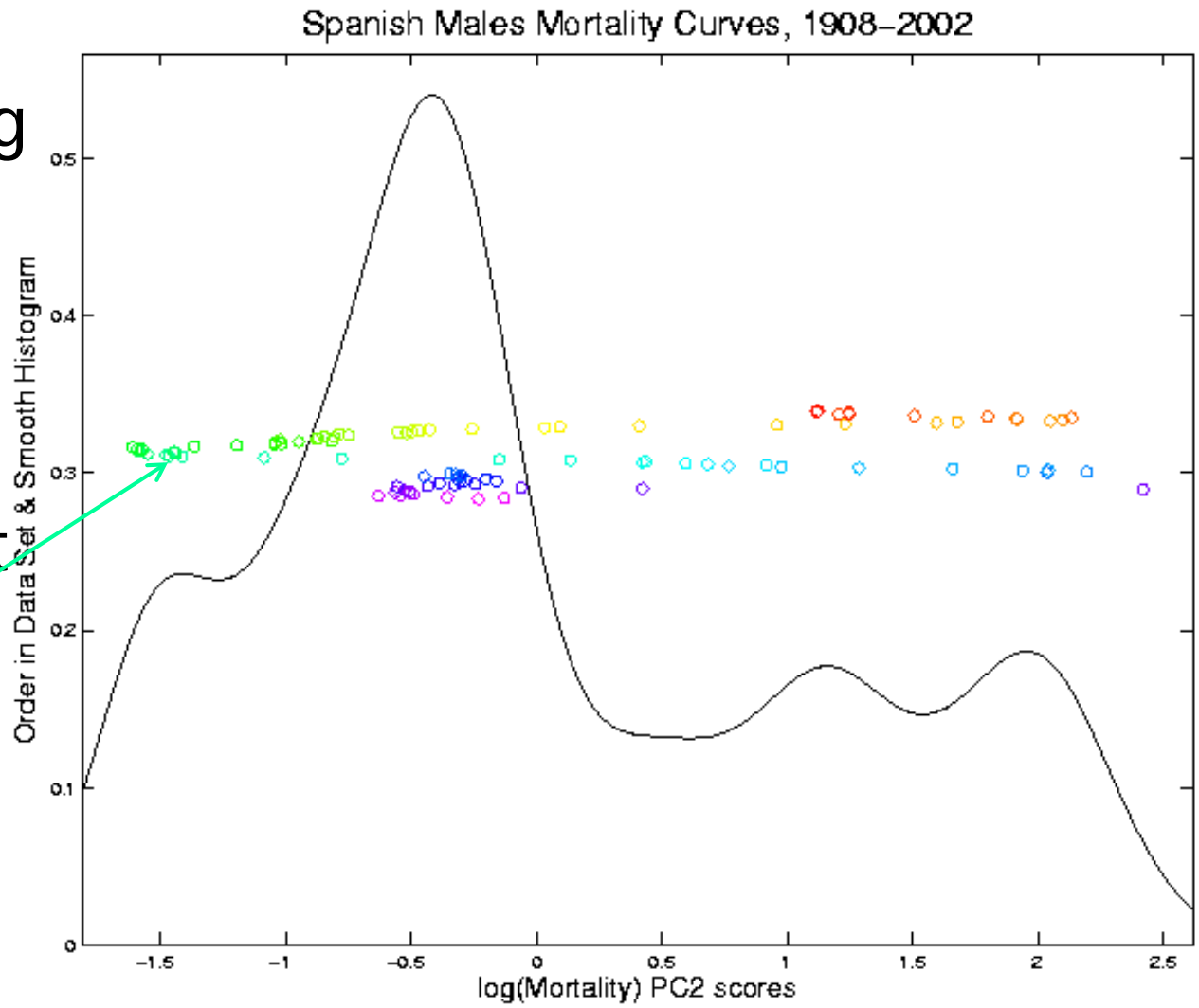
Spanish Civil War Hit Hardest



Mortality Time Series

Explain Using PC 2 Scores

Steady Improvement To mid-50s

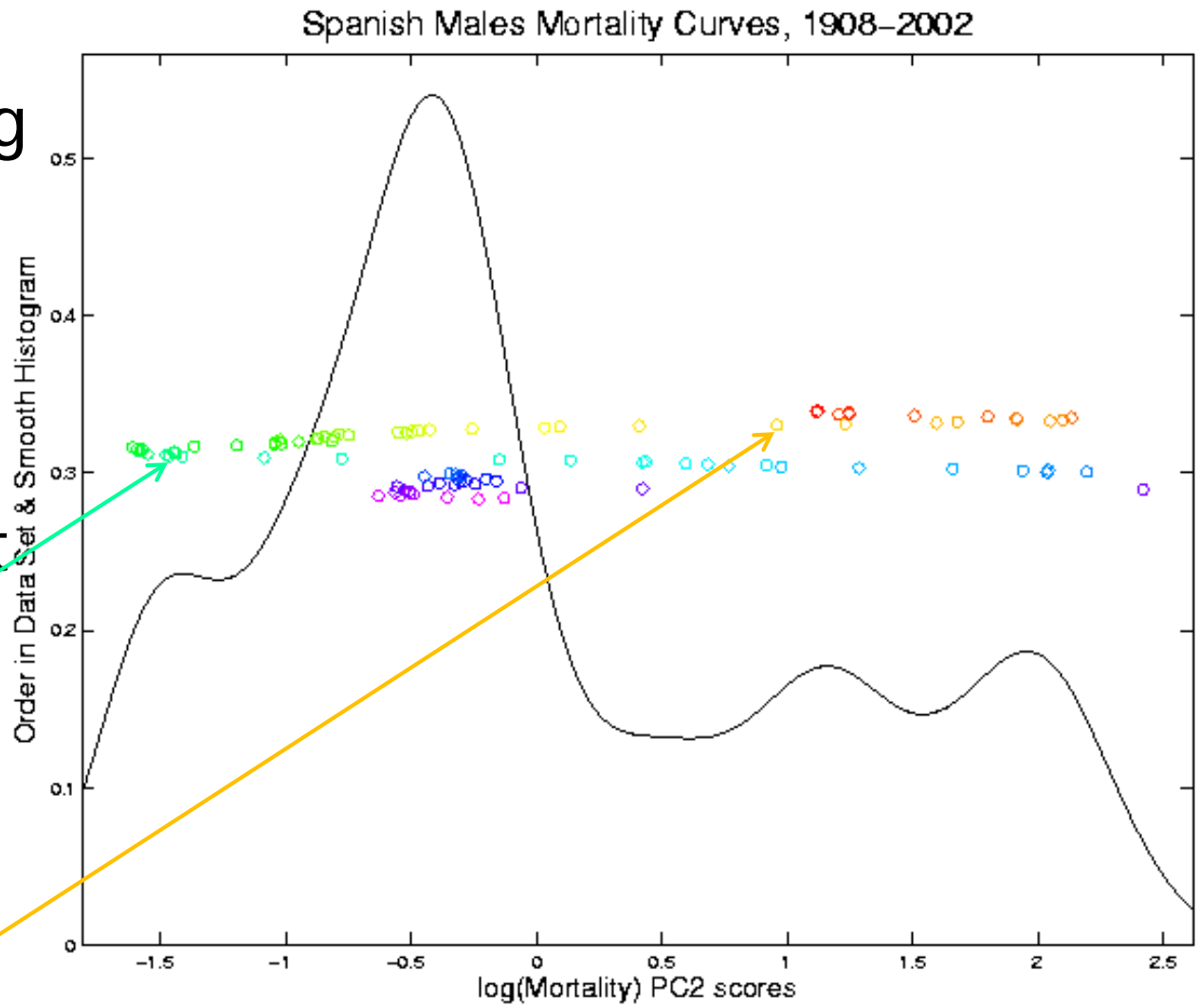


Mortality Time Series

Explain Using PC 2 Scores

Steady Improvement To mid-50s

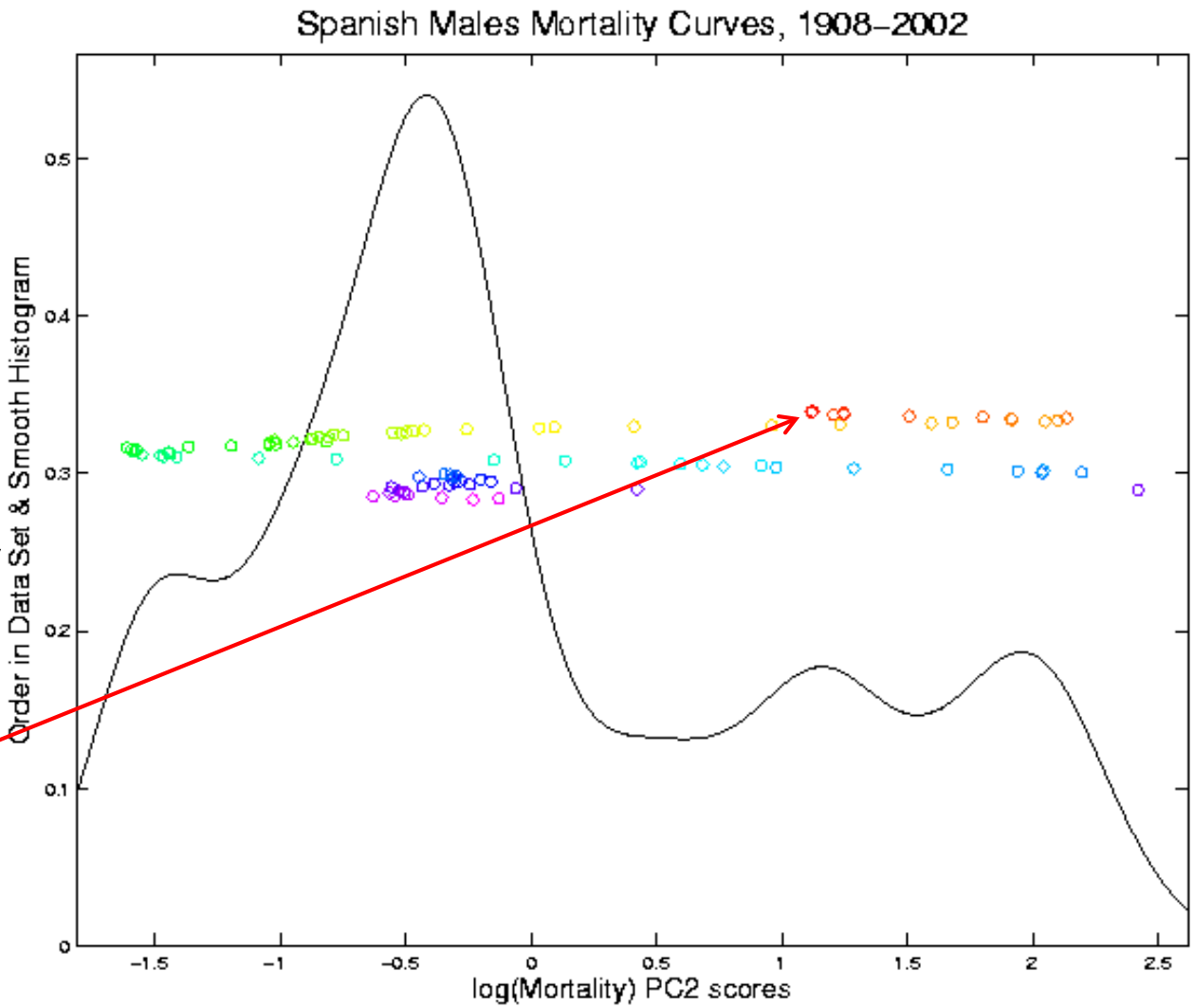
Increasing Automotive Death Rate



Mortality Time Series

Explain Using PC 2 Scores

Corner Finally Turned by Safer Cars & Roads



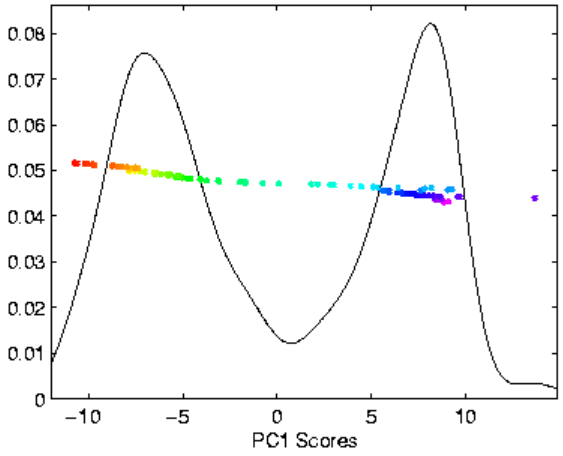
Scores Plot

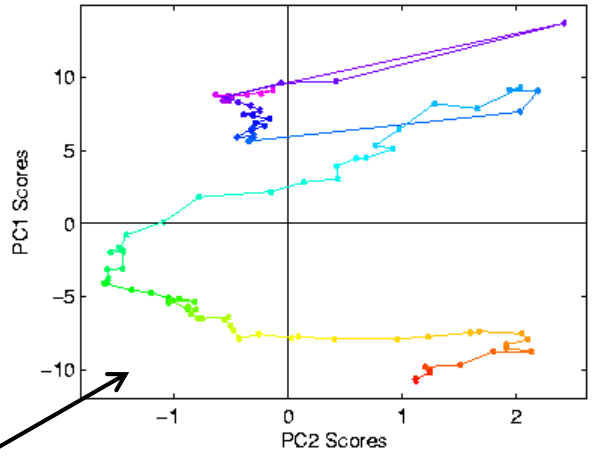
Mortality Time Series

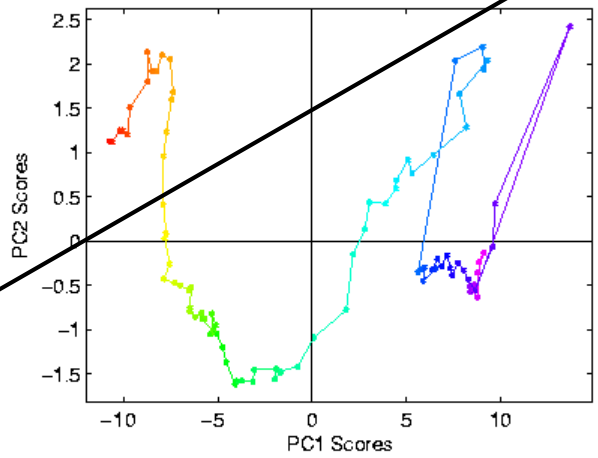
Descriptor (Point Cloud) Space View

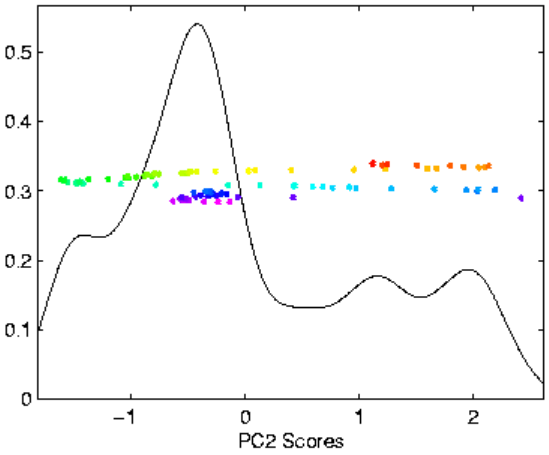
Connecting Lines Highlight Time Order

Good View of Historical Effects









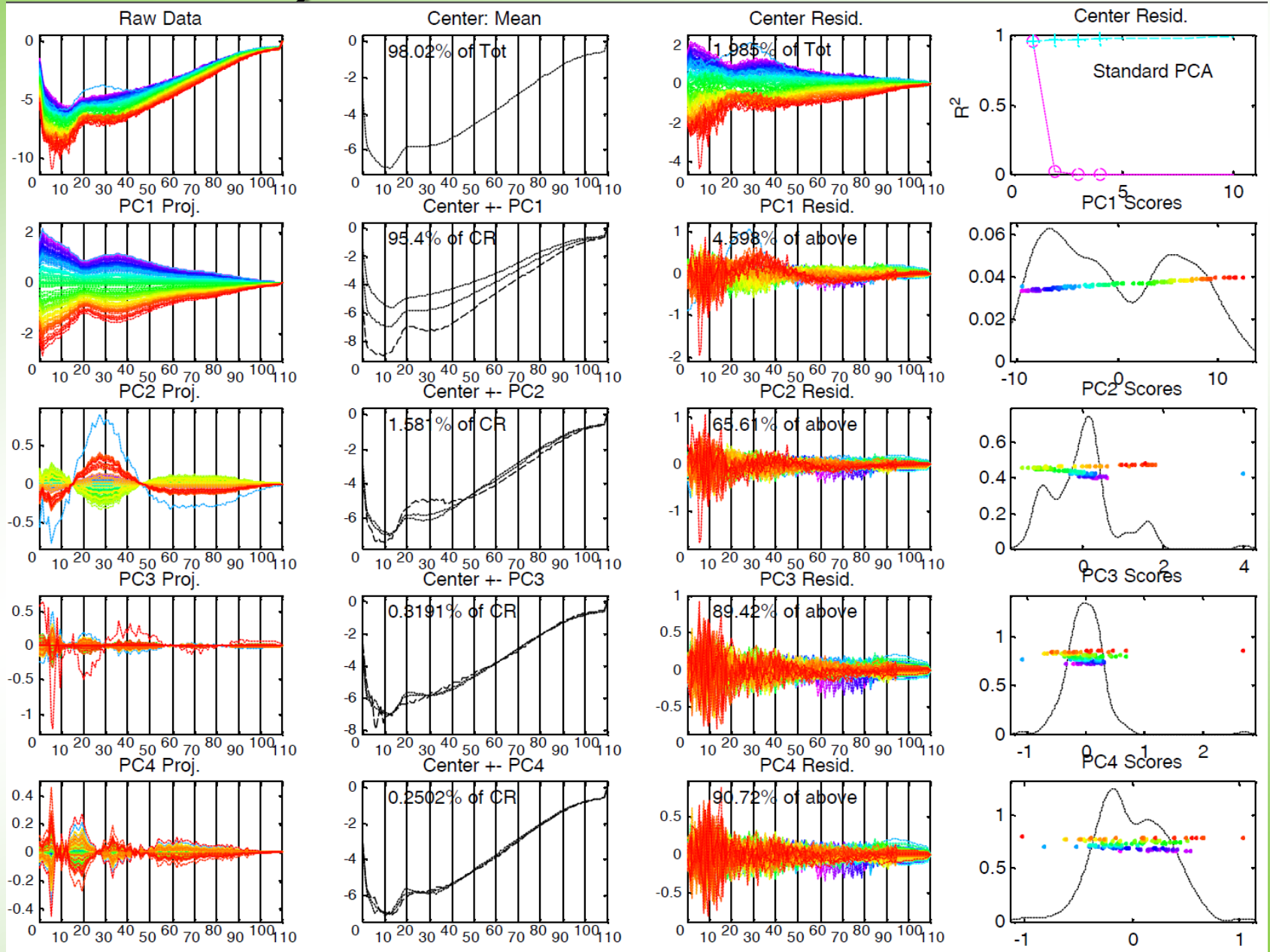
Mortality Time Series

Try a Related Mortality Data Set:

Switzerland

(In Europe, but different history)

Mortality Time Series – Swiss Males



Mortality Time Series – Swiss Males

Some Points Similar to Spain:

- PC1: Overall Improvement
- Better for Young
- PC2: About 20 – 45 vs. Others
- Flu Pandemic
- Automobile Effects

Some Quite Different:

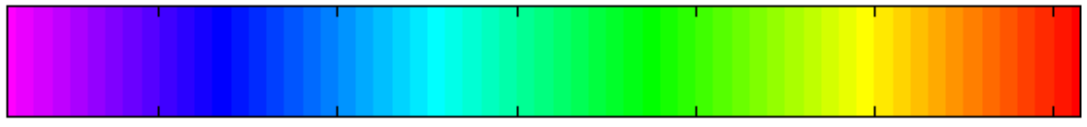
- No Age Rounding
- No Civil War

Time Series of Curves

- Just a “Set of Curves”
- But Time Order is Important!
- Useful Approach (as above):

Use *color* to code for *time*

Start



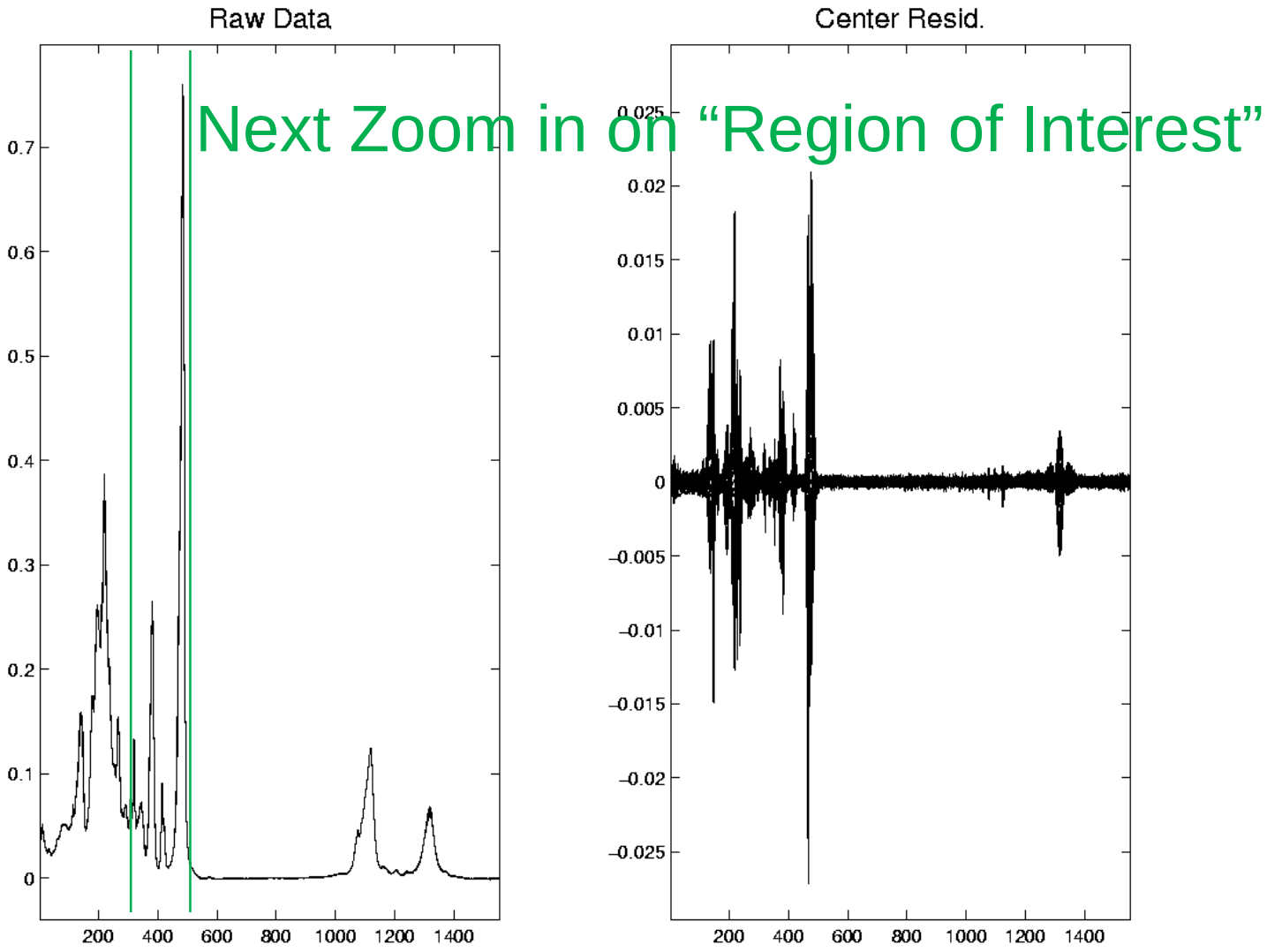
End

Chemo-metric Time Series, HA 27

Raw Data:

- All 60 spectra essentially the same
- “Scale” of mean is much bigger than variation about mean
- Hard to see structure of all 1600 freq’s

Chemo-metric Time Series, HA 27

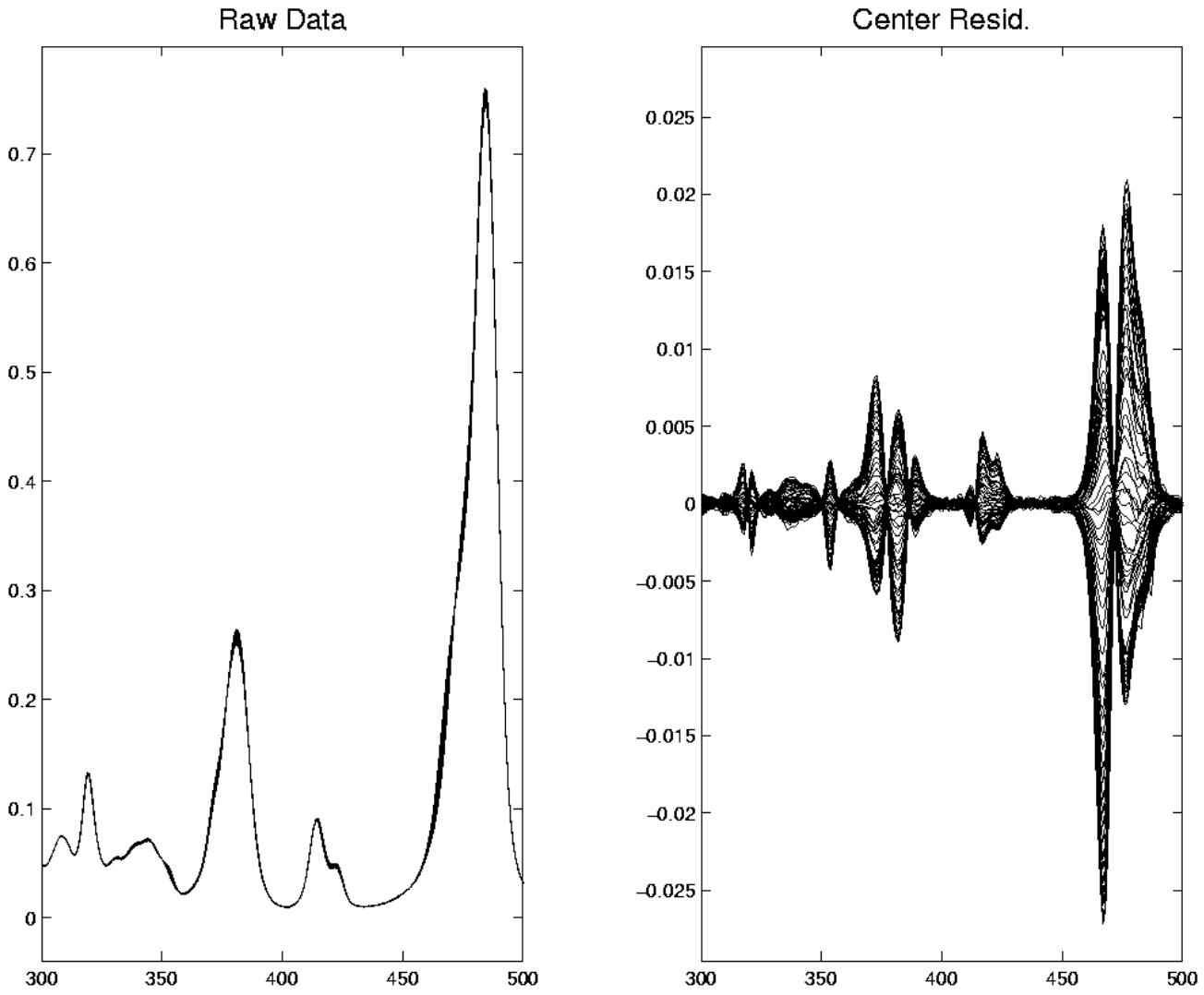


Chemo-metric Time Series, HA 27

Centered Data:

- Now can see different spectra
- Since mean subtracted off
- Note much smaller vertical axis

Chemo-metric Time Series, HA 27



Chemo-metric Time Series, HA 27

Data zoomed to “important” freq’s:

Raw Data:

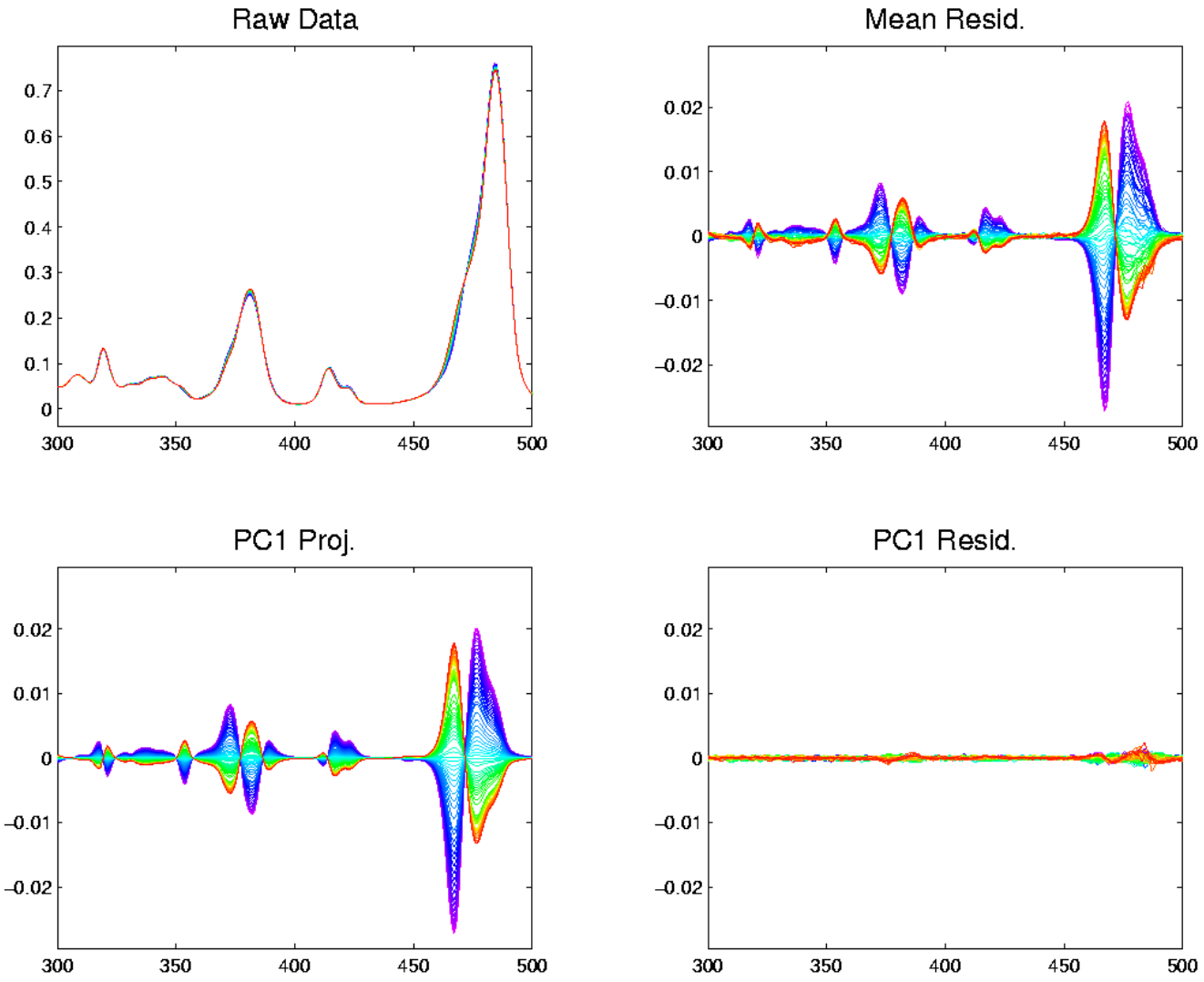
- Now see slight differences
- Smoother “natural looking” spectra

Centered Data:

- Differences in spectra more clear
- Maybe now have “real structure”

Scale is important

Chemo-metric Time Series, HA 27



Chemo-metric Time Series, HA 27

Use of Time Order Coloring:

Raw Data:

- Can see a little ordering, not much

Centered Data:

- Clear time ordering
- Shifting peaks? (compare to Raw)

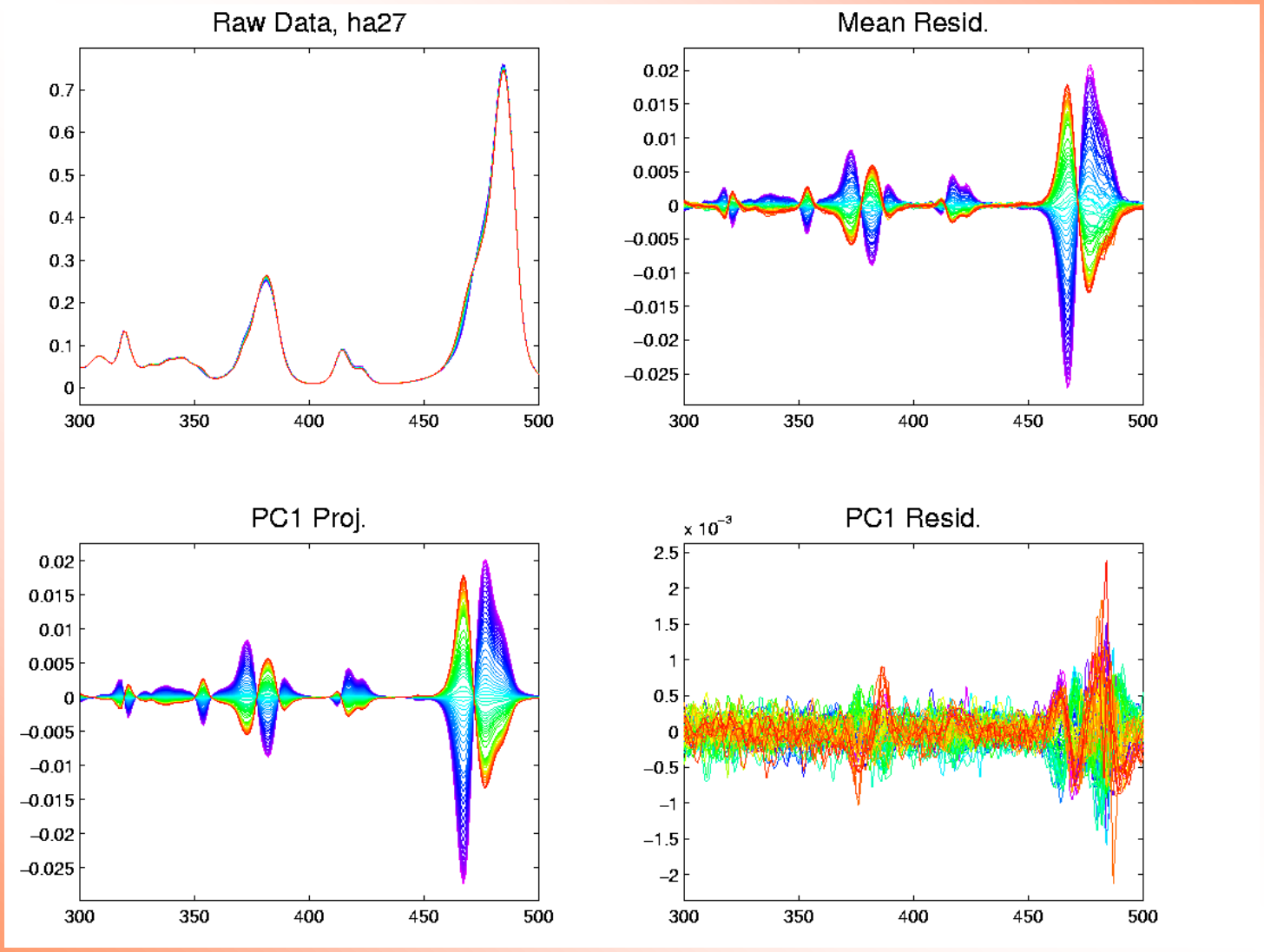
PC1:

- Almost everything?

PC1 Residuals:

- Data nearly linear (same scale import’nt)

Object Space View, HA27



Strong structure in PC1 Resid (d < 2)

Functional Data Analysis

Interesting Real Data Example

- Genetics (Cancer Research)
- RNAseq (Next Gener'n Sequen'g)
- Deep look at "gene components"

Microarrays: Single number (per gene)

RNAseq: Thousands of measurements

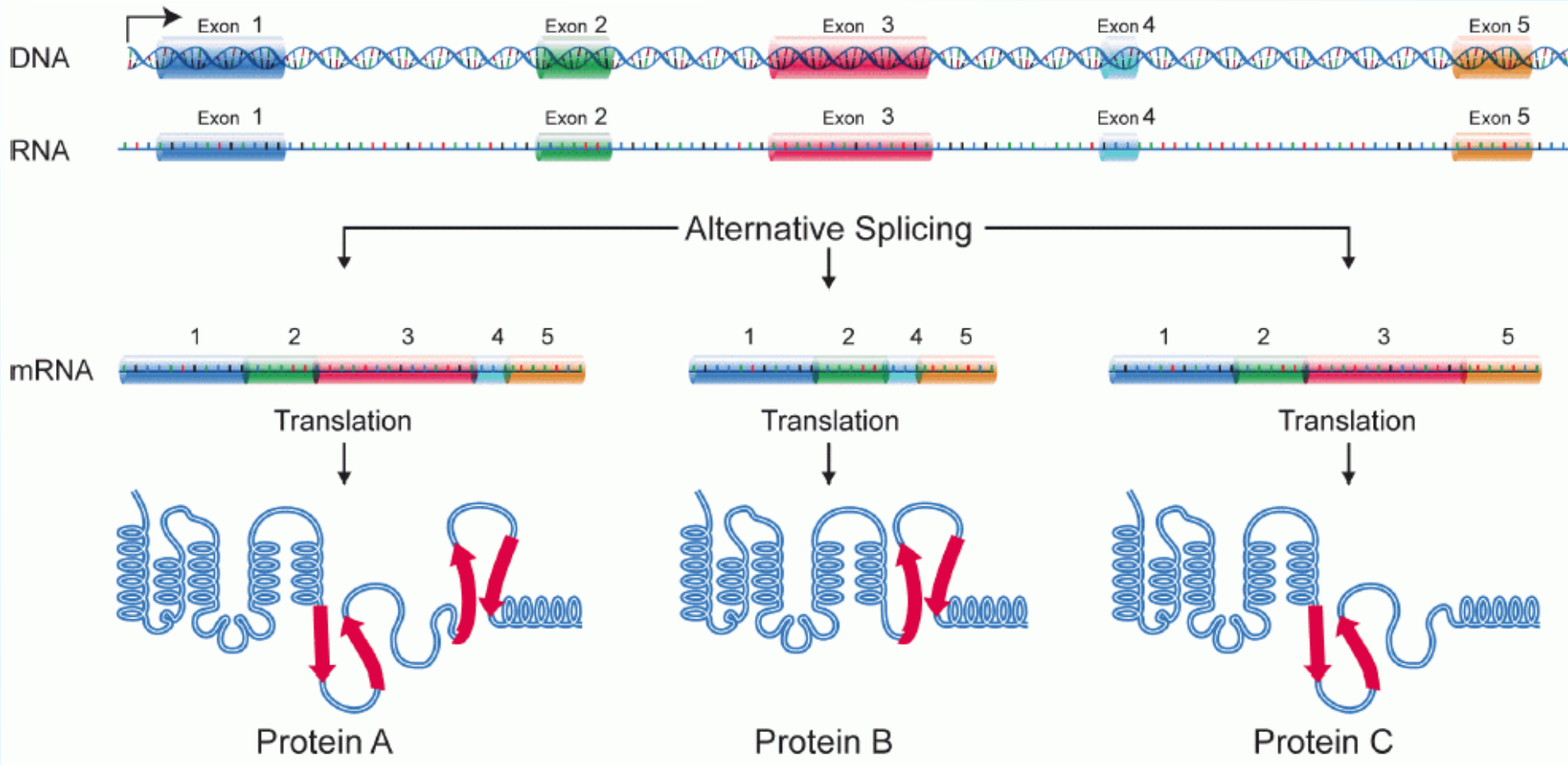
Functional Data Analysis

Interesting Real Data Example

- Genetics (Cancer Research)
- RNAseq (Next Gener'n Sequen'g)
- Deep look at "gene components"

- Gene studied here: CDNK2A
- Goal: *Study Alternate Splicing*
- Sample Size, n = 180
- Dimension, d = ~1700

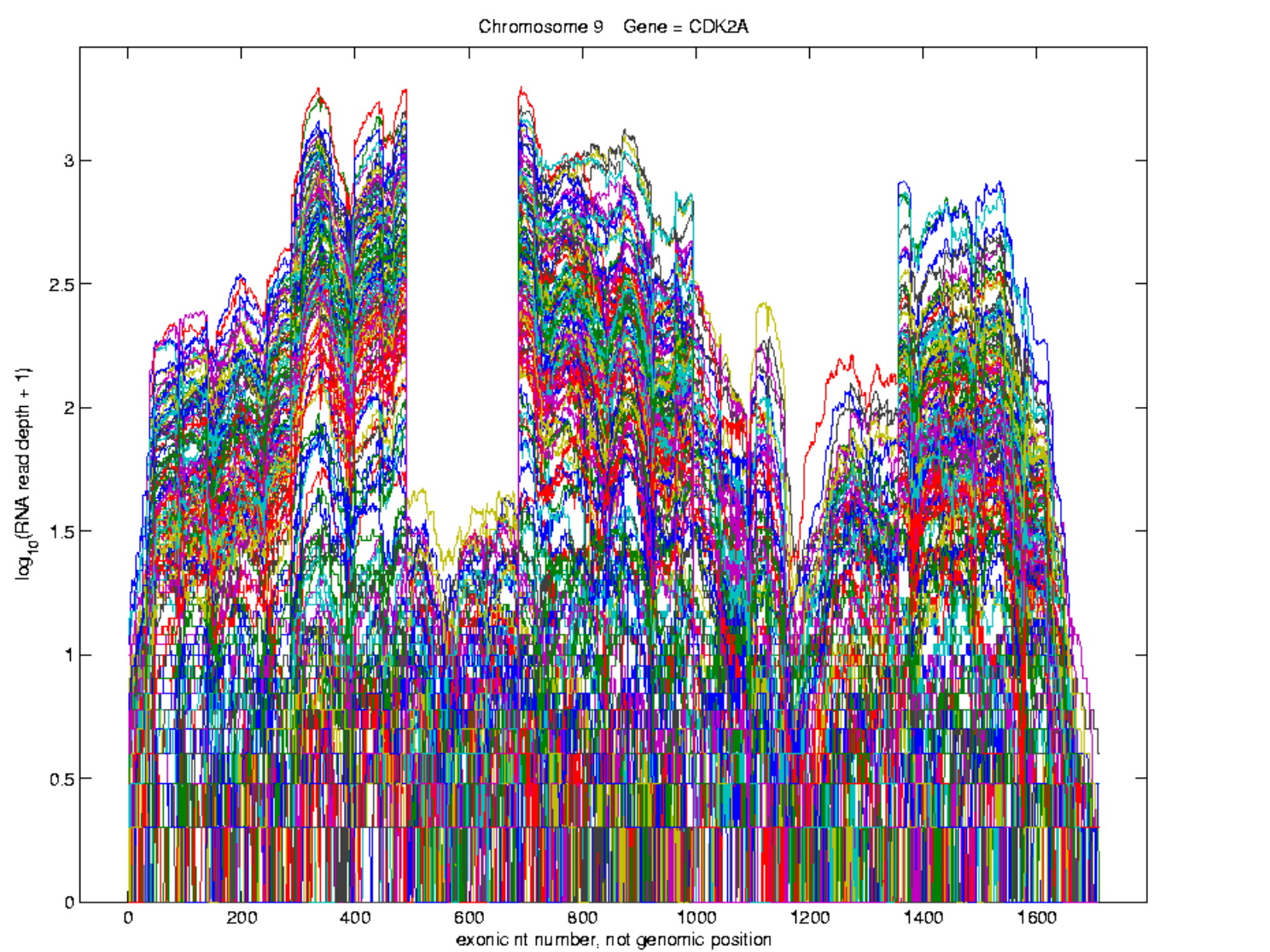
Alternate Splicing



Functional Data Analysis

Simple 1st View: Curve Overlay

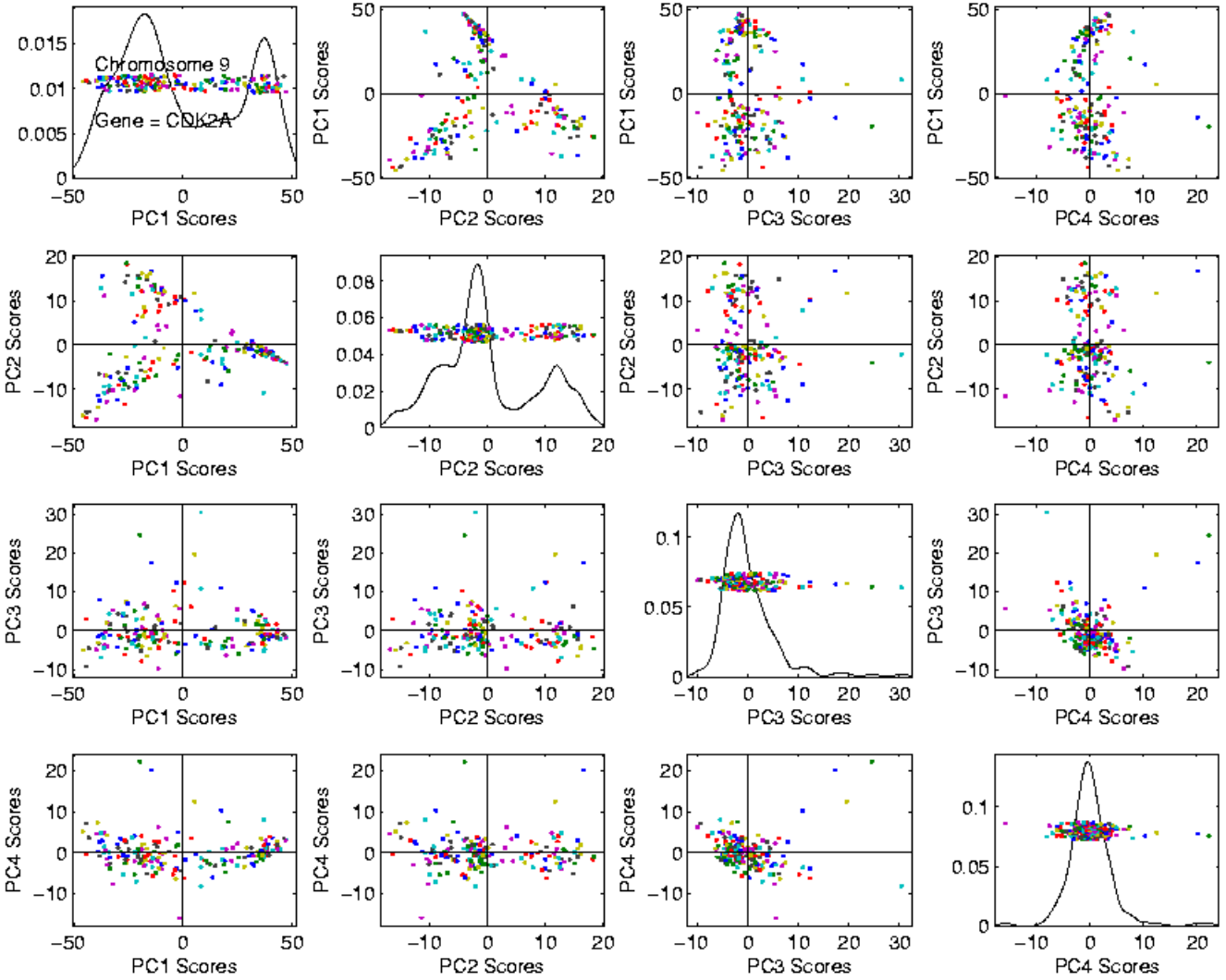
(log scale)



Functional Data Analysis

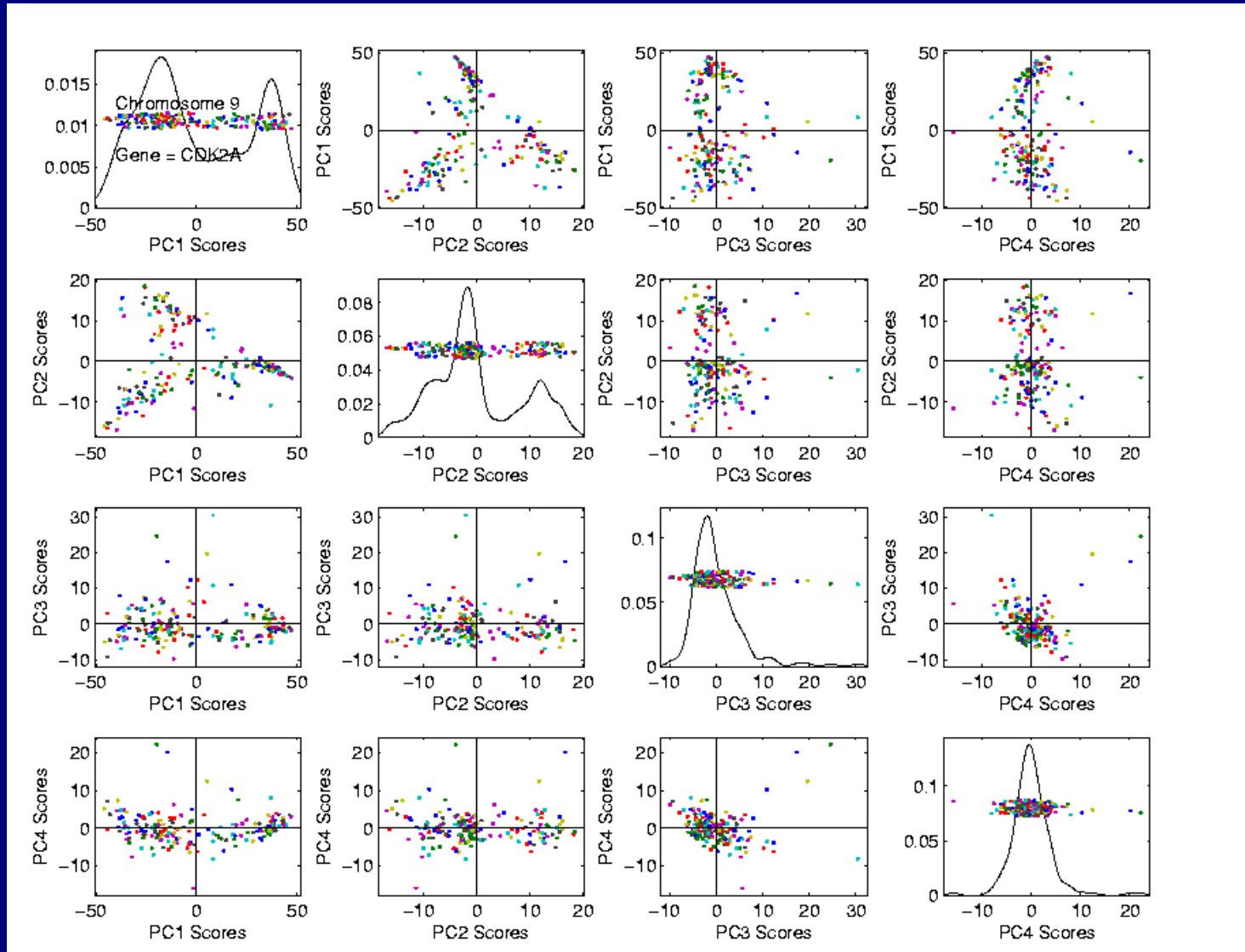
Often Useful Population View:

PCA Scores



Functional Data Analysis

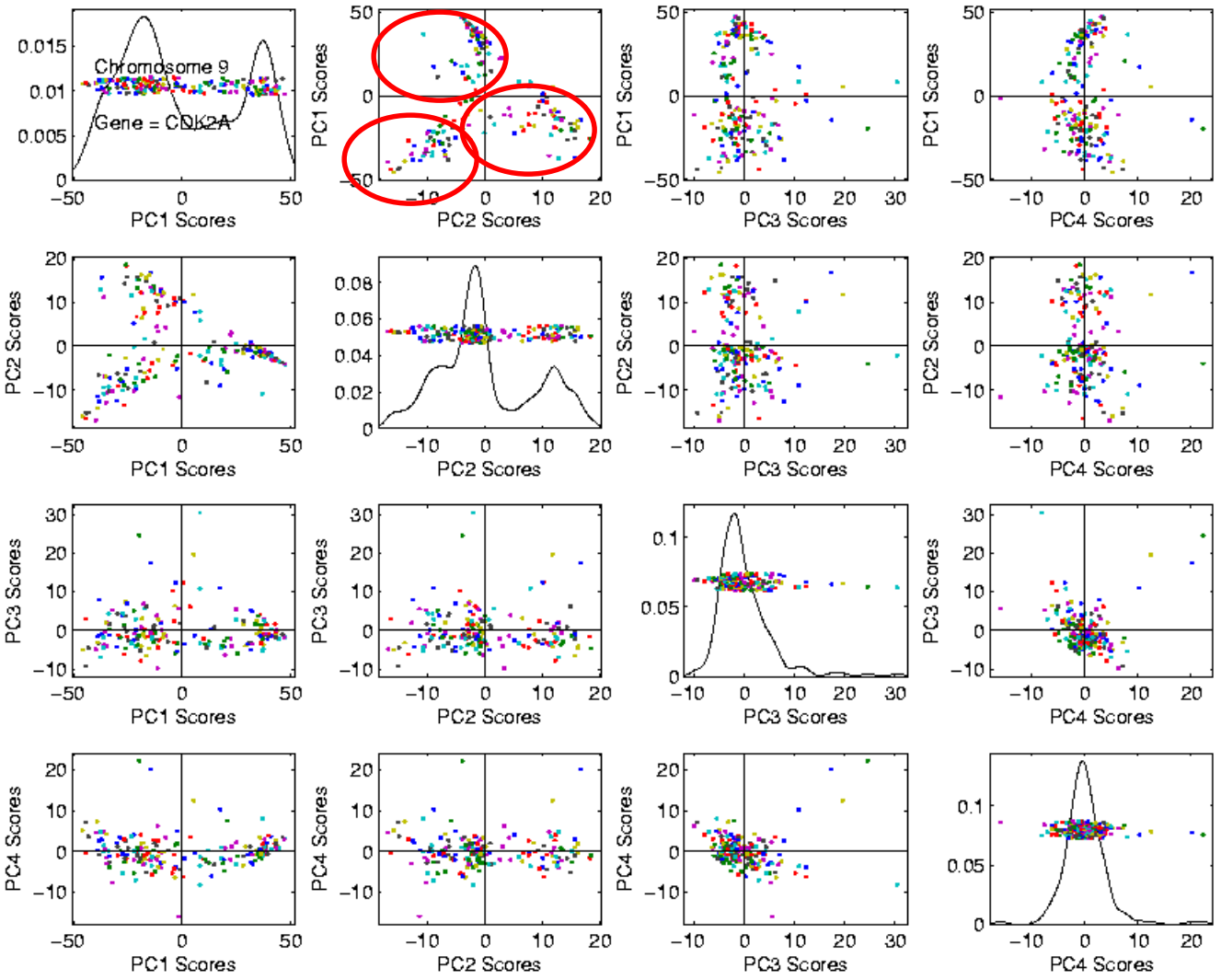
Suggestion Of Clusters?



Functional Data Analysis

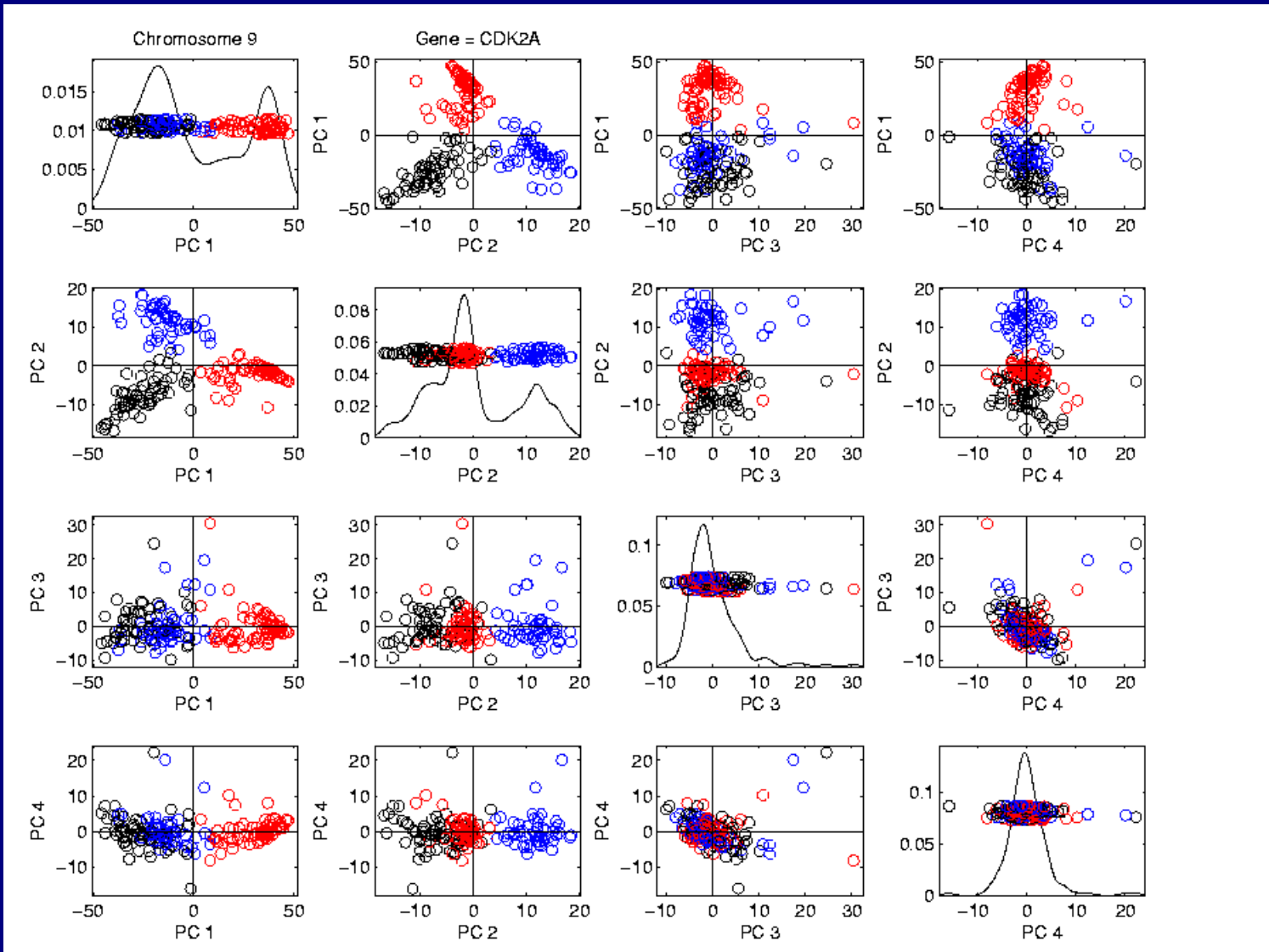
Suggestion Of Clusters

Which Are These?



Functional Data Analysis

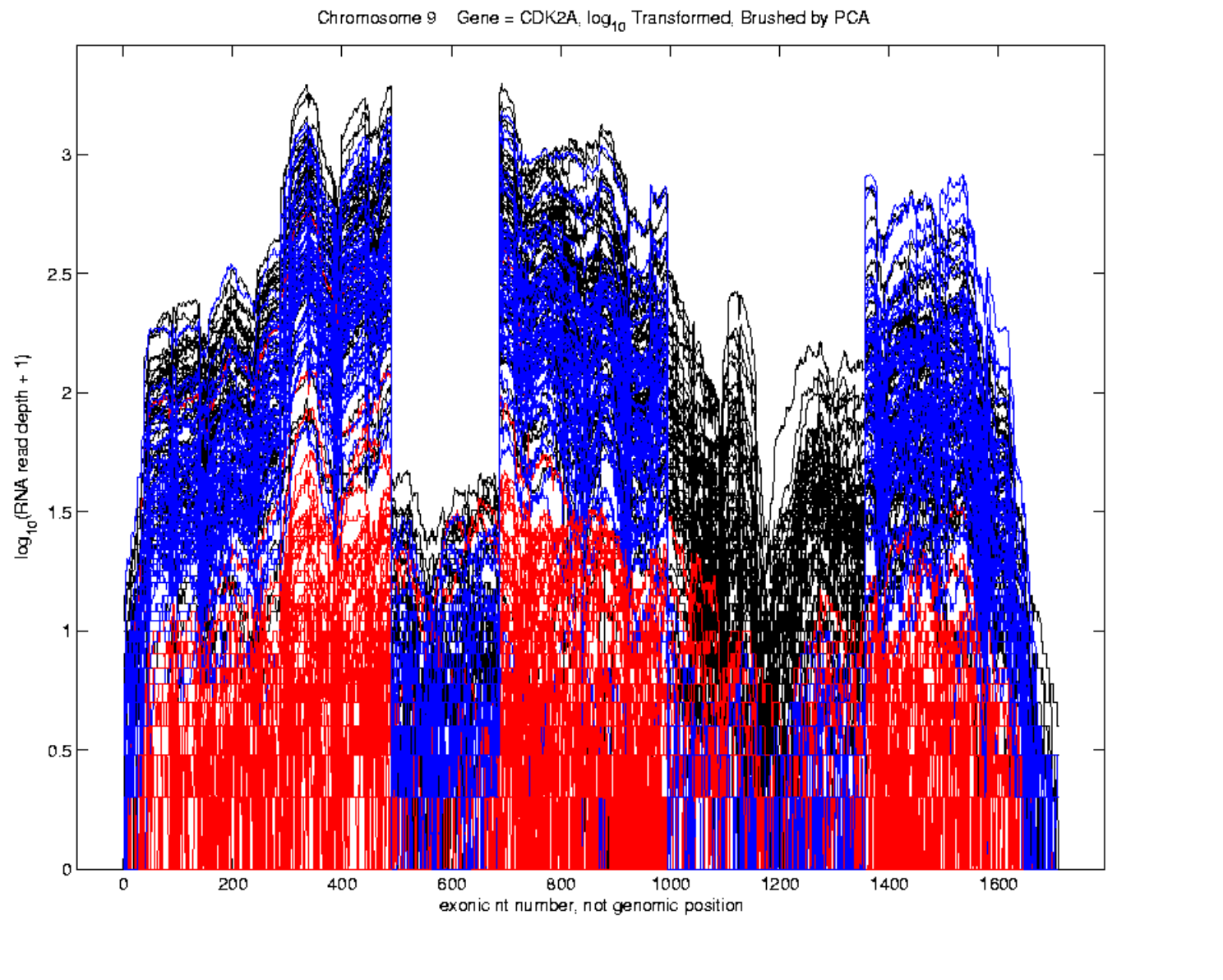
Manually "Brush" Clusters



Functional Data Analysis

Manually Brush Clusters

Clear Alternate Splicing



Functional Data Analysis

Important Points

- ✓ PCA found *Important Structure*
- ✓ In High Dimensional Data Analysis

$$d \sim 1700$$

Limitation of PCA

PCA can provide useful projection directions

But can't "see everything"…

Reason:

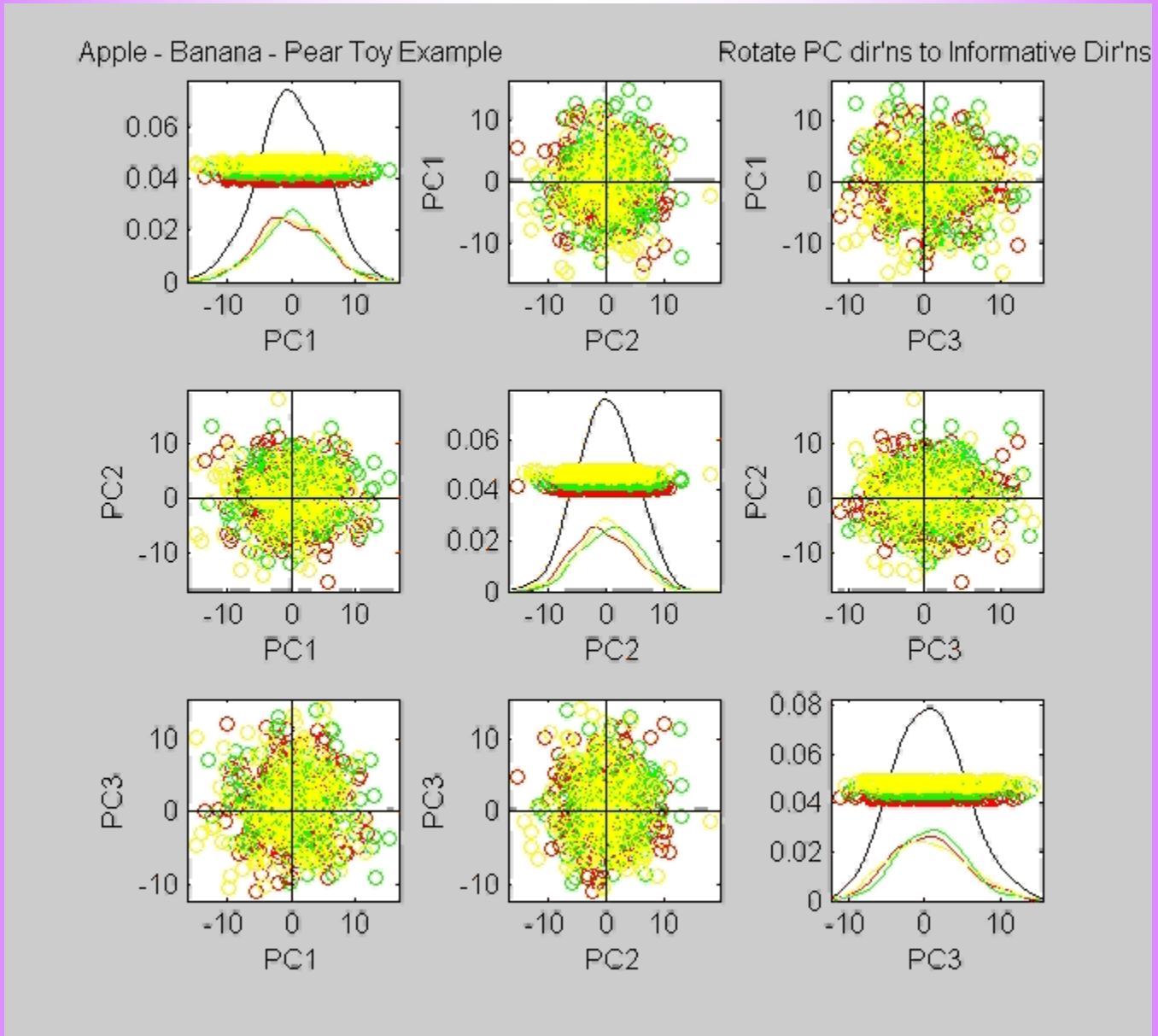
- PCA finds dir'ns of *maximal variation*
- Which may obscure interesting structure

Limitation of PCA

Toy Example:

- Apple – Banana – Pear

Limitation of PCA, Toy EX



Limitation of PCA

Toy Example:

- Apple – Banana – Pear
- Obscured by “noisy dimensions”
- First 3 PC directions only show noise
- Study some rotations, to find structure

Limitation of PCA

Main Point:

- May be *Important* Data Structure
- Not Visible in First Few PCs

Limitation of PCA, E.g.

Interesting Data Set: NCI-60

- NCI = National Cancer Institute
- 60 Cell Lines (cancer treatment targets)
  For Different Cancer Types
- Measured "Gene Expression"
  = "Gene Activity"
- Several Thousand Genes (Simultaneously)
- Data Objects = Vectors of Gene Exp'n
- Lots of Preprocessing

NCI 60 Data

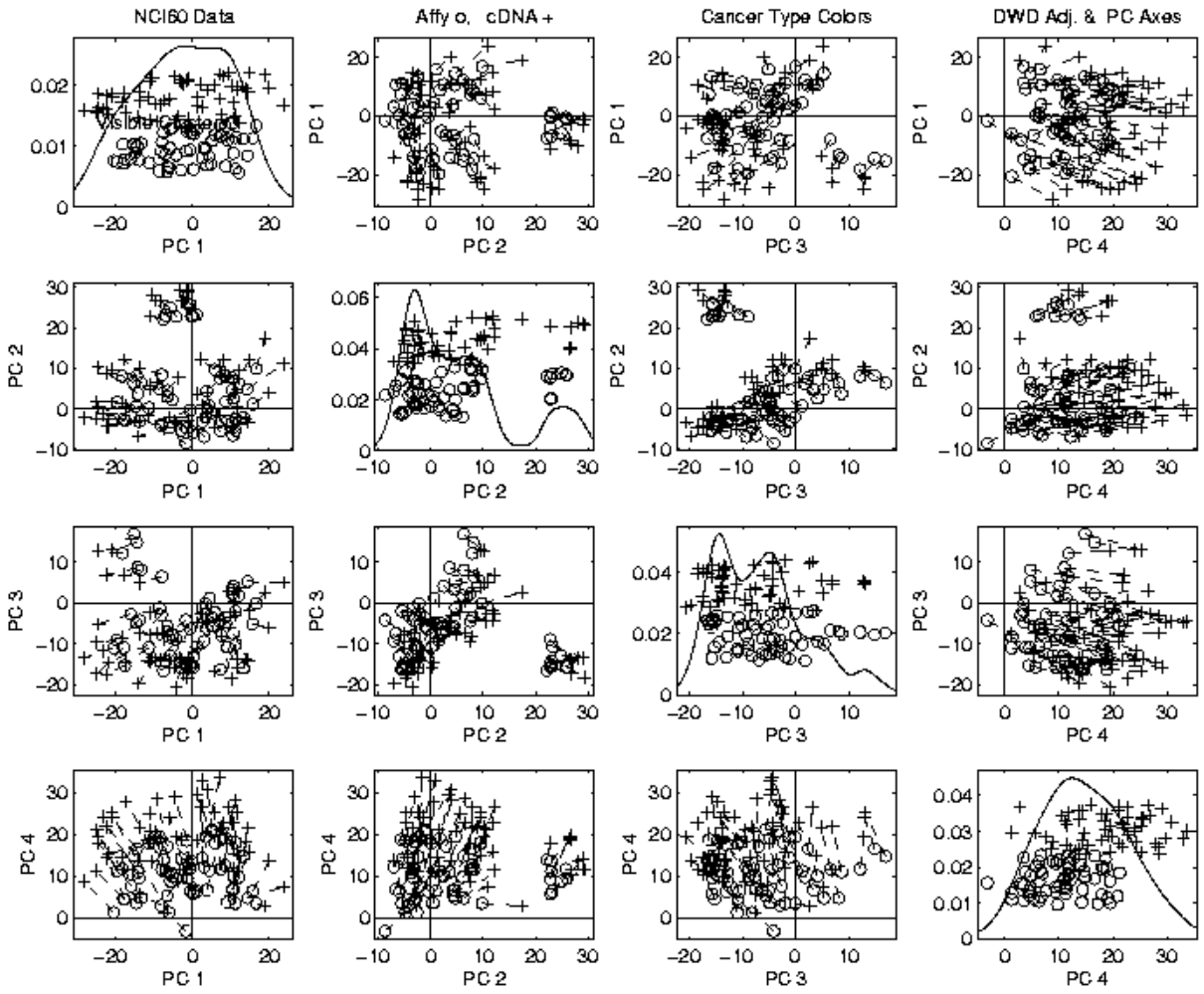
Important Aspect: 8 Cancer Types

- Renal Cancer
- Non Small Cell Lung Cancer
- Central Nervous System Cancer
- Ovarian Cancer
- Leukemia Cancer
- Colon Cancer
- Breast Cancer
- Melanoma (Skin)

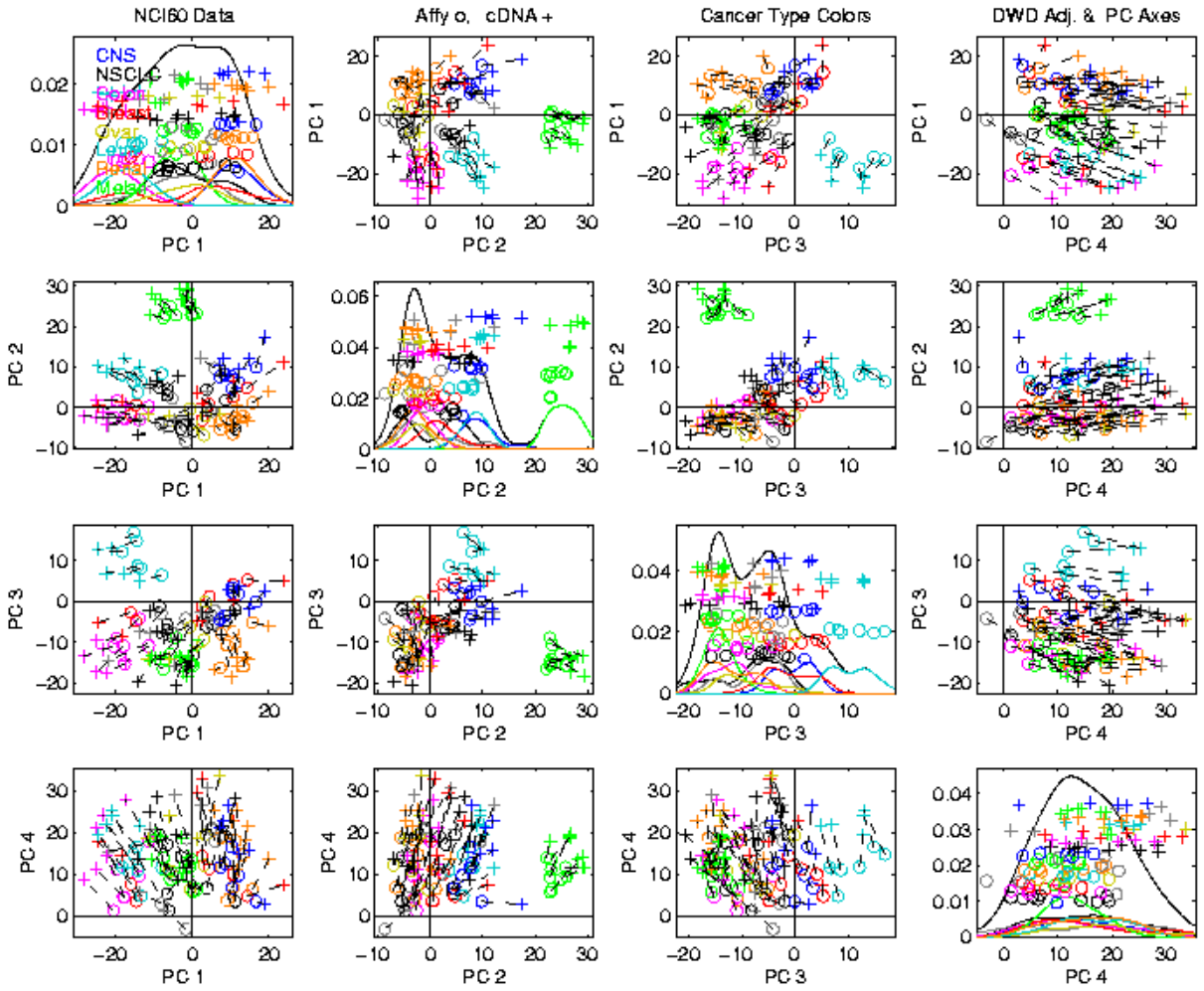
PCA Visualization of NCI 60 Data

- Can we find classes:
  (Renal, CNS, Ovar, Leuk, Colon, Melan)
- Using PC directions?
- First try “unsupervised view”
- I.e. switch off class colors
- Then turn on colors, to identify clusters
- I.e. look at “supervised view”

NCI 60: Can we find classes Using PCA view?



NCI 60: Can we find classes Using PCA view?



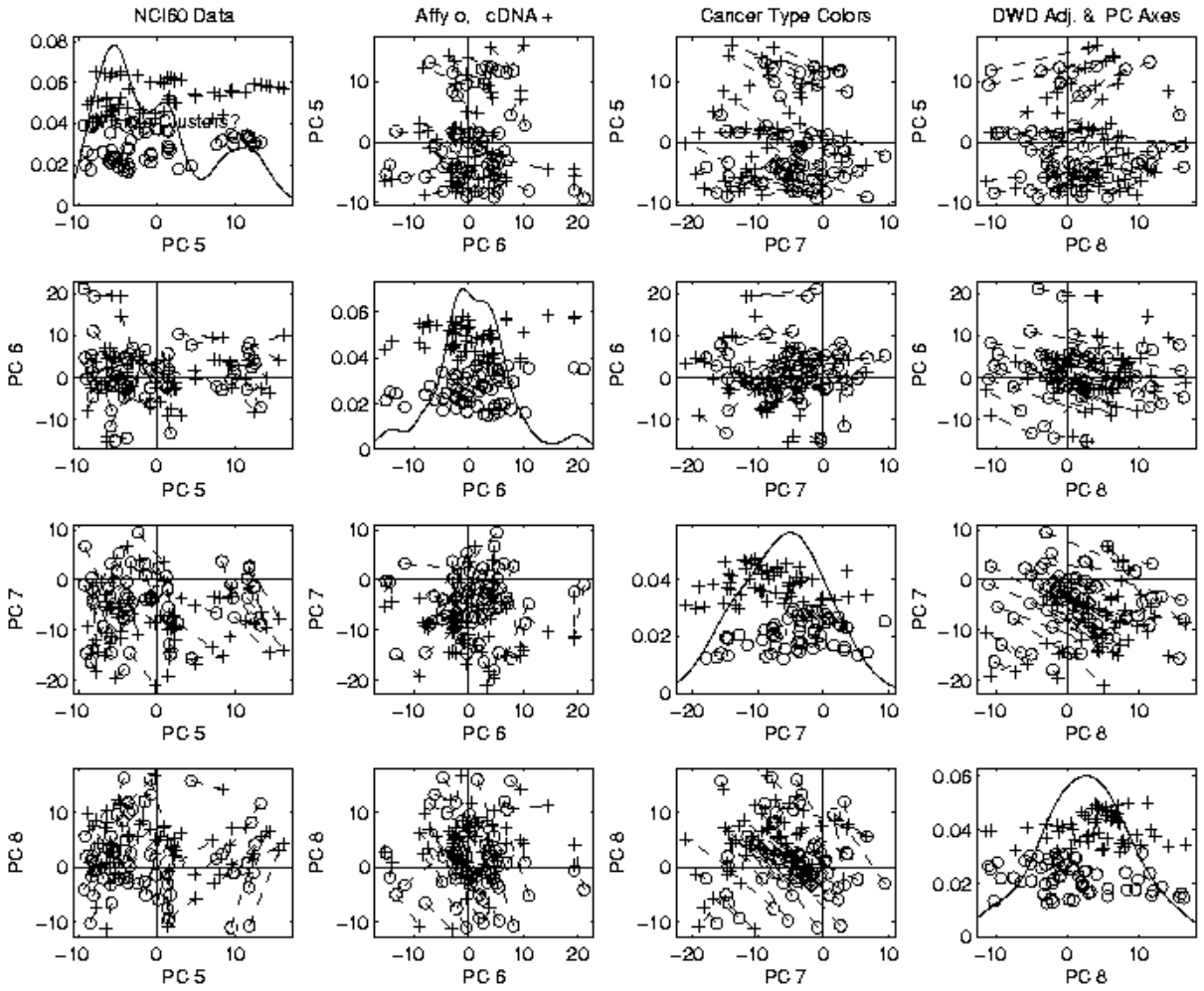
PCA Visualization of NCI 60 Data

Maybe need to look at more PCs?

Study array of such PCA projections:

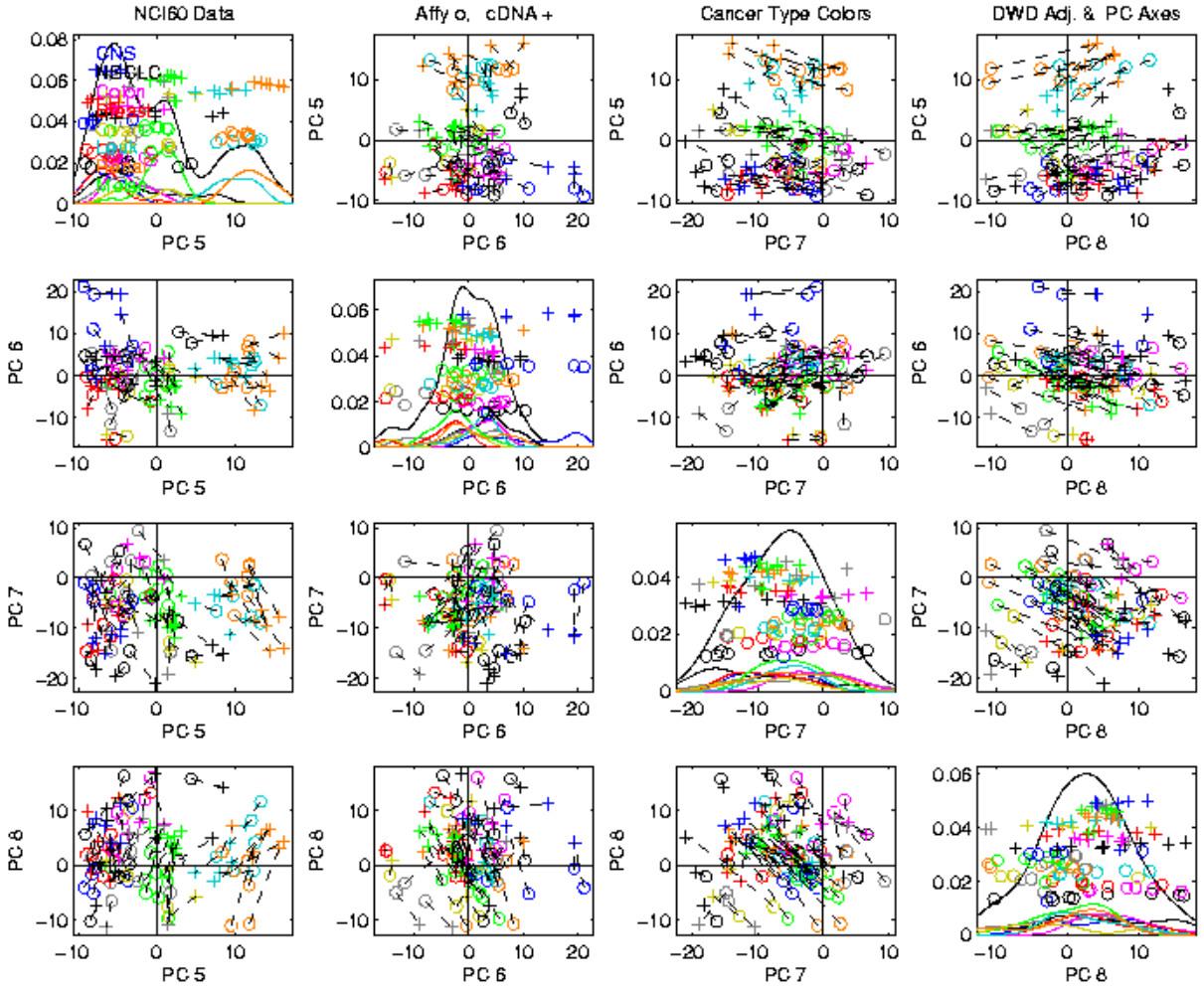
$$[PC1-4] \quad [1-4\ vs\ 5-8] \quad [1-4\ vs\ 8-12]$$
$$[PC5-8] \quad [5-8\ vs\ 9-12]$$
$$[PC9-12]$$

NCI 60: Can we find classes Using PCA 5-8?



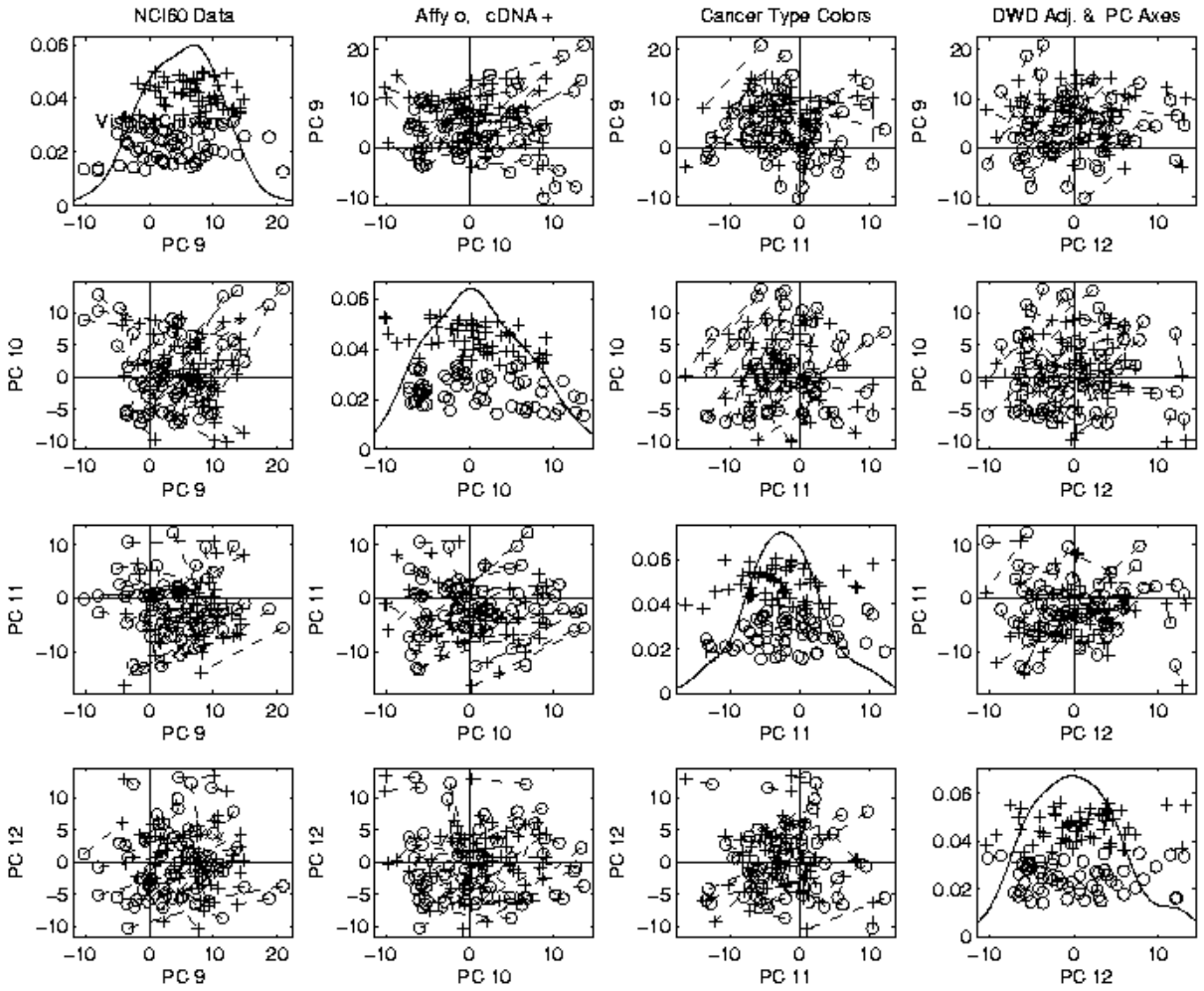
[ ] [ ] [ ]
[X] [ ]
[ ]

NCI 60: Can we find classes Using PCA 5-8?



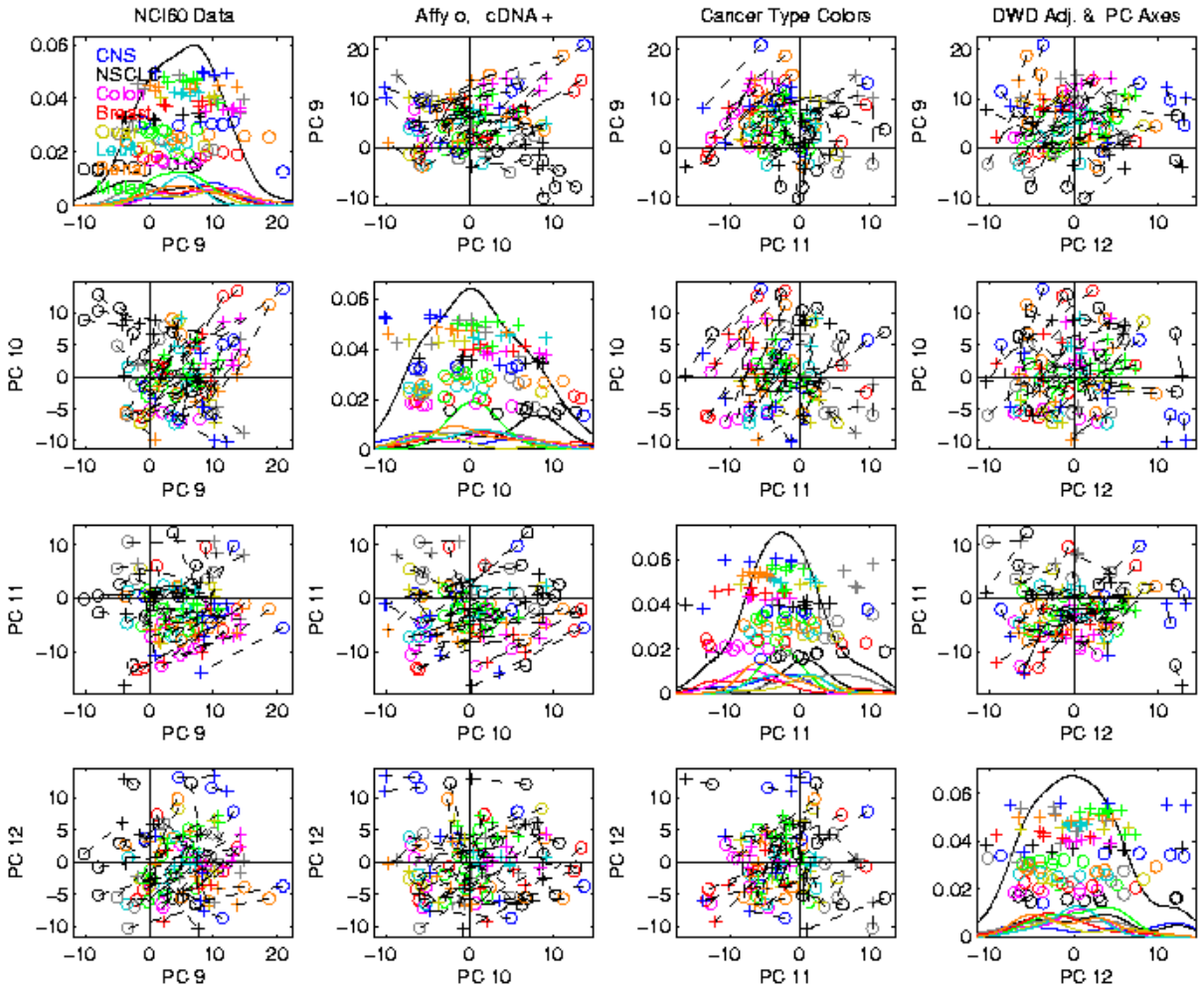
[ ] [ ] [ ]
[X] [ ]
[ ]

NCI 60: Can we find classes Using PCA 9-12?



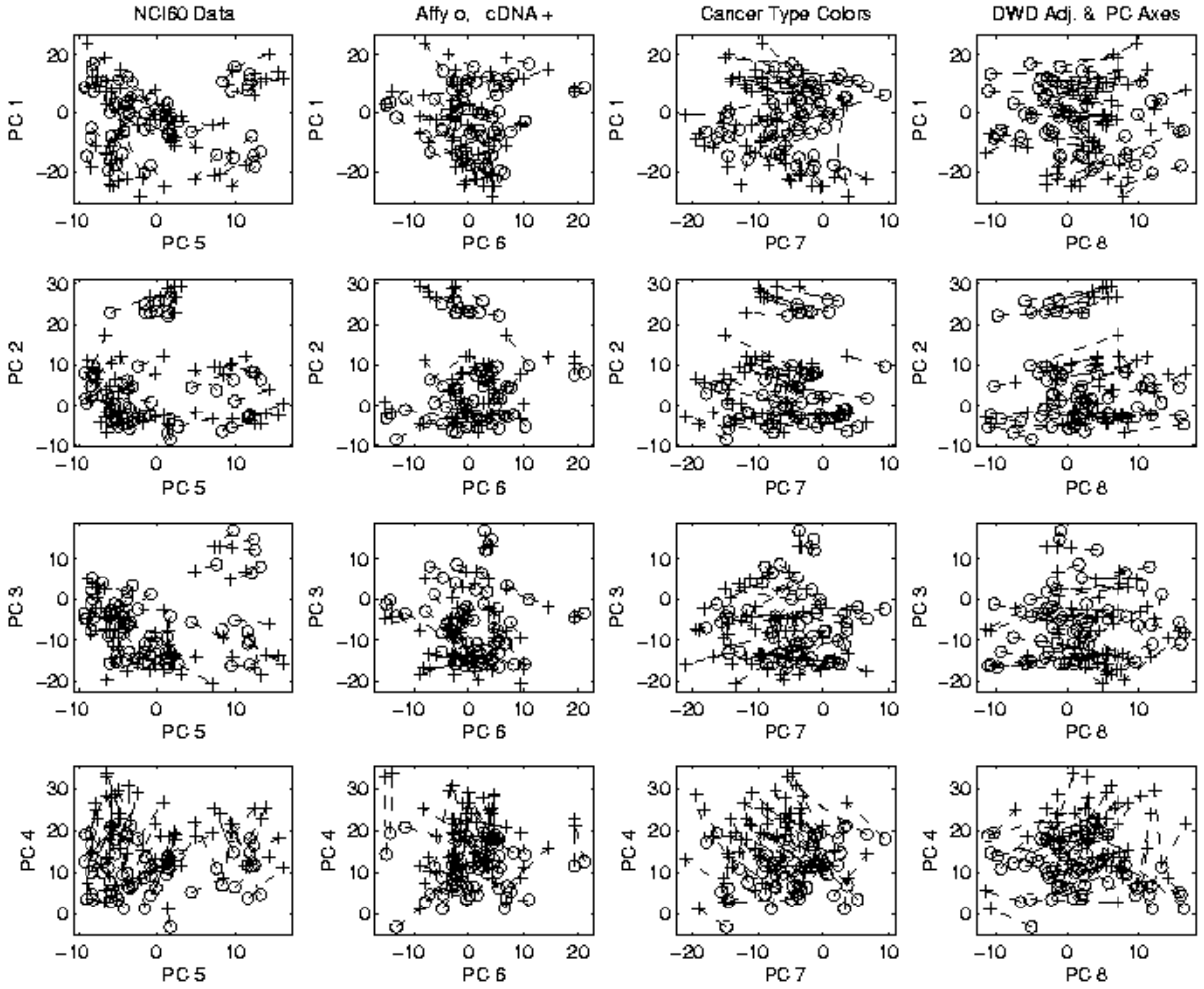
$$\begin{bmatrix} & \\ & \end{bmatrix} \begin{bmatrix} & \\ & \end{bmatrix} \begin{bmatrix} & \\ & \end{bmatrix} \begin{bmatrix} & \\ & \end{bmatrix} \begin{bmatrix} X \end{bmatrix}$$

NCI 60: Can we find classes Using PCA 9-12?



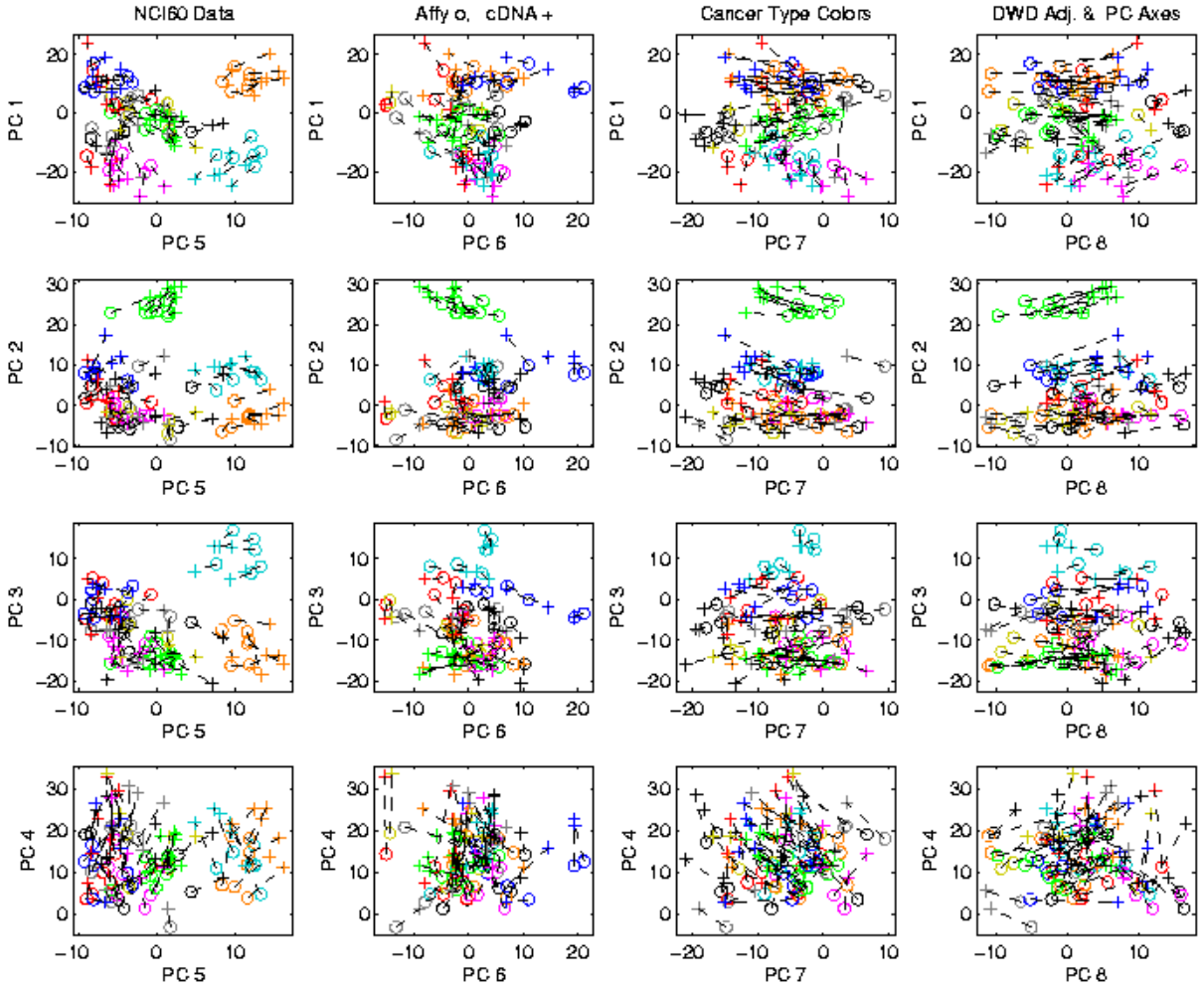
$$\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \quad \begin{bmatrix} & & \\ & & \end{bmatrix} \quad [X]$$

NCI 60: Can we find classes Using PCA 1-4 vs. 5-8?



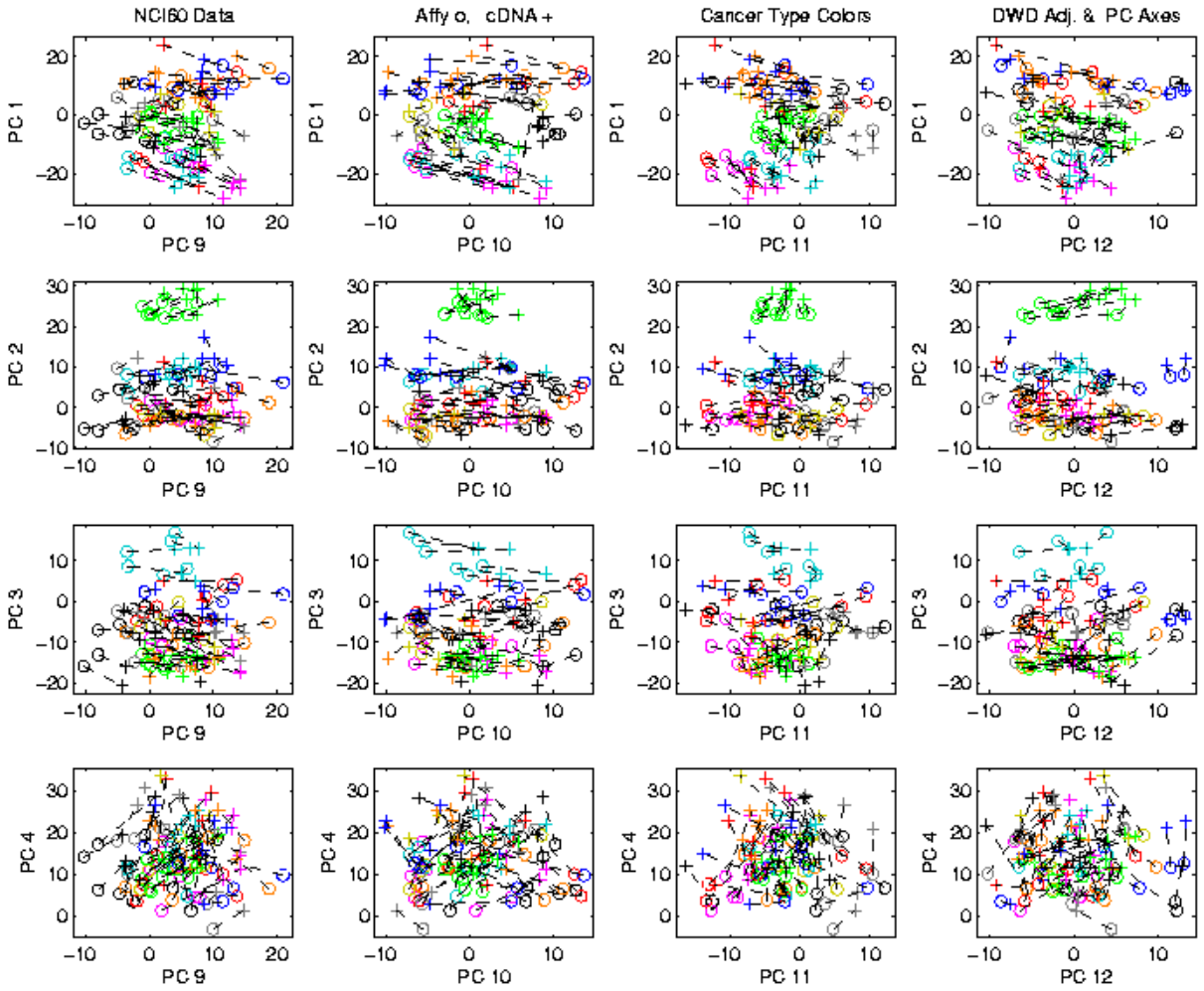
[ ] [X] [ ]
[ ] [ ]
[ ]

NCI 60: Can we find classes Using PCA 1-4 vs. 5-8?



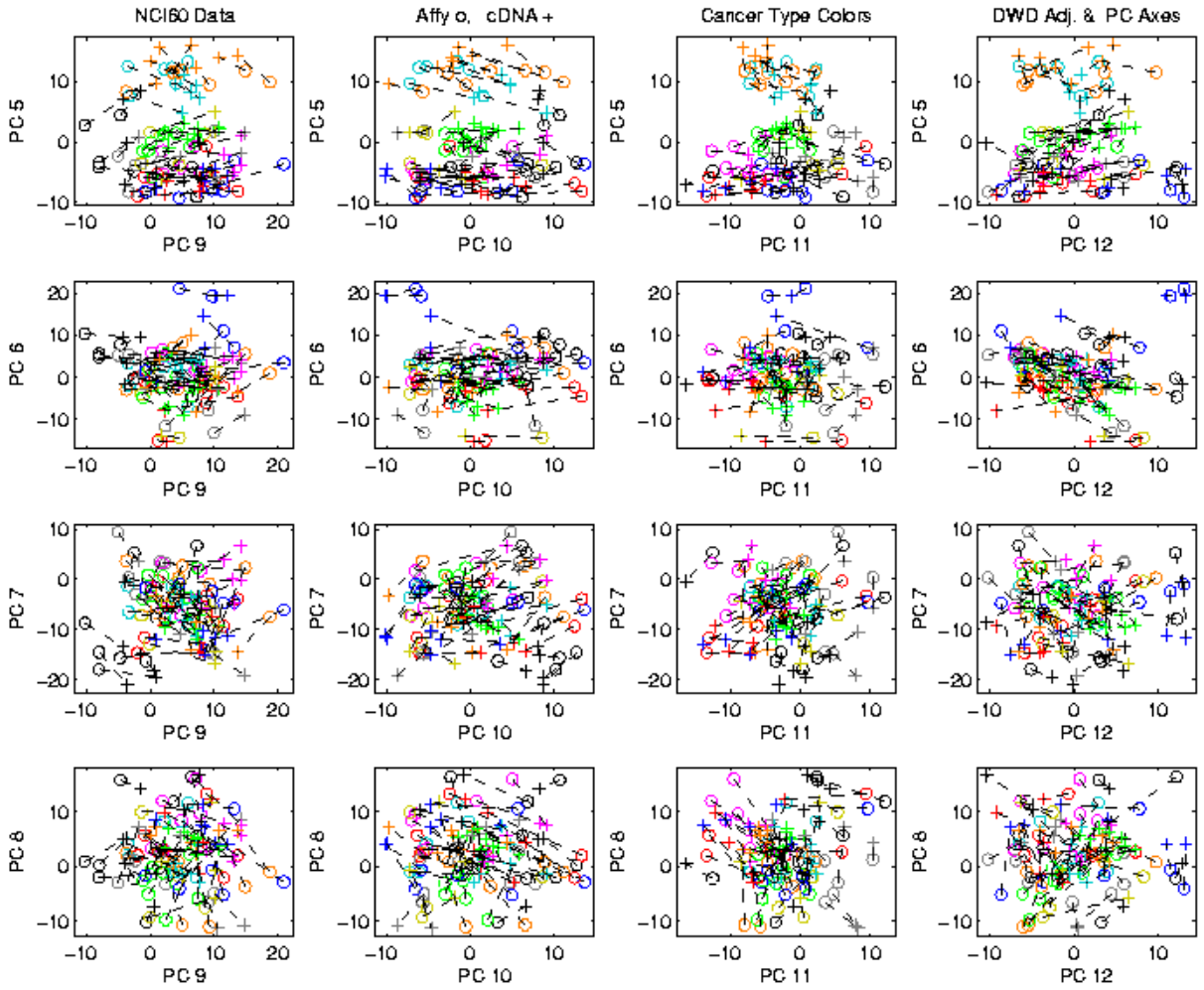
$$\begin{bmatrix} \end{bmatrix} \begin{bmatrix} X \end{bmatrix} \begin{bmatrix} \end{bmatrix}$$
$$\begin{bmatrix} \end{bmatrix} \begin{bmatrix} \end{bmatrix}$$
$$\begin{bmatrix} \end{bmatrix}$$

NCI 60: Can we find classes Using PCA 1-4 vs. 9-12?



$$\begin{bmatrix} \ \end{bmatrix} \quad \begin{bmatrix} \ \end{bmatrix} \quad \begin{bmatrix} X \end{bmatrix}$$
$$\begin{bmatrix} \ \end{bmatrix} \quad \begin{bmatrix} \ \end{bmatrix}$$
$$\begin{bmatrix} \ \end{bmatrix}$$

NCI 60: Can we find classes Using PCA 5-8 vs. 9-12?



$$\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$$
$$\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}$$
$$\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$$

PCA Visualization of NCI 60 Data

Can we find classes using PC directions??

- Found some, but not others
- Nothing after the first five PCs
- Rest seem to be noise driven

Are There Better Directions?

✓ PCA only “feels” maximal variation

✓ *Ignores* Class Labels

✓ How Can We Use Class Labels?