Principal Components Analysis of TOF-SIMS Data: Step by Step

NESAC/BIO

NESAC/BIO

Daniel J. Graham PhD

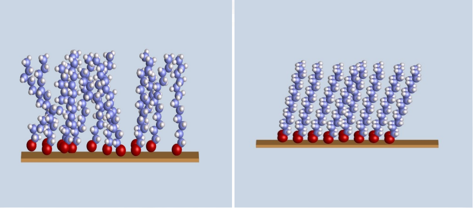
NESAC/BIO

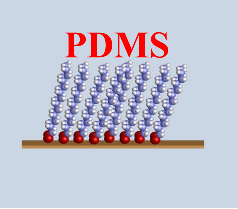
MVSA Website

UNIVERSITY OF WASHINGTON

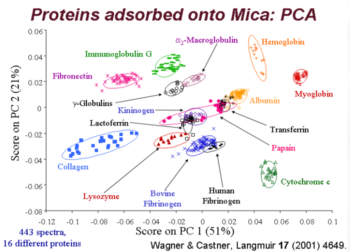
PCA has been successful

- **Monolayers**
  - **Graham and Ratner** *Langmuir 18 (2002) 5861-5868*
  - **Graham et. al.** *Langmuir 18 (2002) 1518-1527*





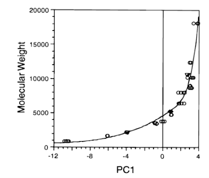
- **Proteins**
  - *M. S. Wagner† and David G. Castner Langmuir 2001, 17, 4649-4660 & Applied Surface Science 203-204 (2003) 698-703*
  - *Lhoest et. al. JBMR 57 (3): (2001) 432-440*



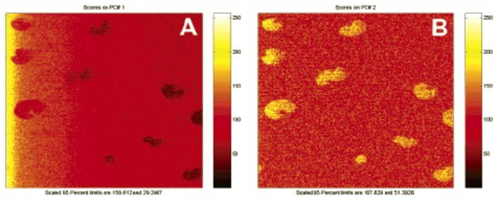
443 spectra,
16 different proteins

Wagner & Castner, Langmuir **17** (2001) 4649.

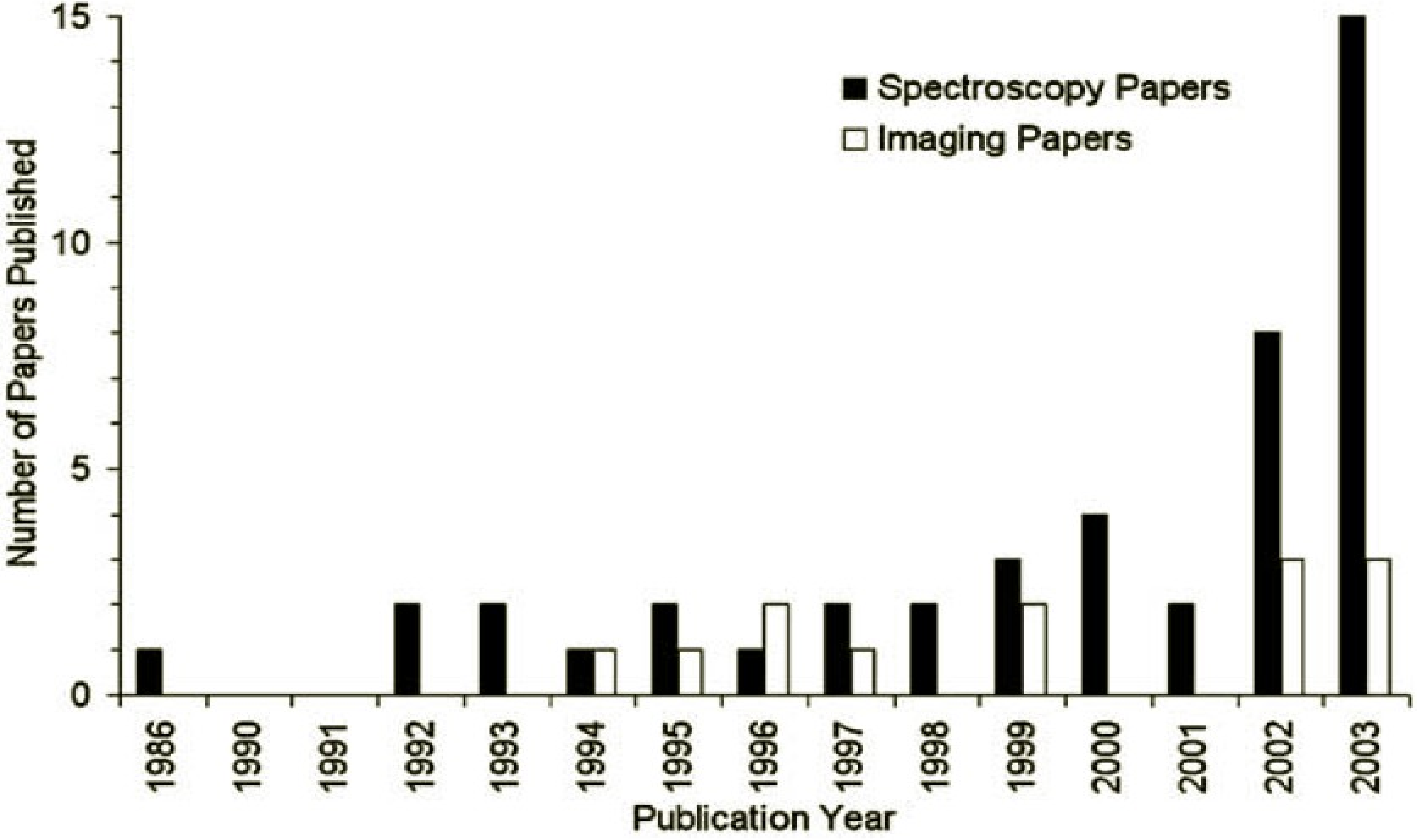
- **Polymers**
  - *Eynde and***Betrand** *Applied Surface Science 141 (1999) 1–20 & SIA 25, (1997)876
888*



- **Imaging**
  - **Wickes** *Surf. Interface Anal.* **2003**; 35: **640- 648**
  - *Biesinger et. al. Anal. Chem. 2002, 74, 5711-5716*



The Use of MVA Methods for TOF-SIMS is Increasing



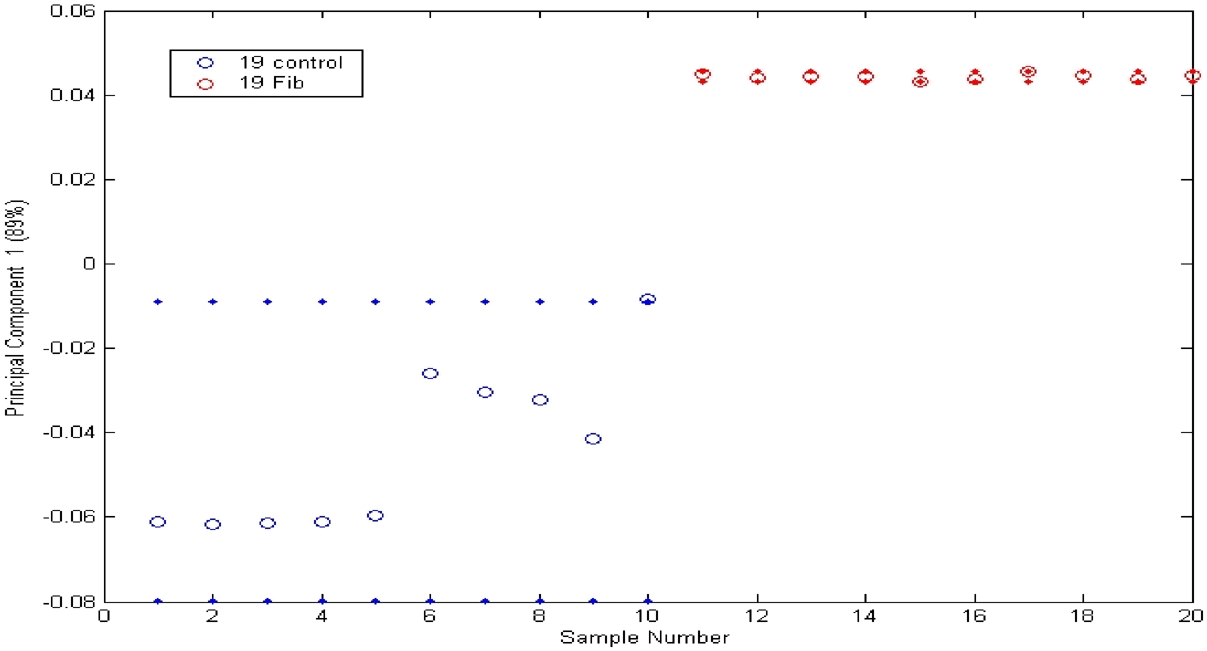
M.S. Wagner et. al., Surface Science 570 (2004) 78–97

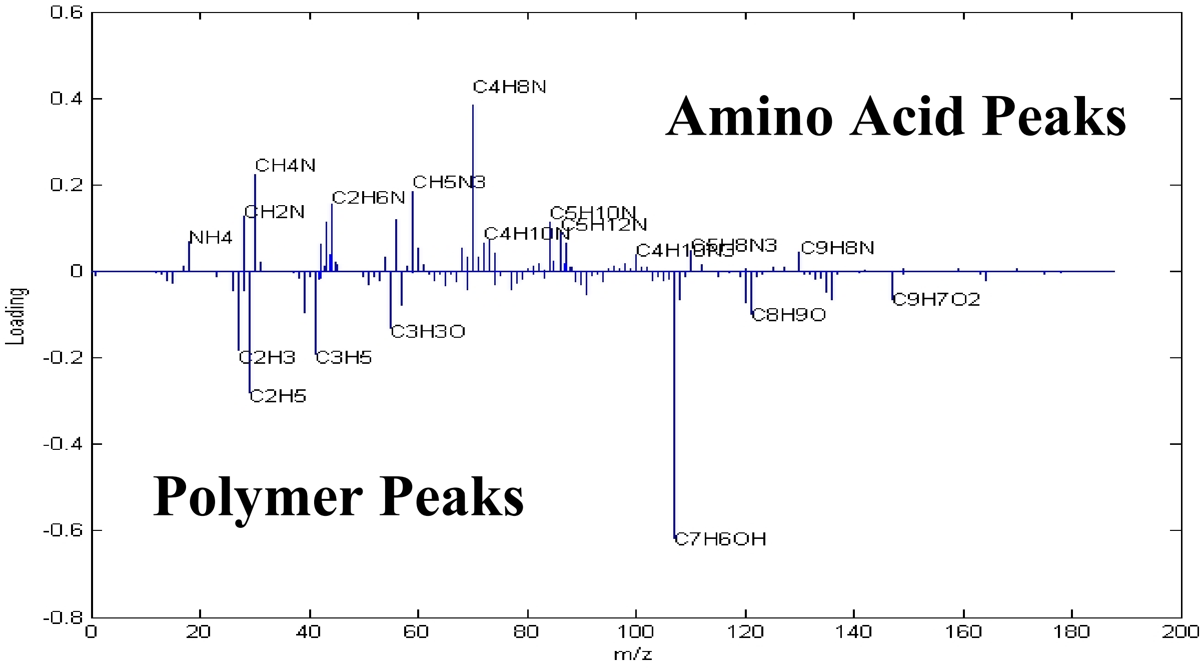
When Should PCA be Used?

- PCA should be used to help answer questions
  - Are surfaces A and B different?
  - How does treatment X change the surface chemistry?
  - How is fragmentation pattern affected by ____?
  - Can TOF-SIMS data distinguish Protein A from Protein B?
- The question should be part of the experimental design and not an afterthought

Example:
Is sample A different from Sample B?

PCA: Polyarylate polymer with and without Fibrinogen

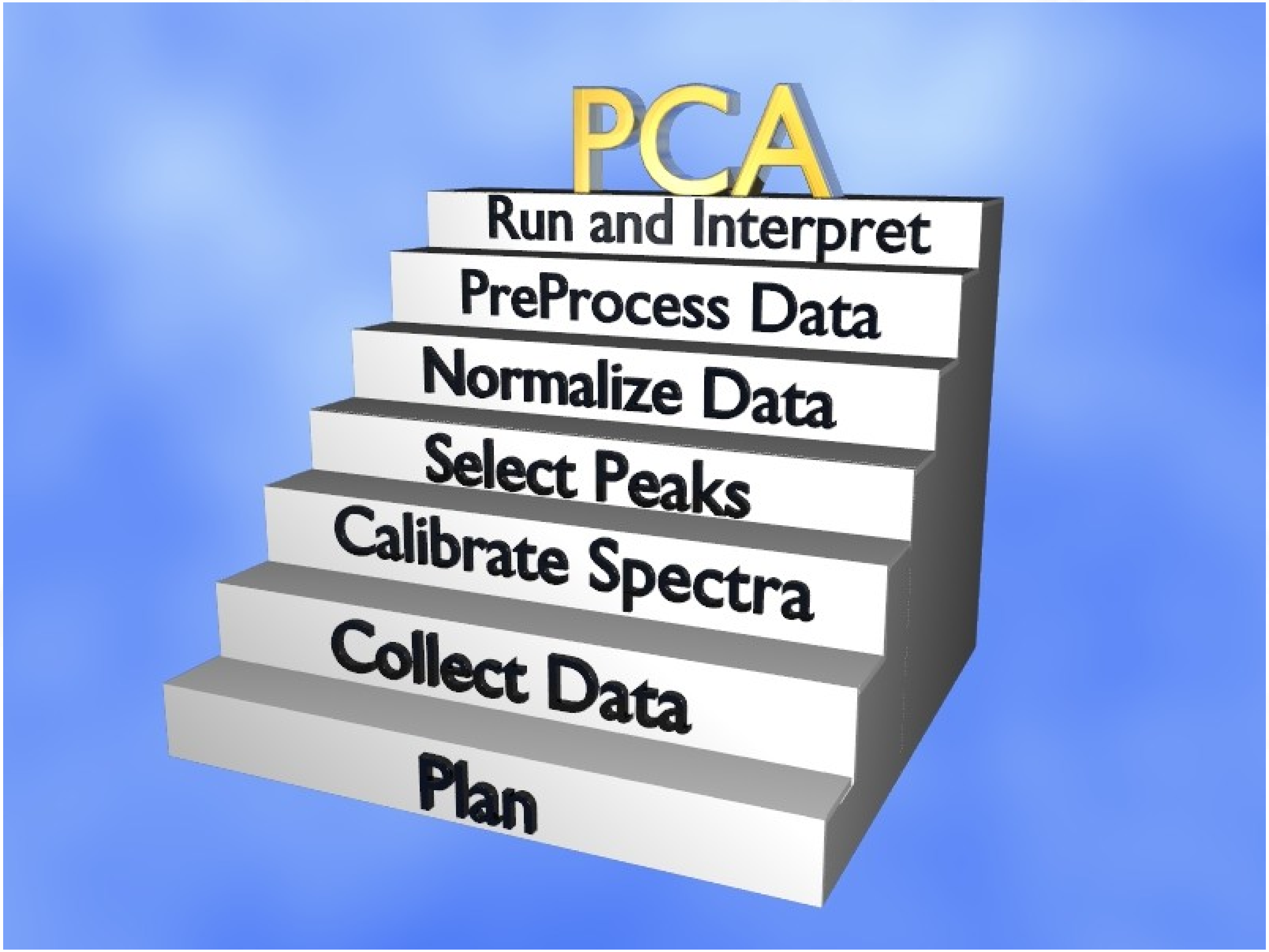




Amino Acid Peaks

Polymer Peaks

Steps to PCA



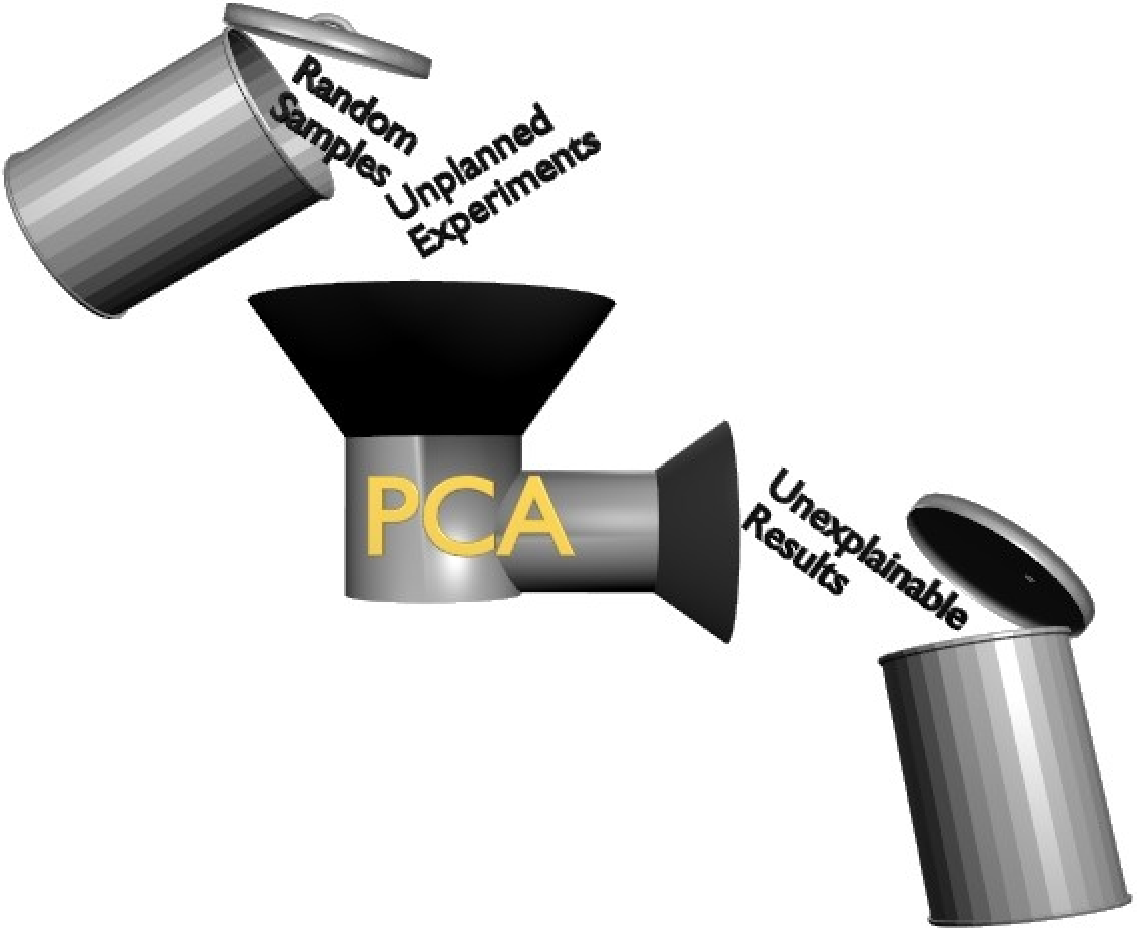
PCA
Run and Interpret
PreProcess Data
Normalize Data
Select Peaks
Calibrate Spectra
Collect Data
Plan

Plan

- What is the question you want to answer?
- What samples do you need to answer that question?
- How many samples/ replicates do you need?

Remember PCA will find the main differences between any samples

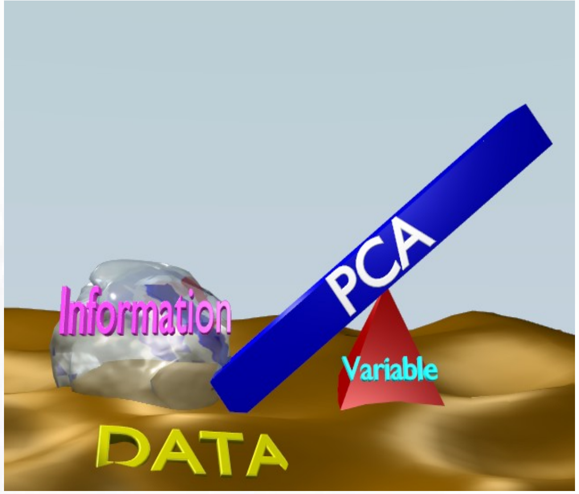
If you input garbage in



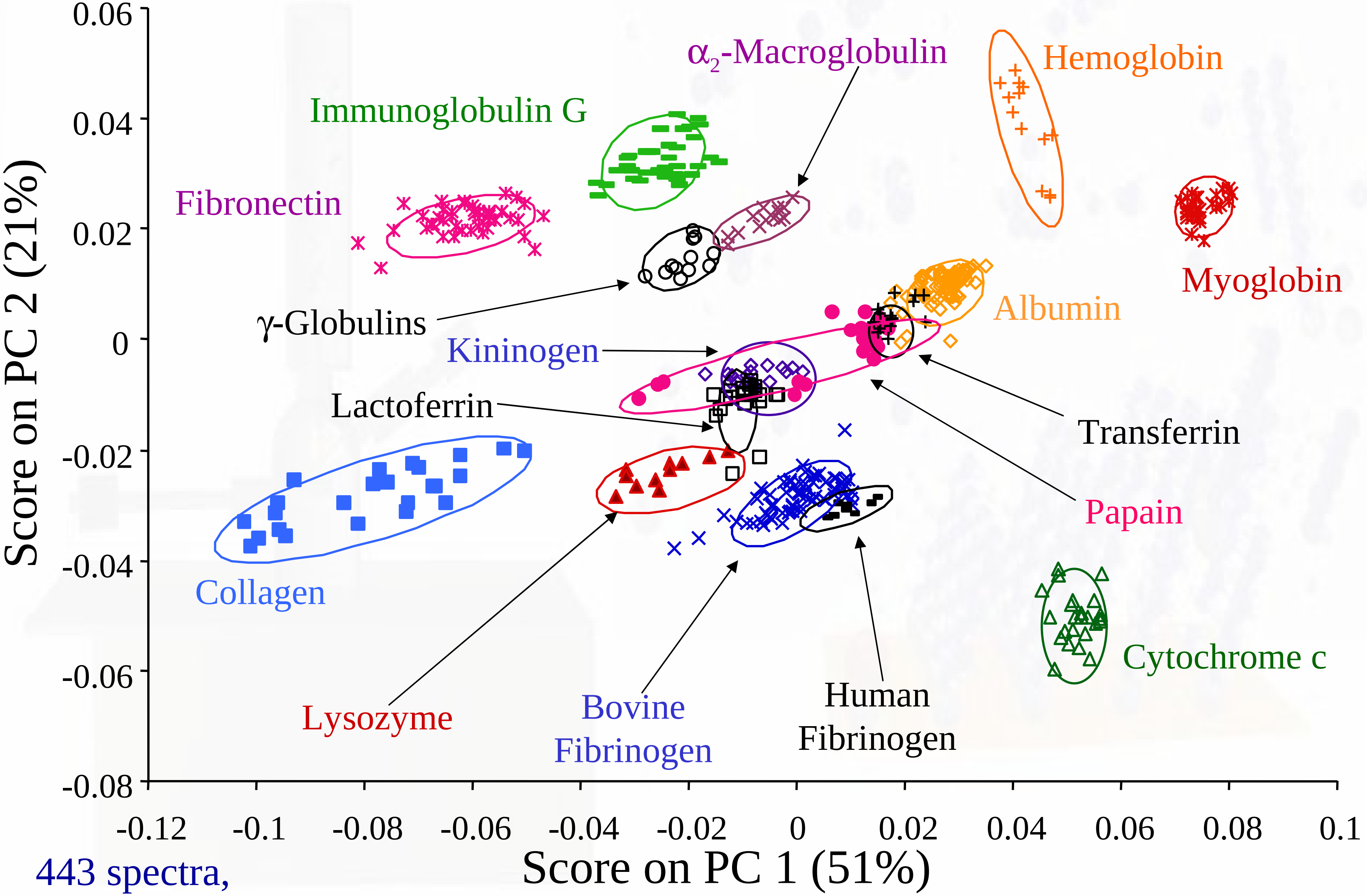
You will get garbage out!!!

Experimental Design/Data Collection

- Not all systems are well defined, but your experimental design can be:
  - Think about what you want to learn from SIMS
  - Simplify the number of variables you are dealing with per experiment
  - Plan appropriate controls
  - Run enough replicates to determine reproducibility
    - Homogeneous => 3 to 5 spots on 2 samples
    - Inhomogeneous => 5 to 7 spots on 3 to 5 samples



Proteins adsorbed onto Mica: PCA

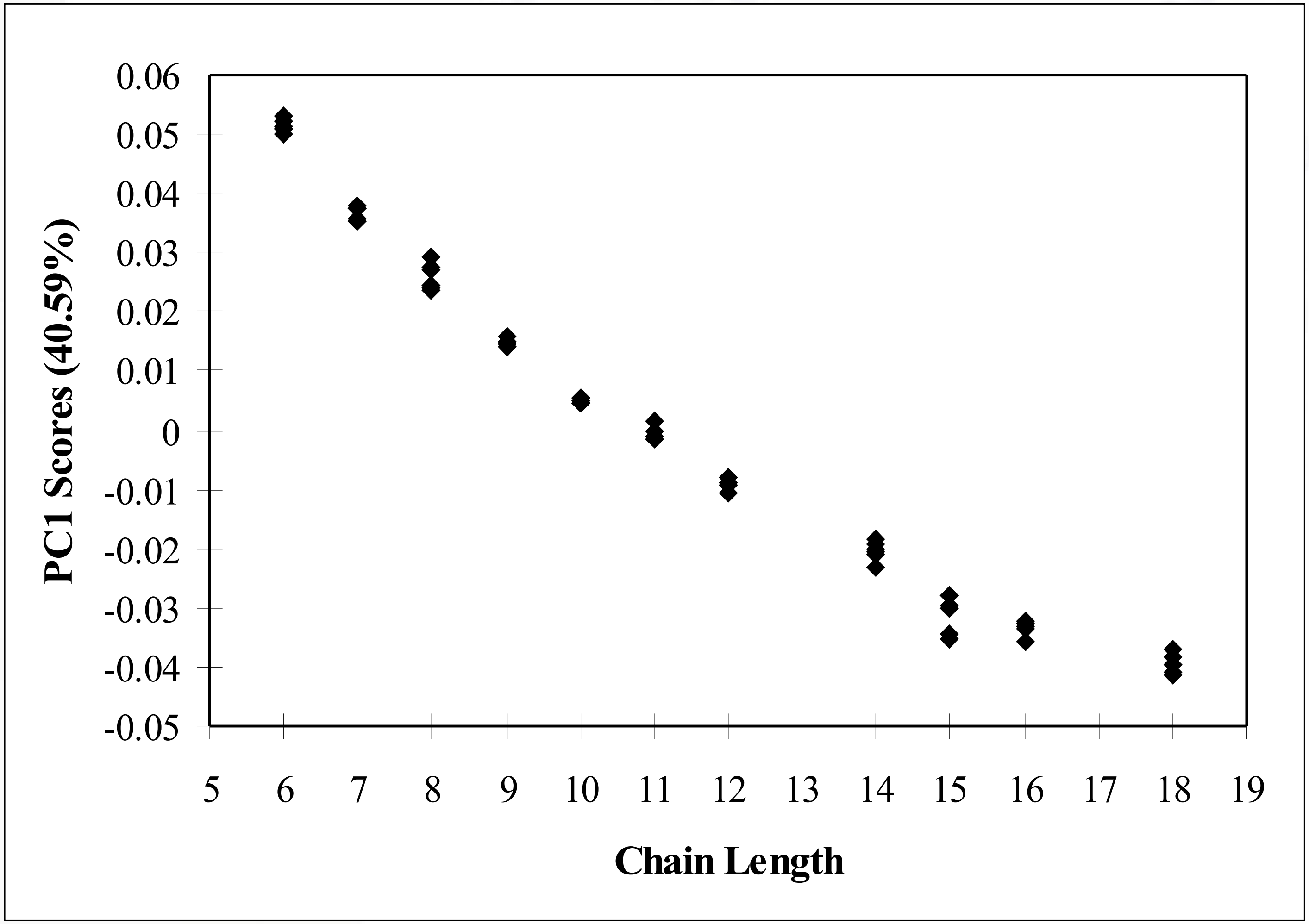


443 spectra,
16 different proteins

Wagner & Castner, Langmuir **17** (2001) 4649.

SAMs – typically very homogeneous

- Different chain length SAMs
- 6 spectra per chain length
- Most data points overlap showing high reproducibility

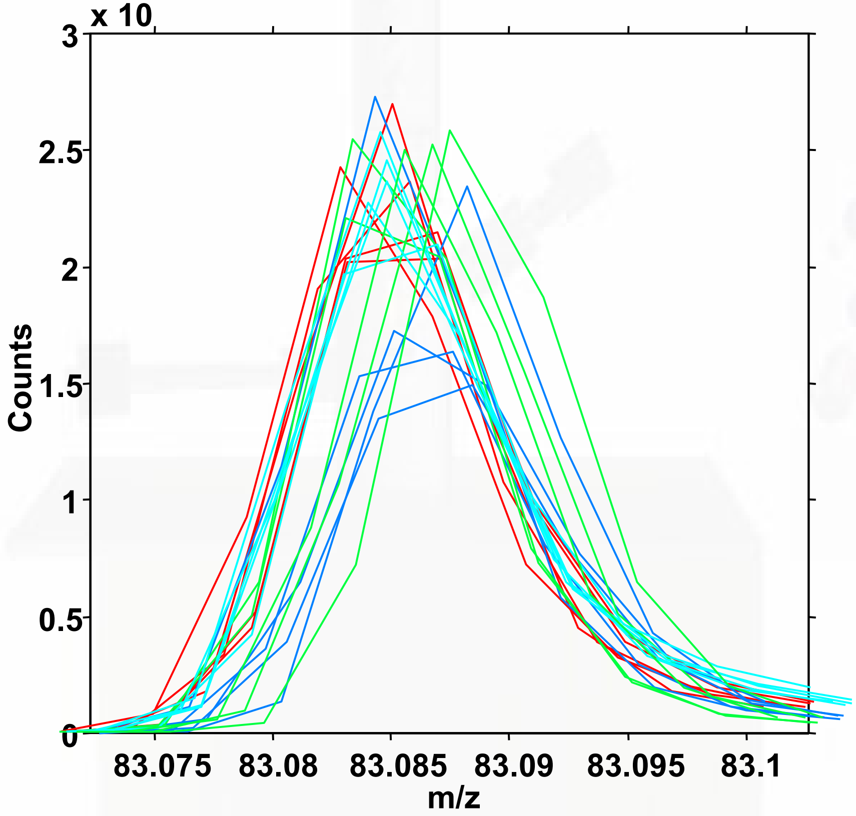


Data calibration

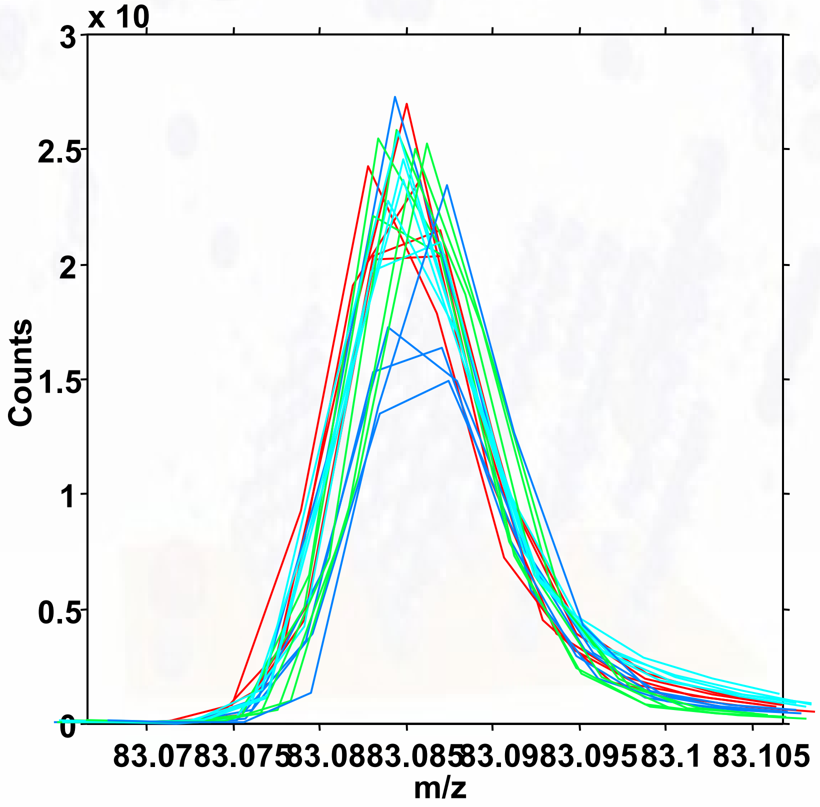
- All spectra in the data set should be calibrated to the same peak set
  - Be consistent
- Include a high-mass peak if possible
  - This will increase the accuracy of identifying high mass peaks
- Don't trust autocalibration functions
  - They can make mistakes

Calibration example

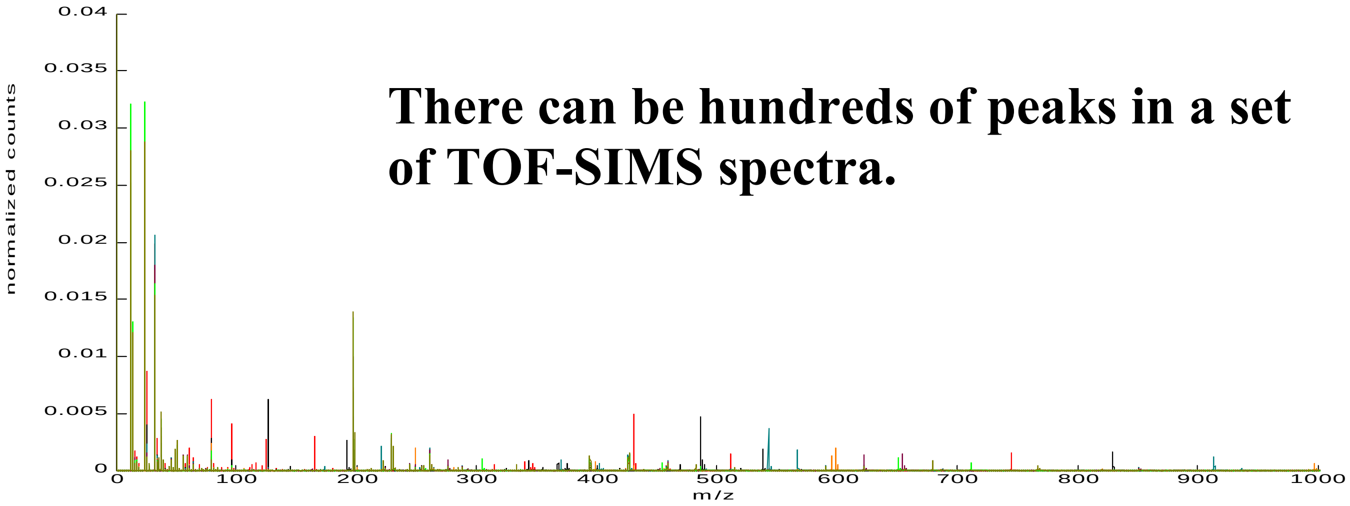
Initial Calibration



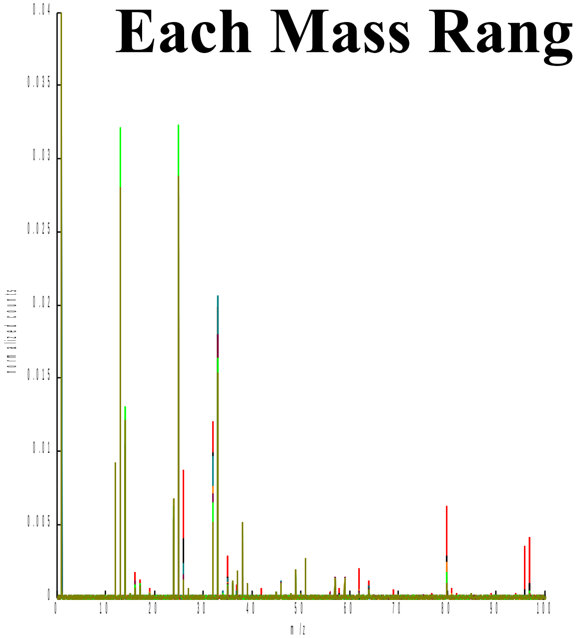
After Checking Calibration to assure consistency



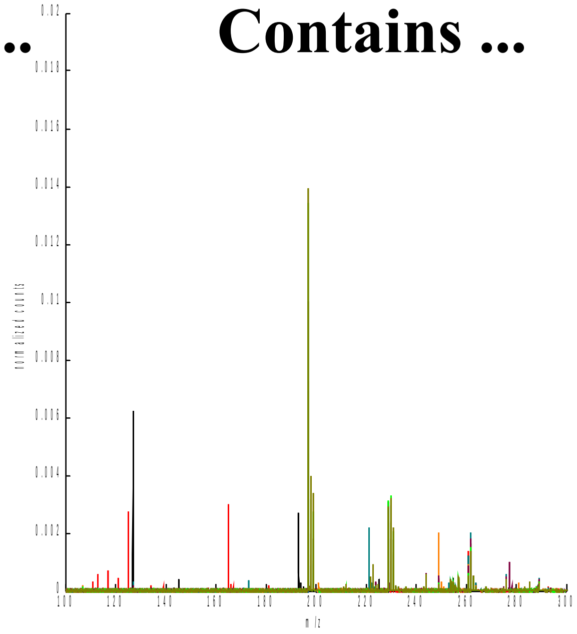
Peak Selection-Which Peaks should you select?



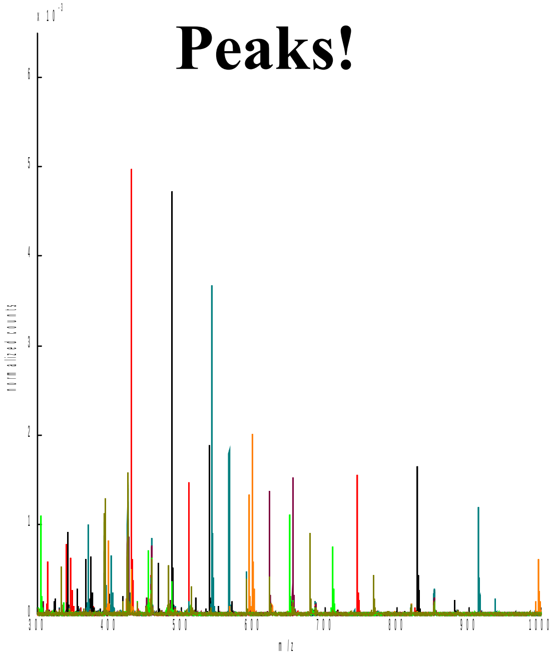
There can be hundreds of peaks in a set of TOF-SIMS spectra.



Each Mass Range...



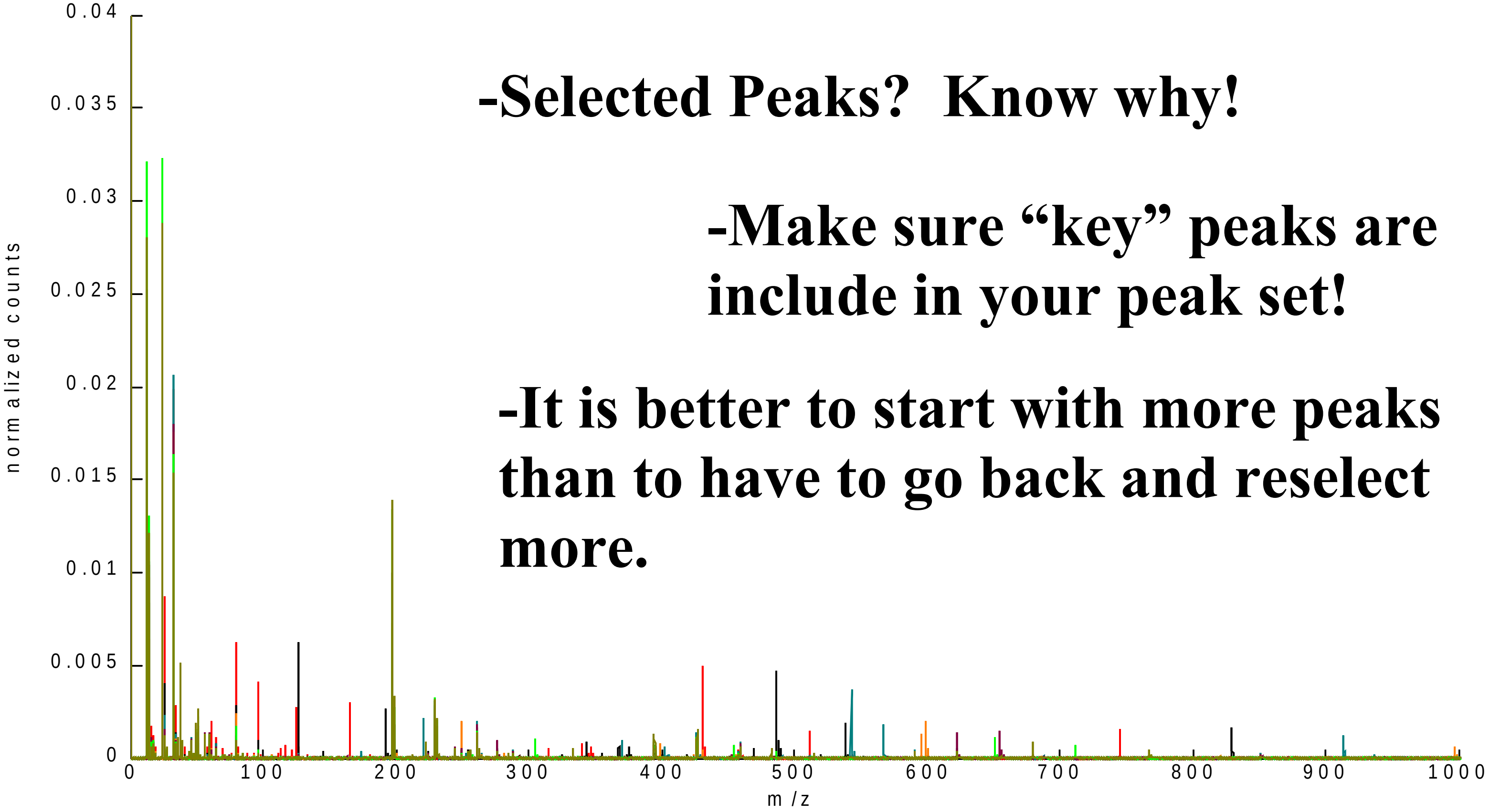
Contains ...



Peaks!

Peak Selection-Which Peaks should you select?

-All Peaks?

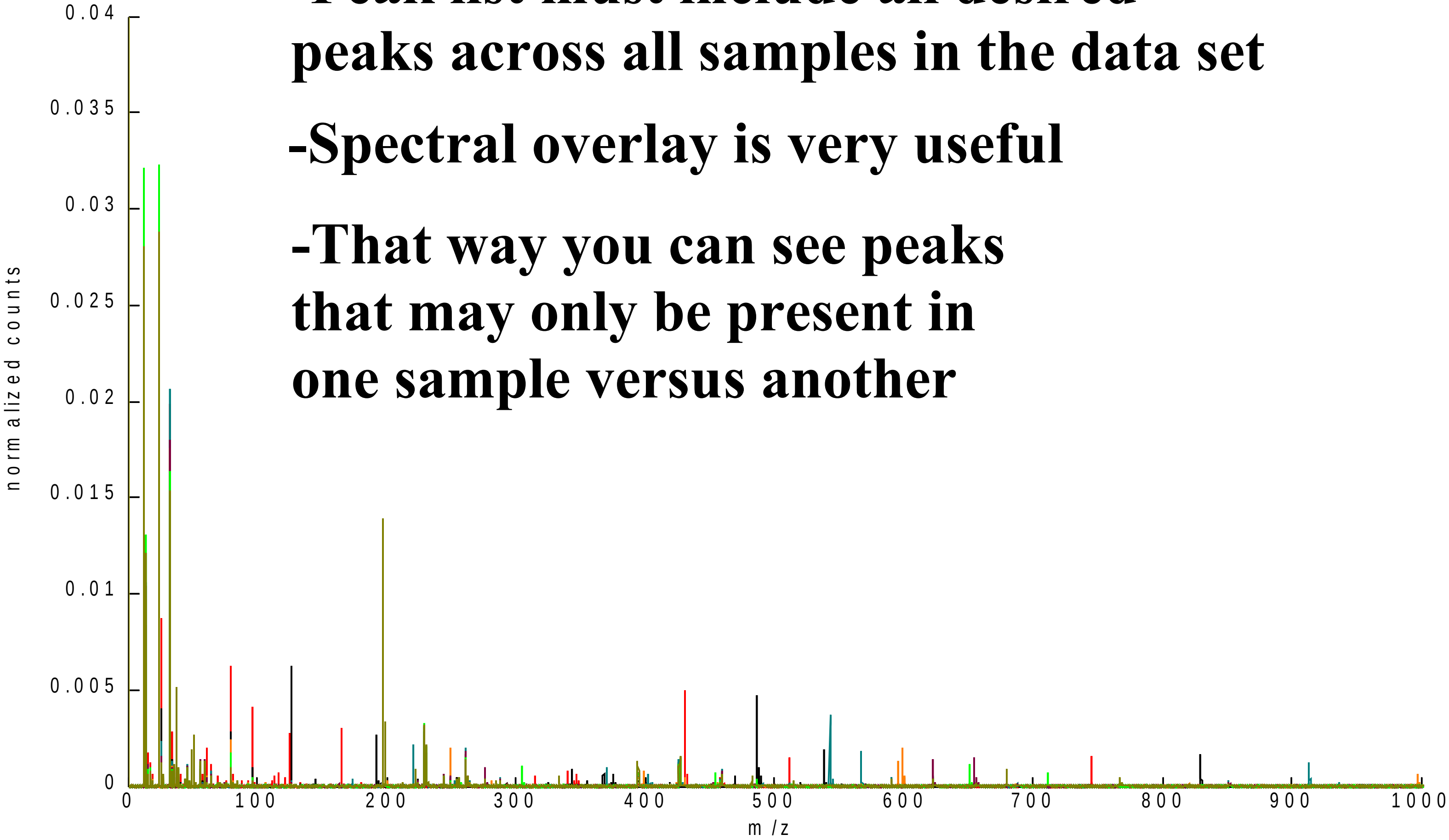


-Selected Peaks? Know why!

-Make sure "key" peaks are include in your peak set!

-It is better to start with more peaks than to have to go back and reselect more.

Peak Selection-Which Peaks should you select?



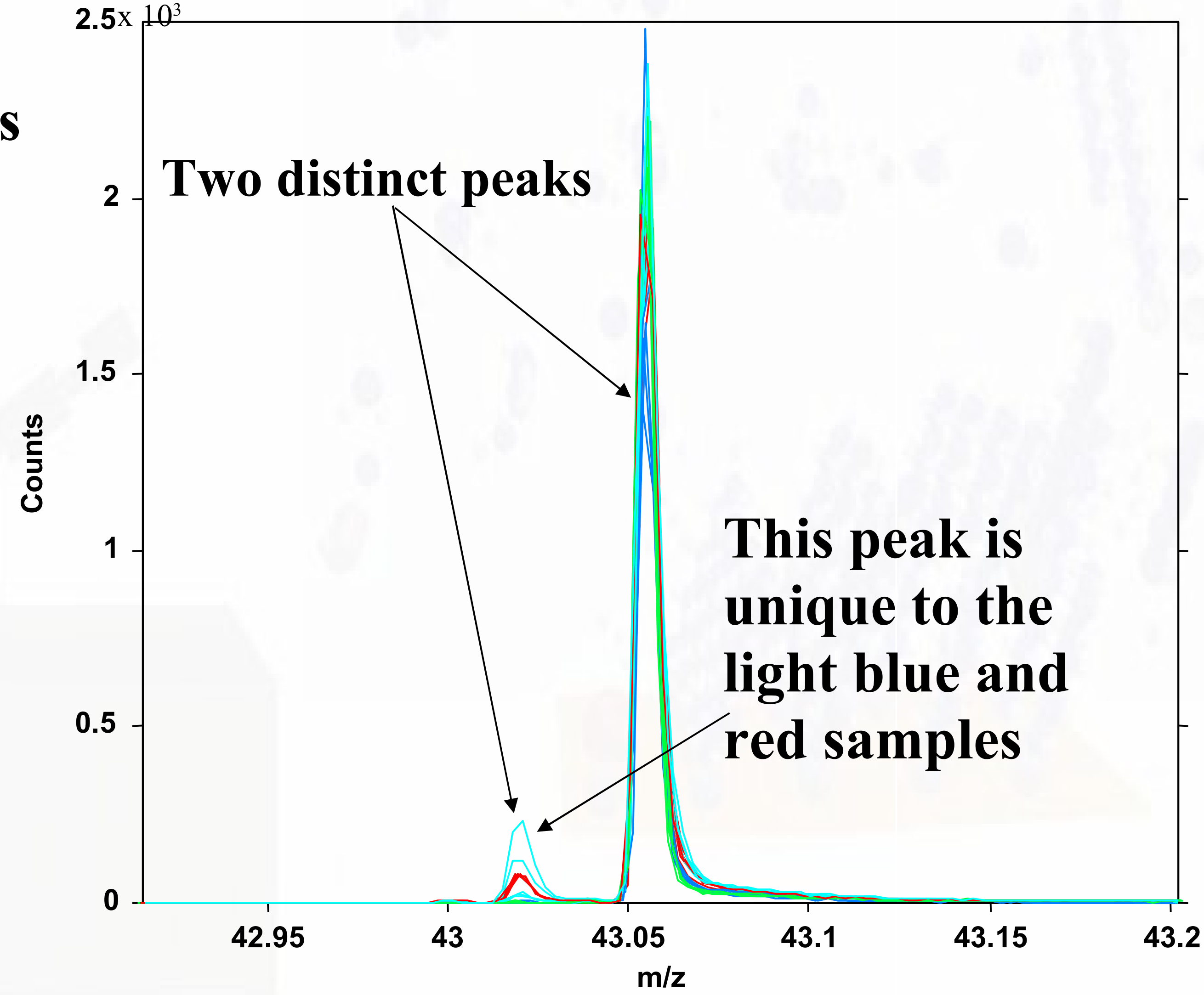
-Peak list must include all desired peaks across all samples in the data set

-Spectral overlay is very useful

-That way you can see peaks that may only be present in one sample versus another

Peak Selection Continued

-To keep high-mass resolution of TOF-SIMS you need to select individual peaks



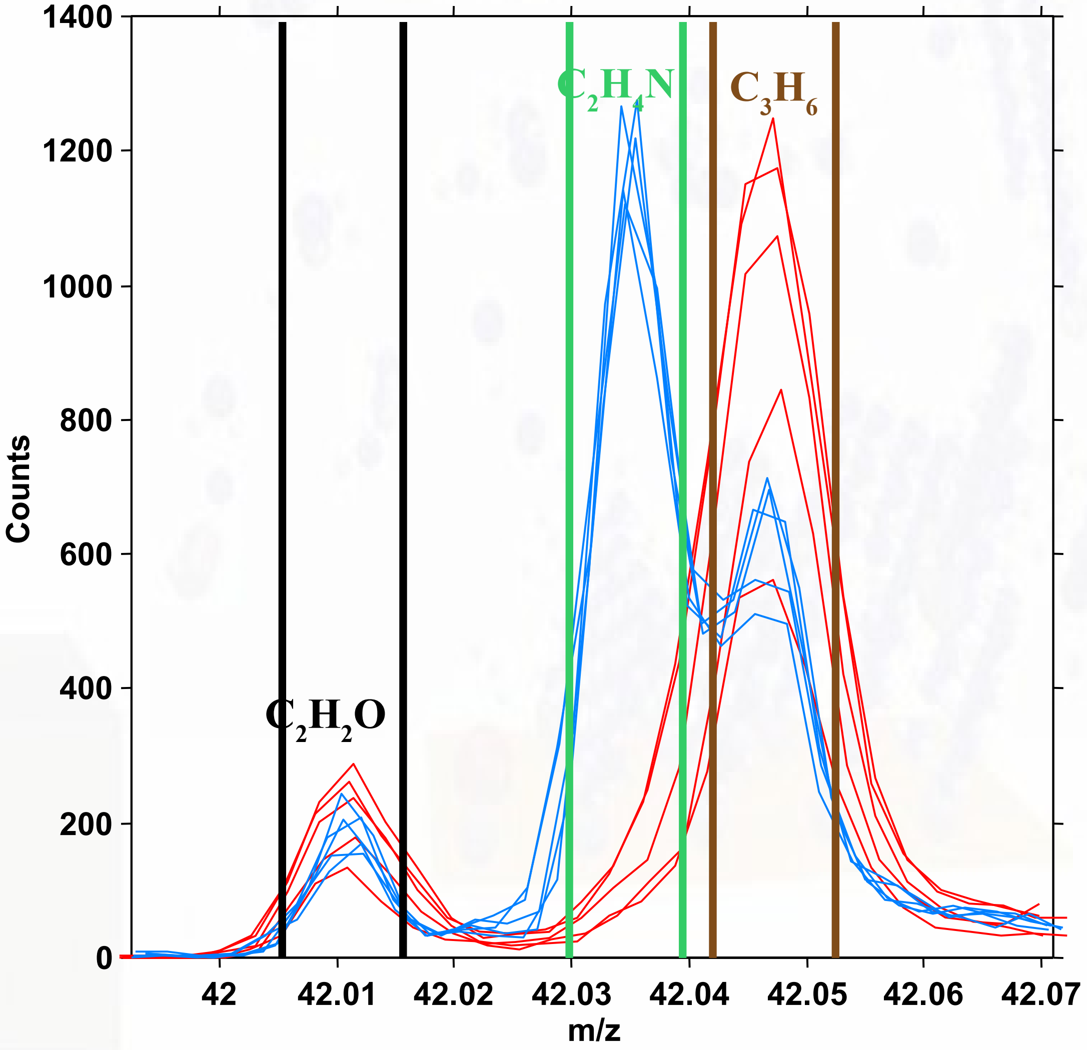
Two distinct peaks

This peak is unique to the light blue and red samples

Carefully Set Integration Limits

If your software allows you to set integration limits manually:

- **Overlay spectra so you can set limits properly for all samples**
- **Set the limits tightly around the peaks**
- **Set all limits consistently**



Data Pretreatment

- Typical data pretreatments include
  - Normalization
  - Centering
  - Scaling
- Pretreatments are done in an attempt to **maximize** differences due to sample differences and minimize differences from other sources
- Know the assumptions being made
  - Are they valid?

Data Normalization

- Normalization is most common data scaling method
- Normalization is typically done to remove differences in the data due to:
  - Sample charging
  - Instrument variations
- Attempts to remove variation in the data not due to sample differences

• Common Normalization Methods

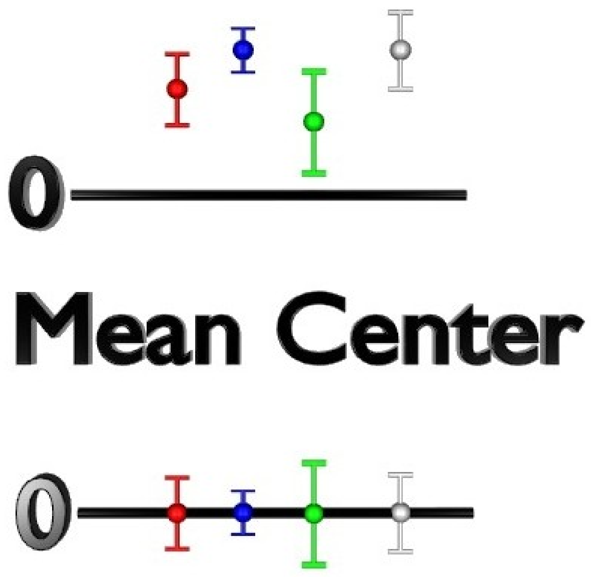
-Total intensity → Good when selecting most all peaks

-Sum of selected peaks → Good when selecting only a subset of peaks (normalization must be redone if you remove peaks from the data set)

-Highest peak in spectra
-User selected peak → Need good reason for peak choice
Can introduce user bias

Data Centering

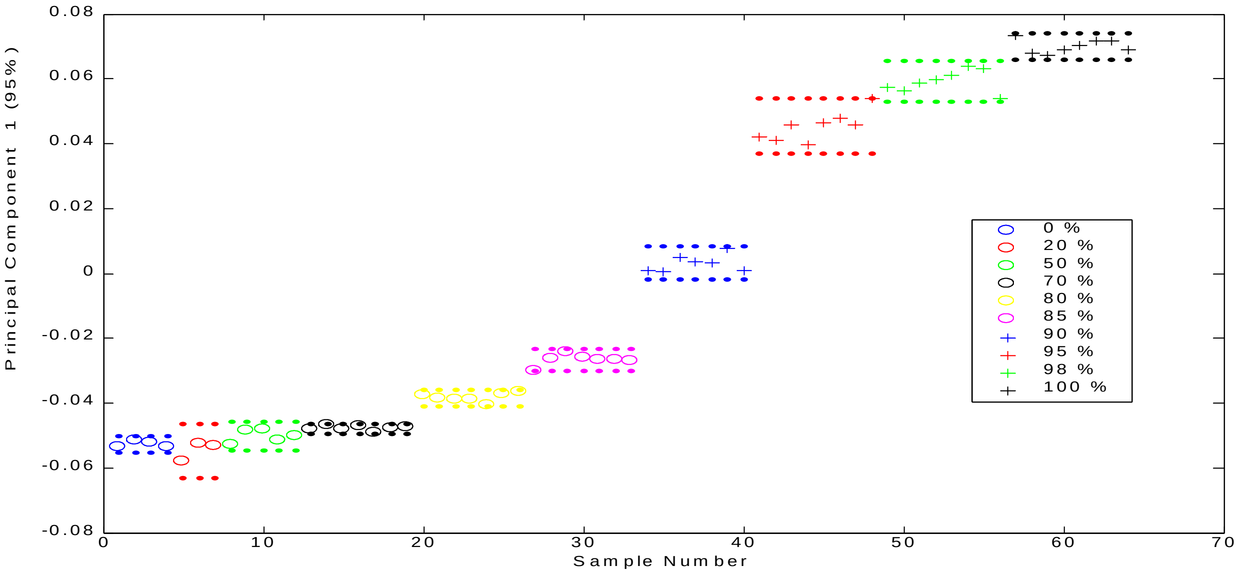
- Centering is done to remove
  - A common offset from the data
  - Differences in the means between samples
- Mean Centering
  - Subtracts the mean of each variable from each measurement from that variable
  - Makes it so data varies across common mean of zero



Other Data Scaling

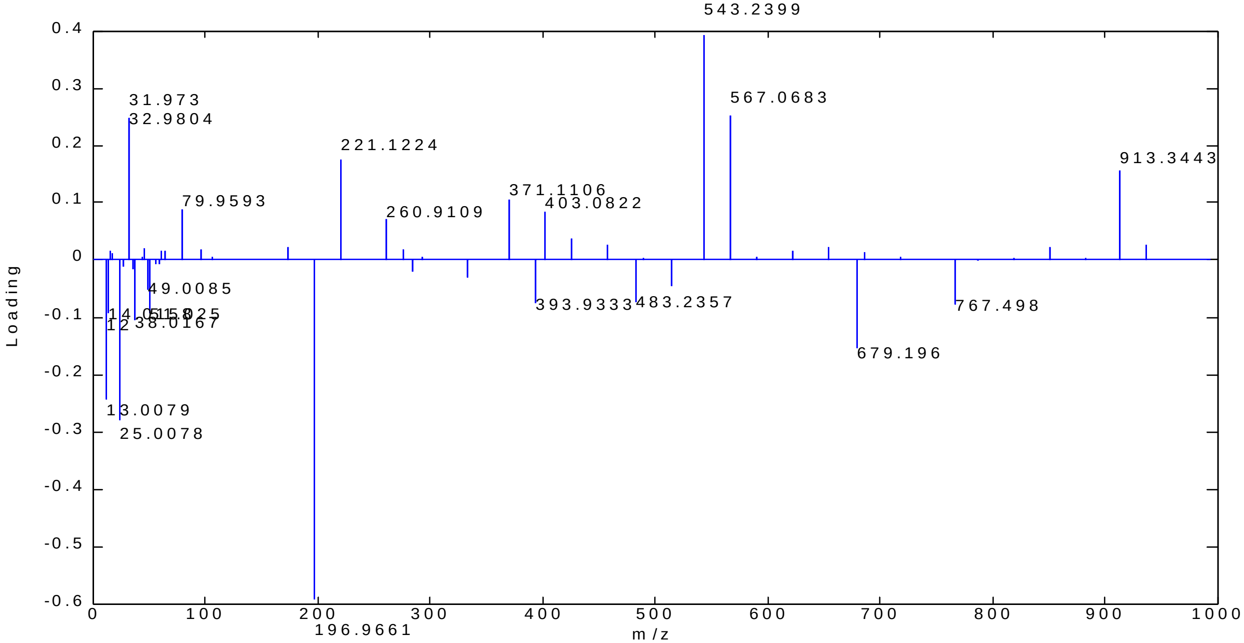
- Autoscaling
  - Divides mean centered data by standard deviation of each variable
  - Creates a data set where all variables vary between +1 and -1
- Non linear scaling
  - Log transformation
- Root Mean Scaling
- Square root scaling
- Optimal scaling

Example :Mixed C10 C18 SAMs



-Normalized Sum of Selected Peaks

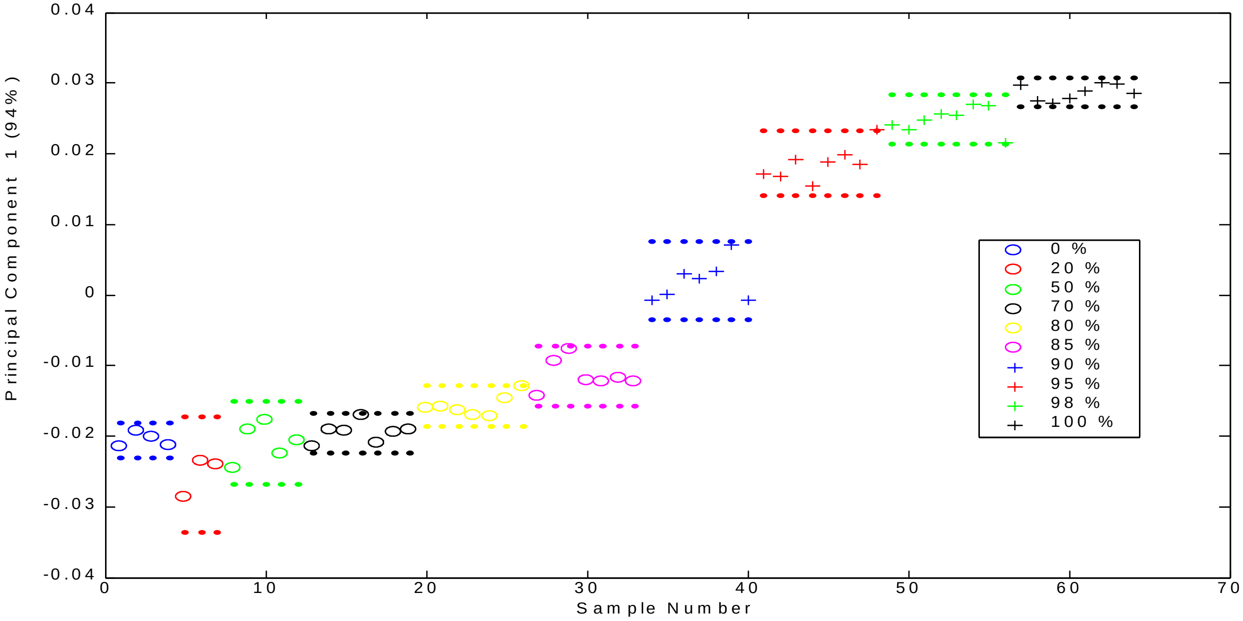
-Mean Centered

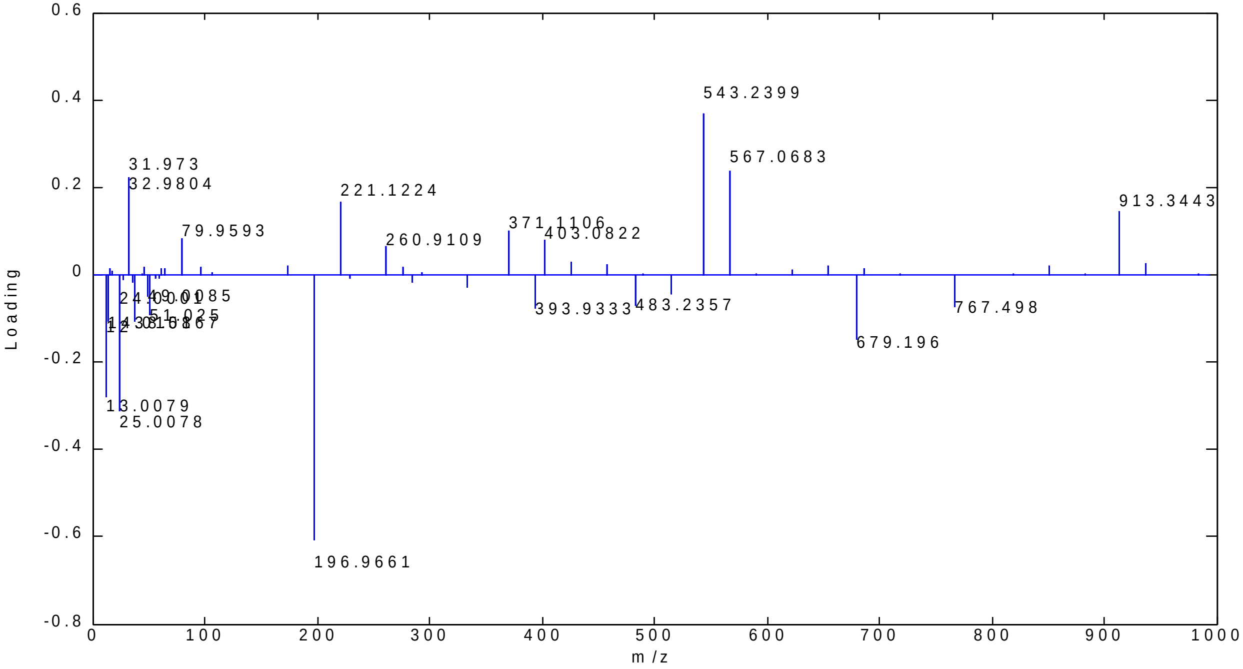


Example :Mixed C10 C18 SAMs

-Normalized Total intensity

-Mean centered



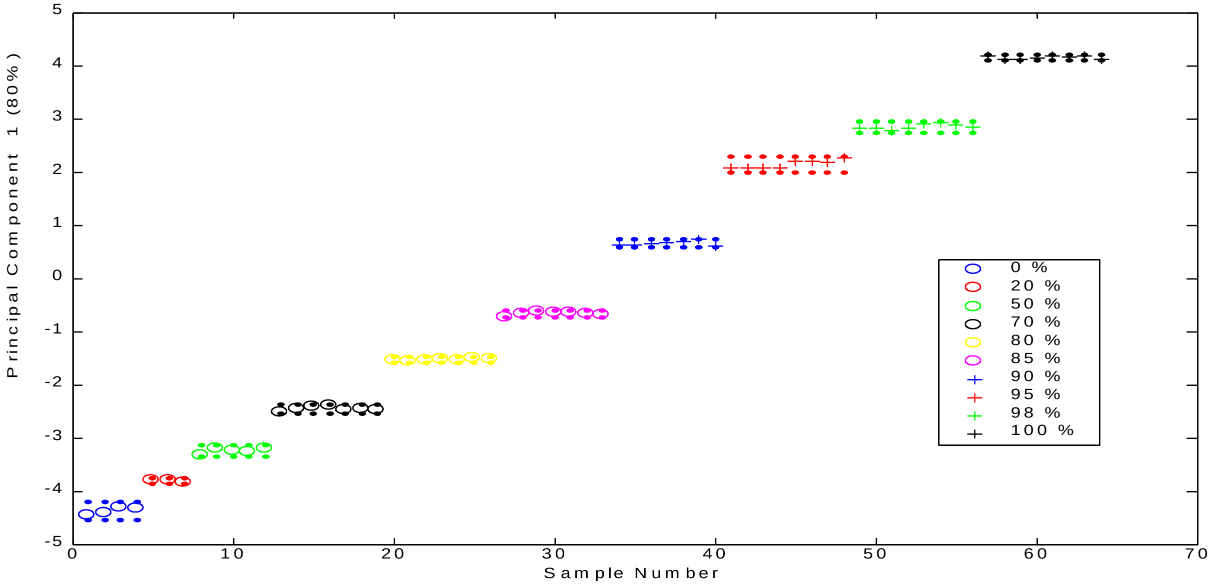


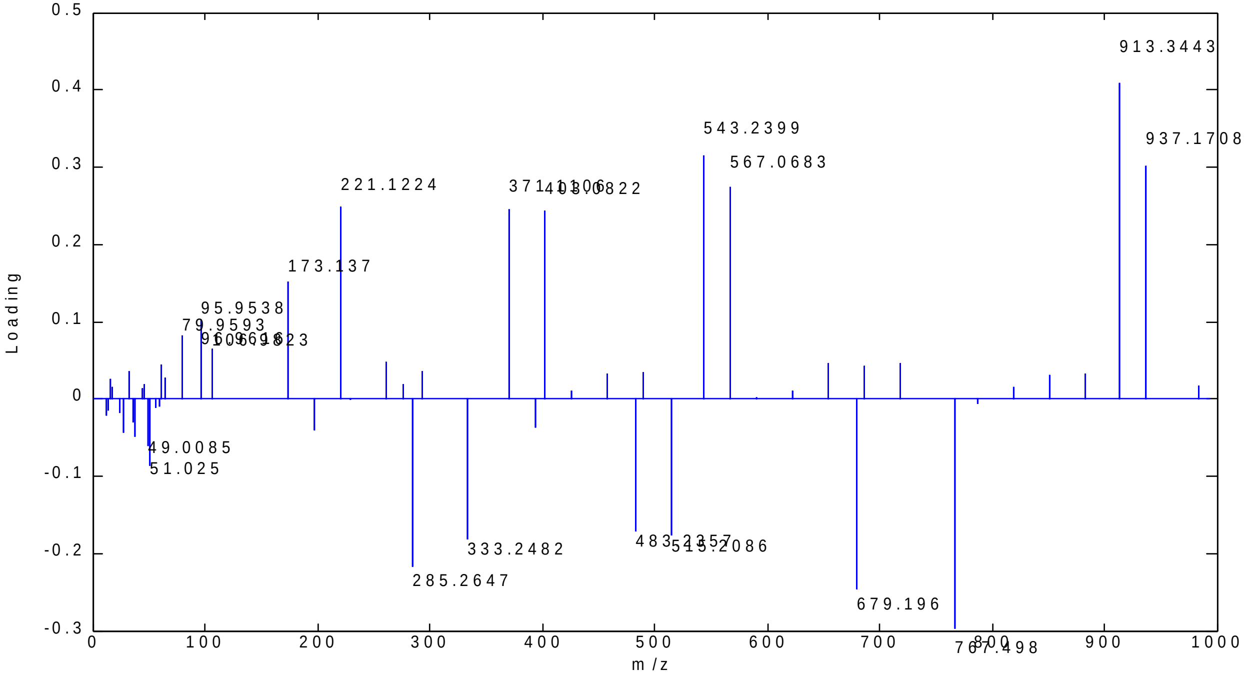
Example :Mixed C10 C18 SAMs

-Normalized Total intensity

-log10 transformed

-mean centered

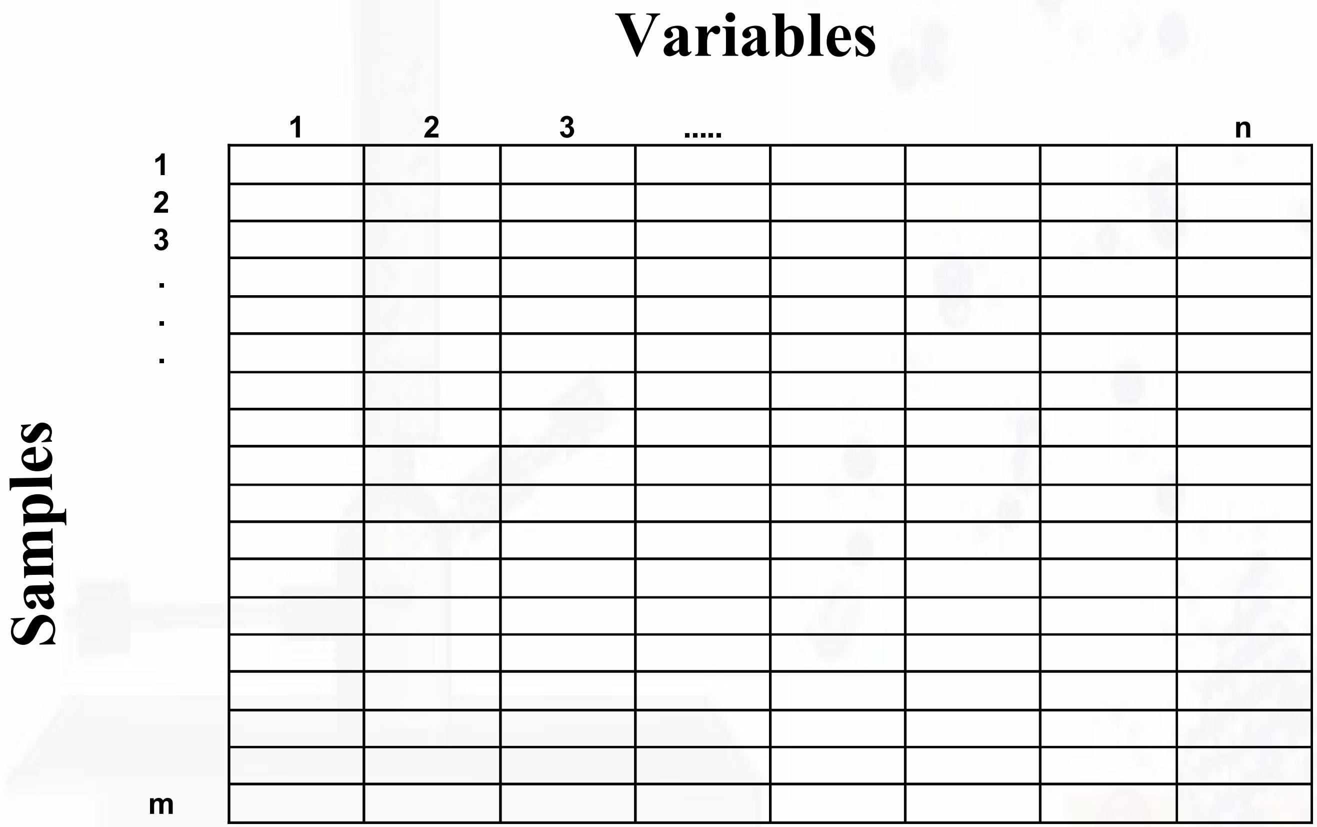




Running PCA

- Data organized in data matrix
- Data should be normalized before running PCA
- Choose appropriate data pretreatment
  - These are typically options in the PCA programs
- Run Program
- Extract the information
  - Scores
  - Loads
  - % variance captured for each PC

Data Matrix



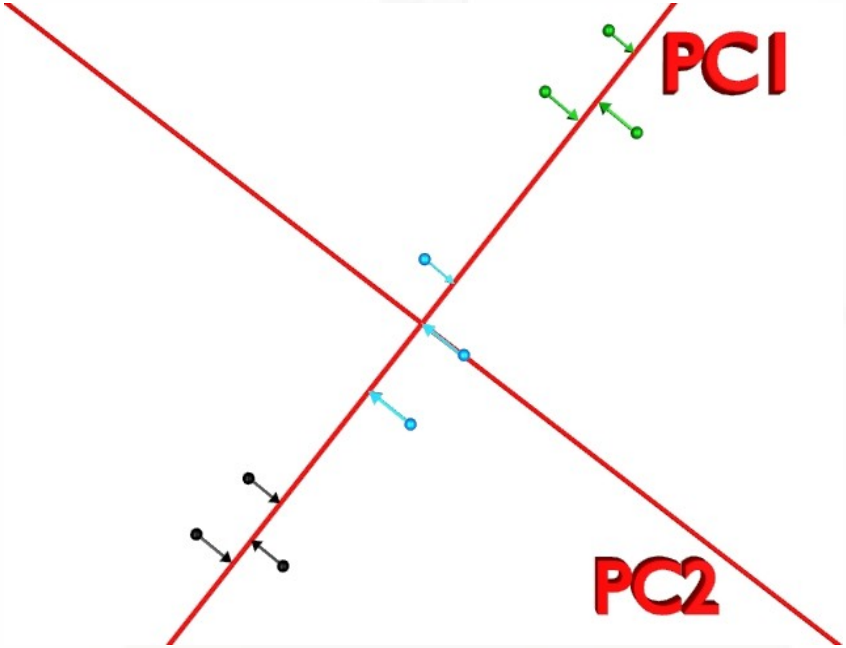
For SIMS data the “samples” are SIMS spectra, or more typically the integrated areas for all peaks for a given spectra

- For SIMS data, the “variables” are the peaks selected from the spectra
- If an entire spectrum is read in to a matrix then, the variables are the individual data bins

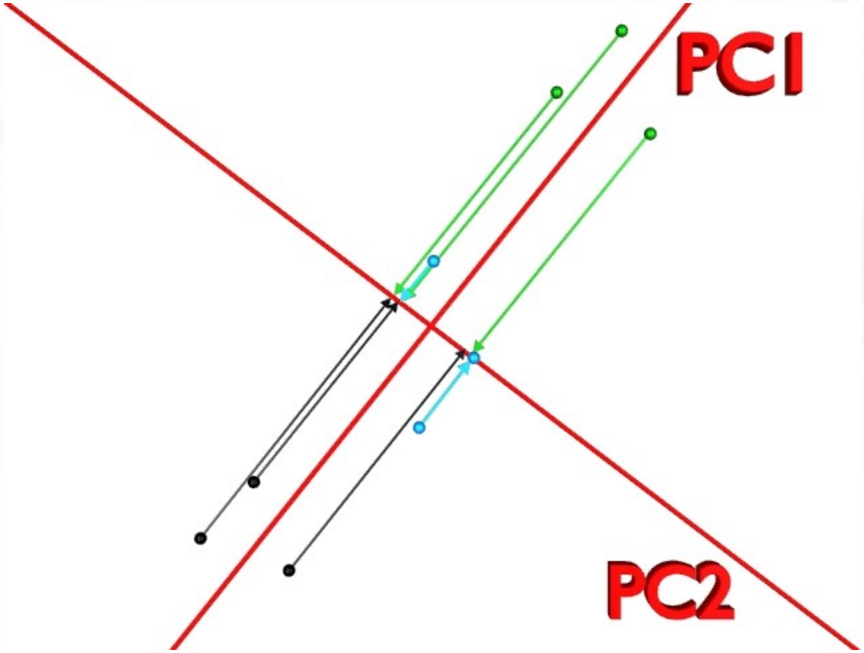
PCA Scores

The Scores are a projection of the samples onto the new PC axes

Scores tell the relationship (spread) between the samples



Projection onto PC1

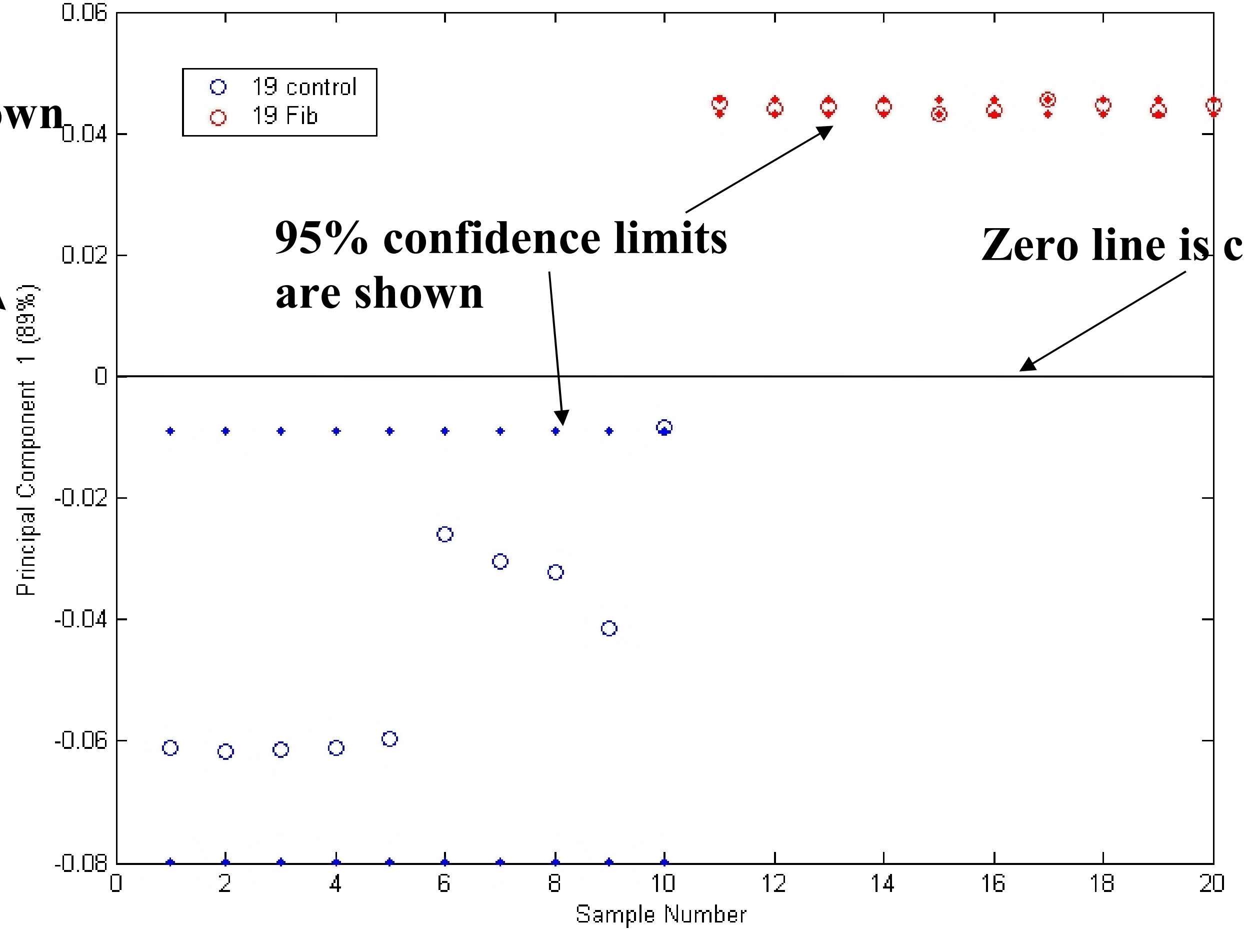


Projection onto PC2

Plotting Scores

- Plotting software may vary
- It is easiest to interpret data in 1 dimension at a time
  - Plot PC vs Sample
  - If samples vary in systematic way you can plot PC vs variable of interest
- Sometimes it is necessary to plot 2 PCs against each other to see sample separation
- Always show % variance captured for each PC
- Always show where zero is
- Use 95% confidence limits to show significance of sample separation

PCA Scores Example 1



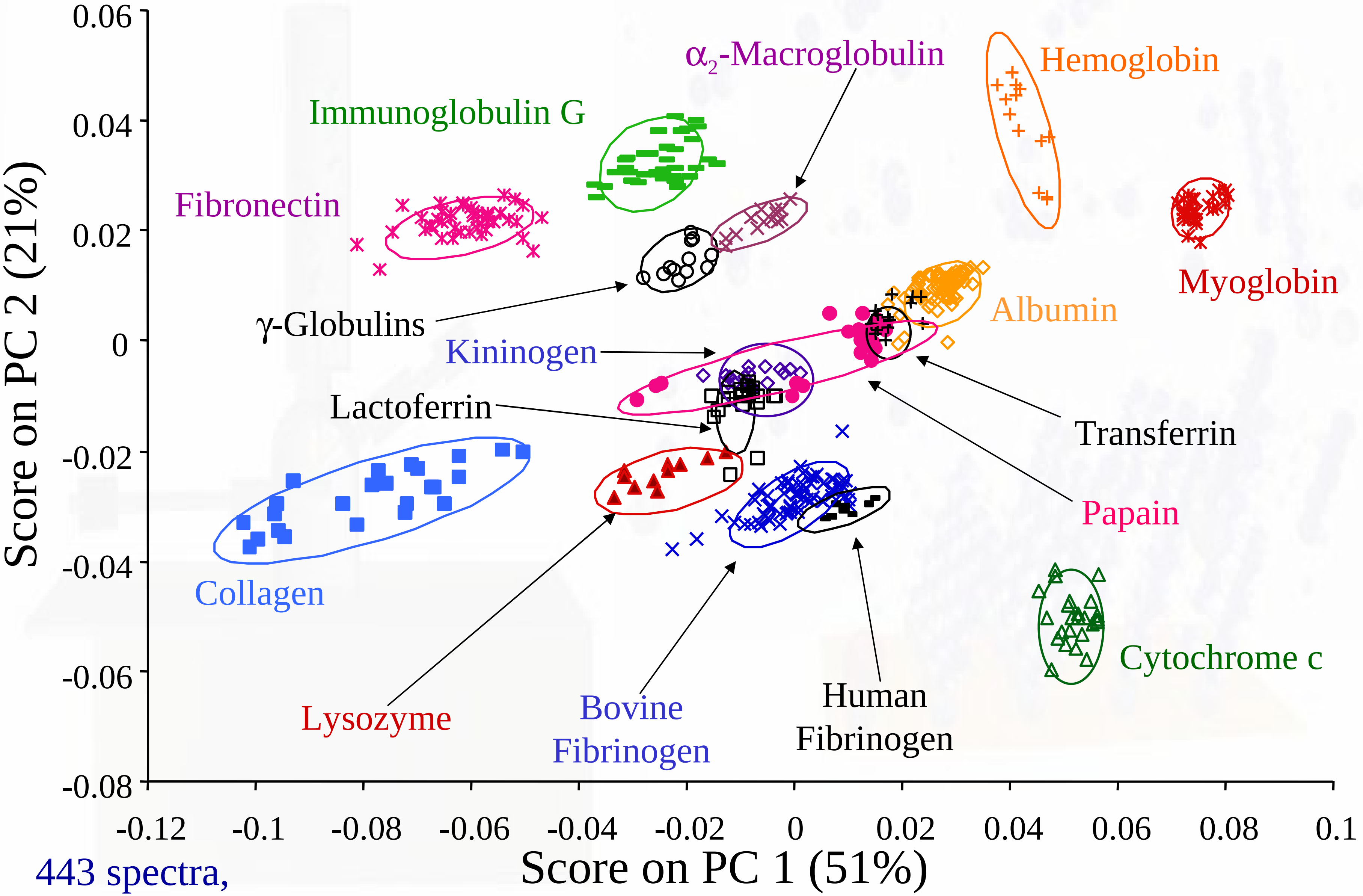
% Variance is shown

95% confidence limits are shown

Zero line is clearly shown

PC scores are plotted against sample number

PCA Scores Example 2

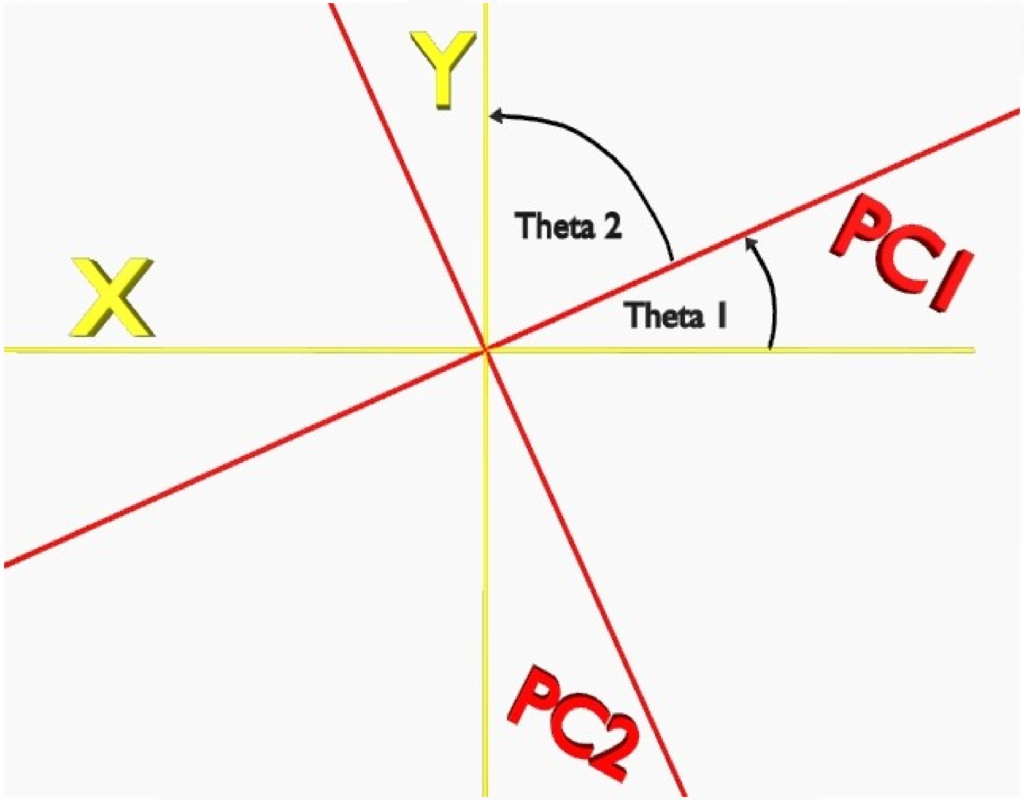


443 spectra,
16 different proteins

Wagner & Castner, Langmuir **17** (2001) 4649.

PCA Loadings

The loadings are the direction cosines between the new axes and the original variables



- **Cos(90) = 0**
  - **Large angle low loading**
- **Cos(0) = 1**
  - **Small angle high loading**

High Loading means that variable had a high influence on the separation of the samples

The loadings tell which variables are responsible for the separation seen between samples

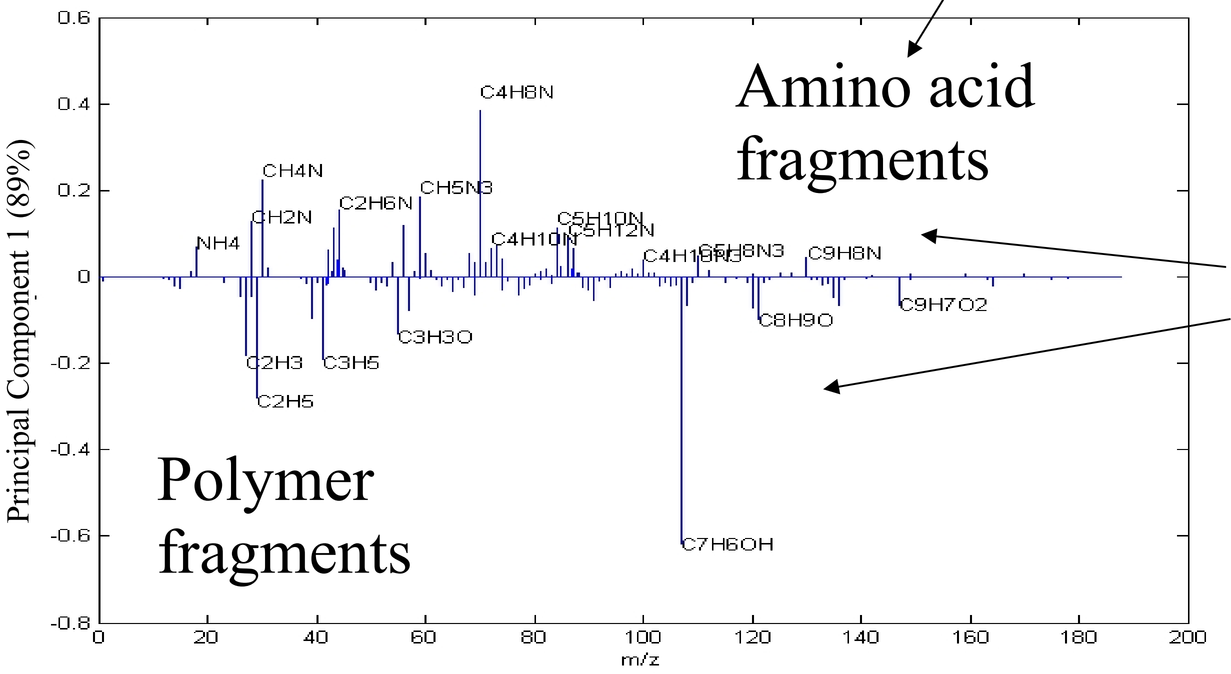
PCA Loadings

- Plotting software may vary
- It is easiest to interpret data in 1 dimension at a time
  - Plot PC vs m/z
    - This makes it so the loadings look more like a mass spectrum
- Always show % variance captured for each PC
- Only label highest loads to maintain clarity
  - You can explain other peak loadings in the text of your paper or report

PCA Loadings

% variance is shown

Descriptor is added to highlight major differences

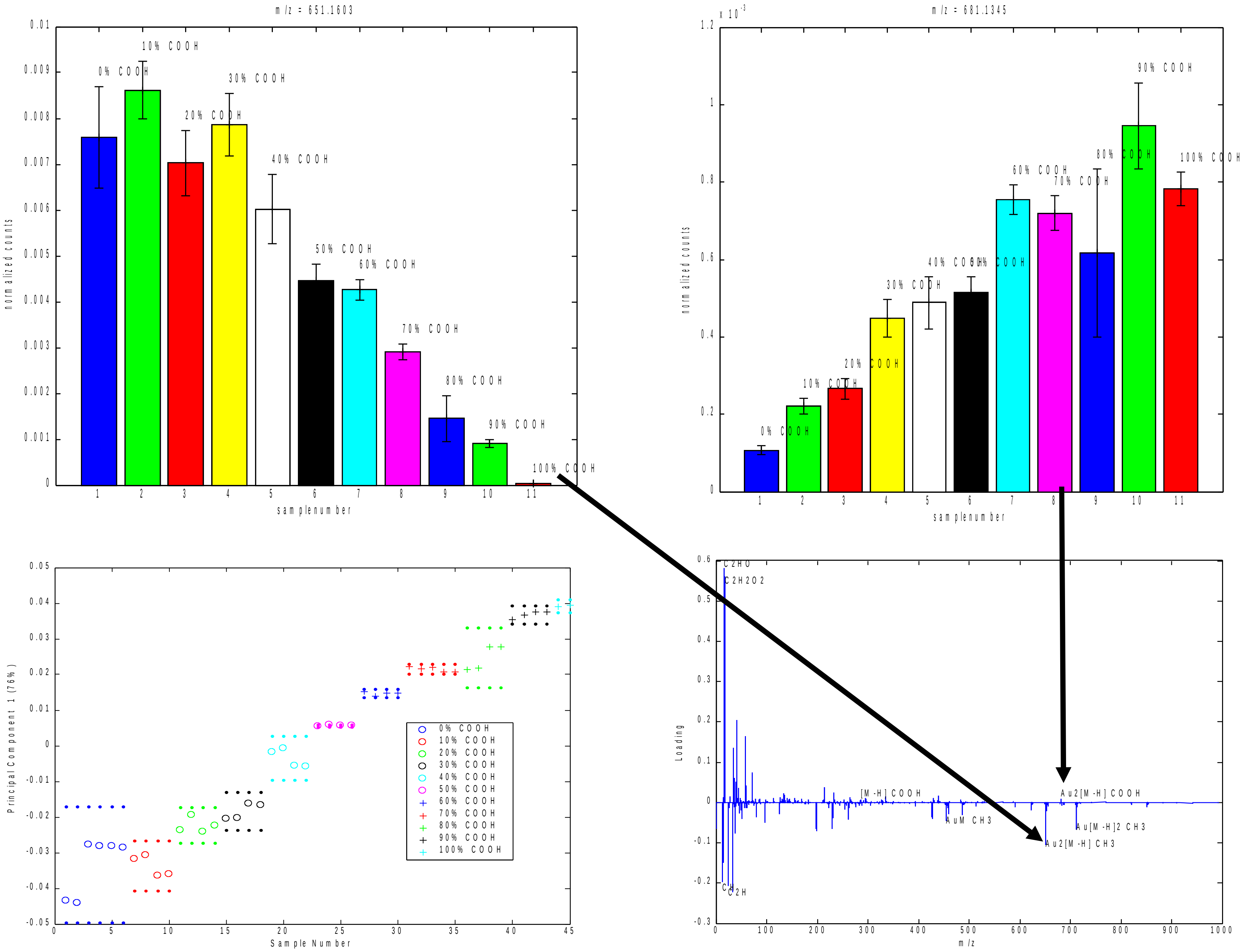


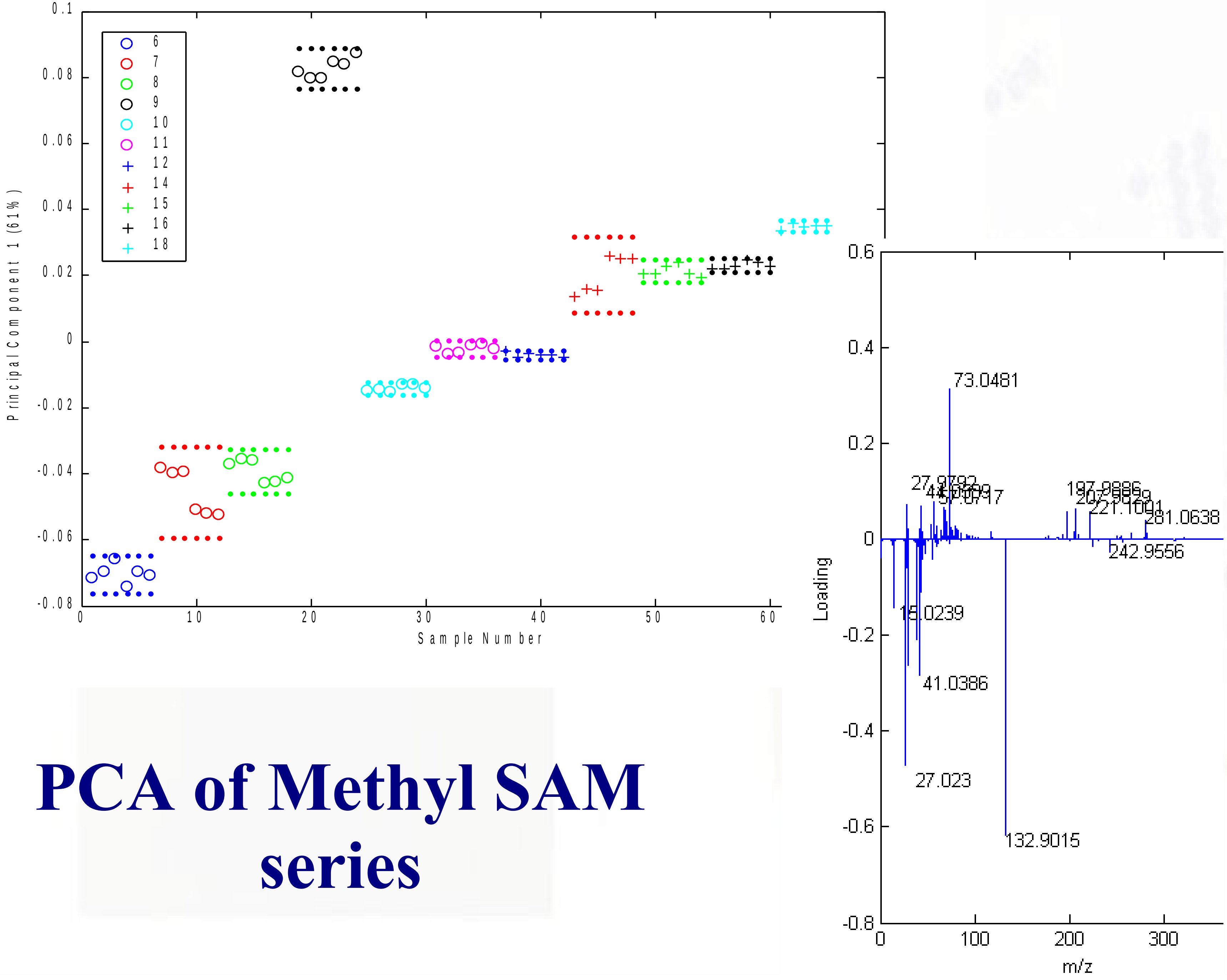
Highest Loadings are labeled

Loadings are plotted versus m/z

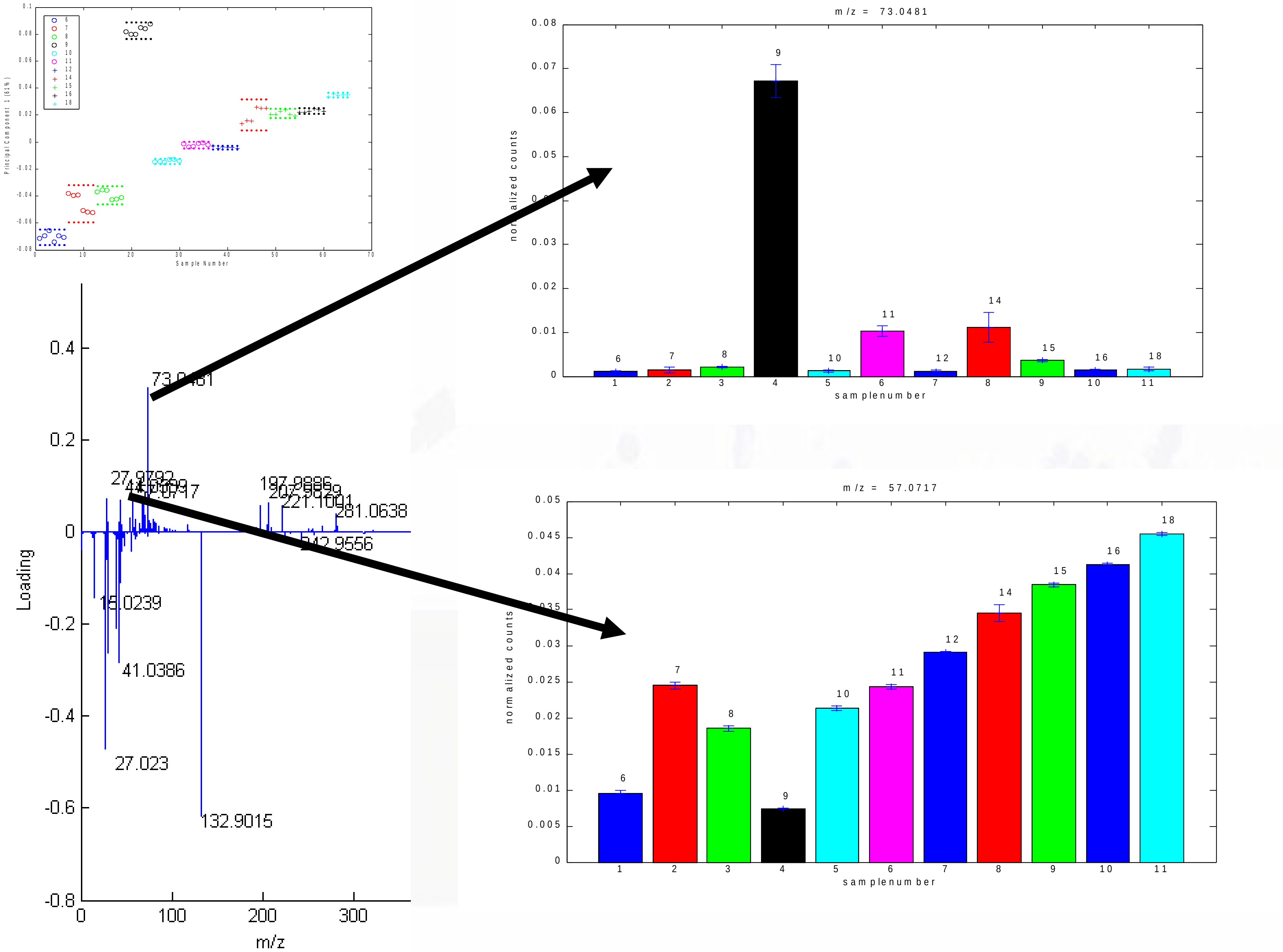
PCA: Interpretation

- Scores and Loadings are interpreted in Pairs
  - PC1 scores with PC1 loadings
  - PC2 scores with PC2 loadings
  - Etc...
- Samples with high positive scores on a given PC are positively correlated with variables with high positive loadings
- This means that in general samples with high positive scores on a given PC will have higher relative intensities for variables with high positive loadings on the same PC


m/z = 651.1603
normalized counts
0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
0% COOH
10% COOH
20% COOH
30% COOH
40% COOH
50% COOH
60% COOH
70% COOH
80% COOH
90% COOH
100% COOH
1
2
3
4
5
6
7
8
9
10
11
samplenumber
m/z = 681.1345
x 10^-3
1.2
1
0.8
0.6
0.4
0.2
0
normalized counts
0% COOH
10% COOH
20% COOH
30% COOH
40% COOH
50% COOH
60% COOH
70% COOH
80% COOH
90% COOH
100% COOH
1
2
3
4
5
6
7
8
9
10
11
samplenumber
Principal Component 1 (76%)
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
0
5
10
15
20
25
30
35
40
45
Sample Number
○ 0% COOH
○ 10% COOH
○ 20% COOH
○ 30% COOH
○ 40% COOH
○ 50% COOH
+ 60% COOH
+ 70% COOH
+ 80% COOH
+ 90% COOH
+ 100% COOH
Loading
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
C2HO
C2H2O2
CH
C2H
[M-H]COOH
AuMCH3
Au2[M-H]COOH
Au[M-H]2CH3
Au2[M-H]CH3
0
100
200
300
400
500
600
700
800
900
1000
m/z



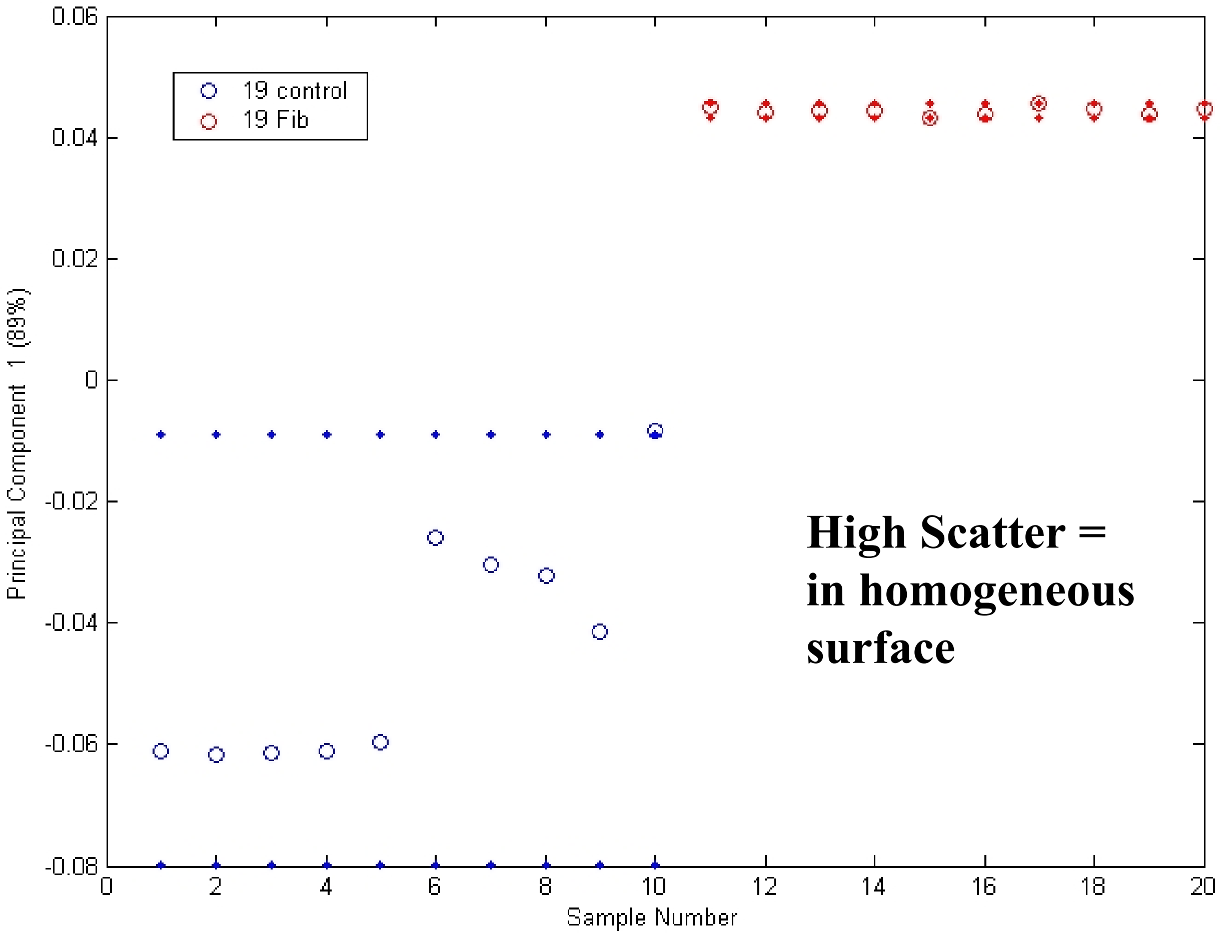
PCA of Methyl SAM series


Principal Component 1 (61%)
Sample Number
6
7
8
9
10
11
12
14
15
16
18
Loading
m/z
73.0481
57.0717
15.0239
41.0386
27.023
132.9015
197.0886
221.1001
281.0638
342.9556
m/z = 73.0481
m/z = 57.0717
normalized counts
samplenumber

PCA Interpretation Continued

- Scores Plots
  - Samples with similar scores are similar (clustered together)
  - Samples with very different scores are different (separated from each other)
- Scatter in the scores for a given sample type suggests inhomogeneities in the sample
- Tight grouping of scores for a given sample suggests a homogeneous surface

PCA Scores Grouping



Low Scatter = homogeneous surface

High Scatter = in homogeneous surface

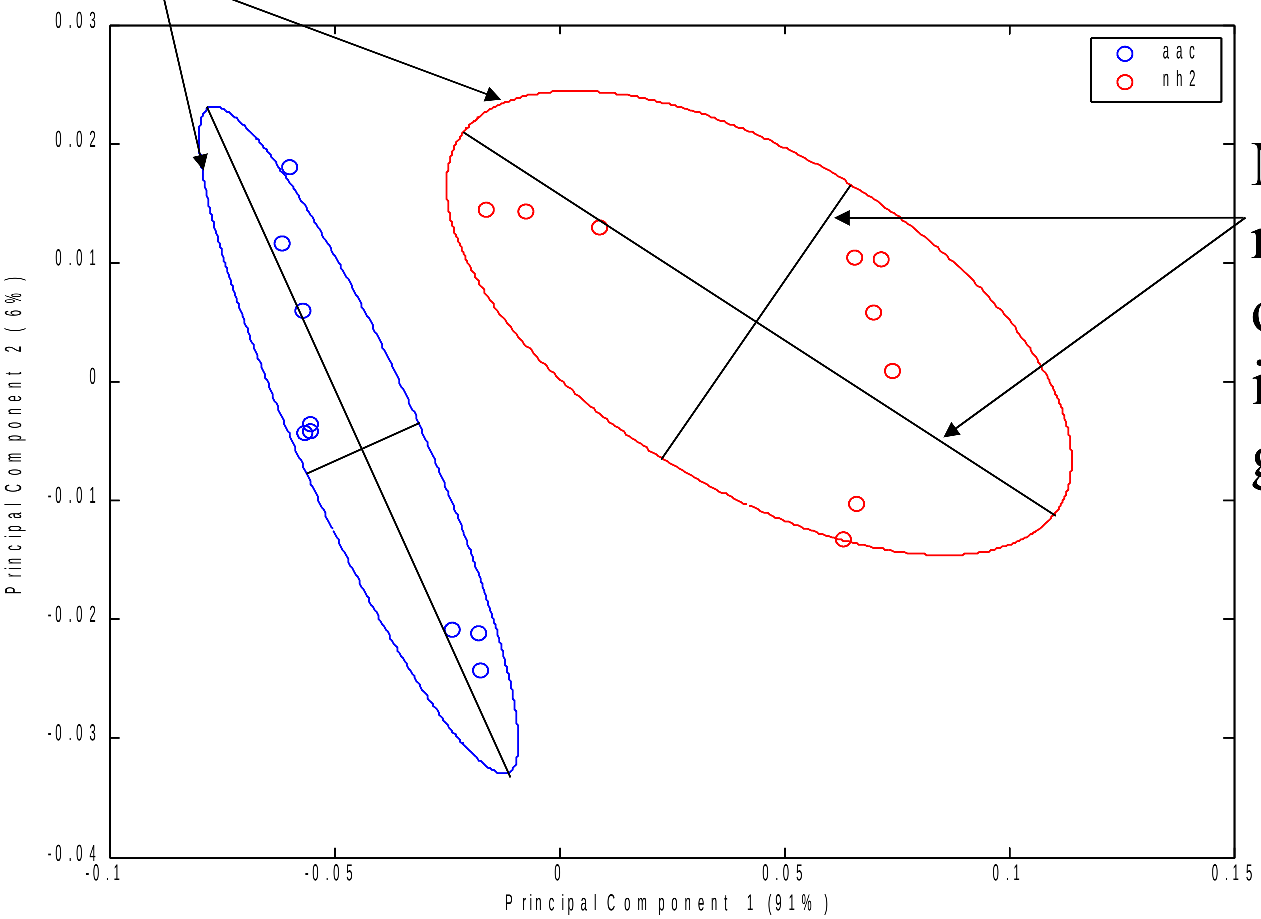
95% Confidence Limits

- Scores are assumed to follow a normal distribution
- t-distribution can be used to calculate confidence limits for a subgroup of scores
  - run PCA on subgroup
  - use eigenvalues from subgroup PCA to determine confidence limits
  - use loadings to rotate confidence limits back to original PC plot
- Shows bounds of groups on PC plots

Wagner, Castner Langmuir 17, 2001, 4649-4660

95% confidence limits

Confidence ellipses

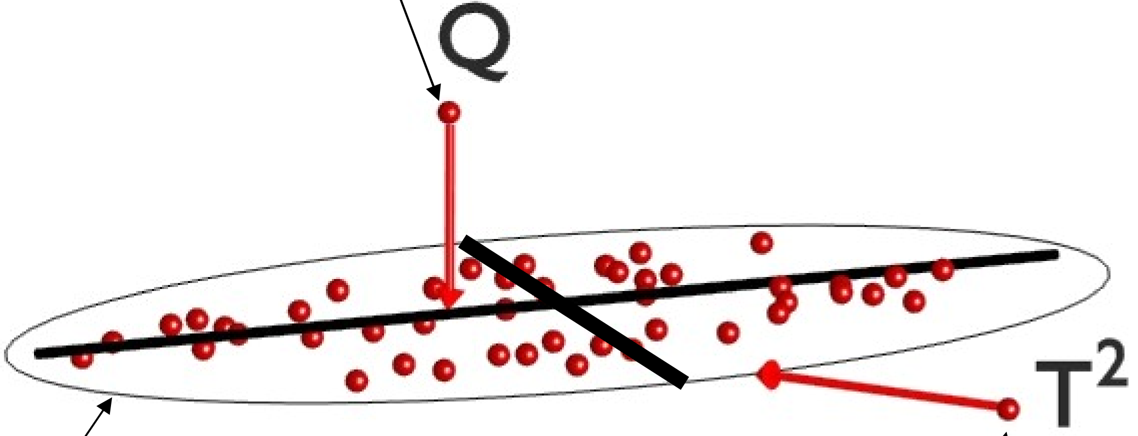


Major and minor axes of scores for individual group

Q and Hotelling's T^2 – Outlier detection

- Q - Variation outside of model
  - sqrt(Q) = euclidean distance from model
- T^2 – Variation inside the model

Sample with high variation outside of model – Large Q



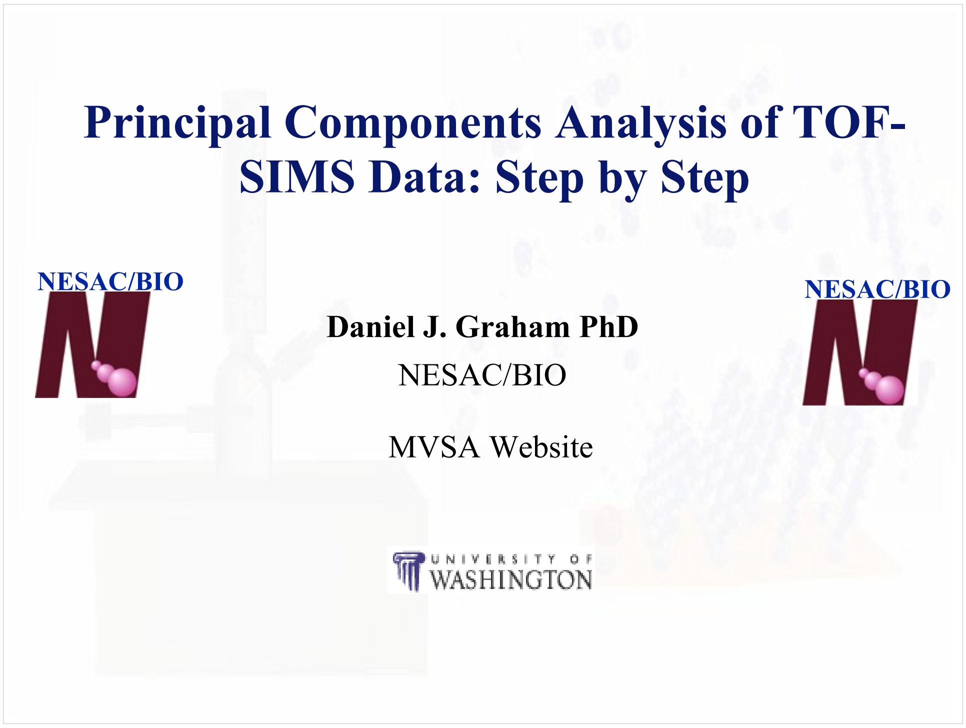
Line of constant T^2

Sample with high variation inside of model – Large T^2

Jackson, J.E. A User's Guide to Principal Components: John Wiley & Sons: New York 1991

Conclusions

- PCA has great potential to aid in spectral interpretation and analysis
  - can aid in determining sample differences
  - requires well thought out experiments
  - cannot do analysis for you
- Plan your experiments with a central question and minimize the number of variables
  - This can greatly simplify the interpretation
  - Can maximize what you get out of your data



Principal Components Analysis of TOF-SIMS Data: Step by Step

NESAC/BIO

NESAC/BIO

Daniel J. Graham PhD

NESAC/BIO

MVSA Website

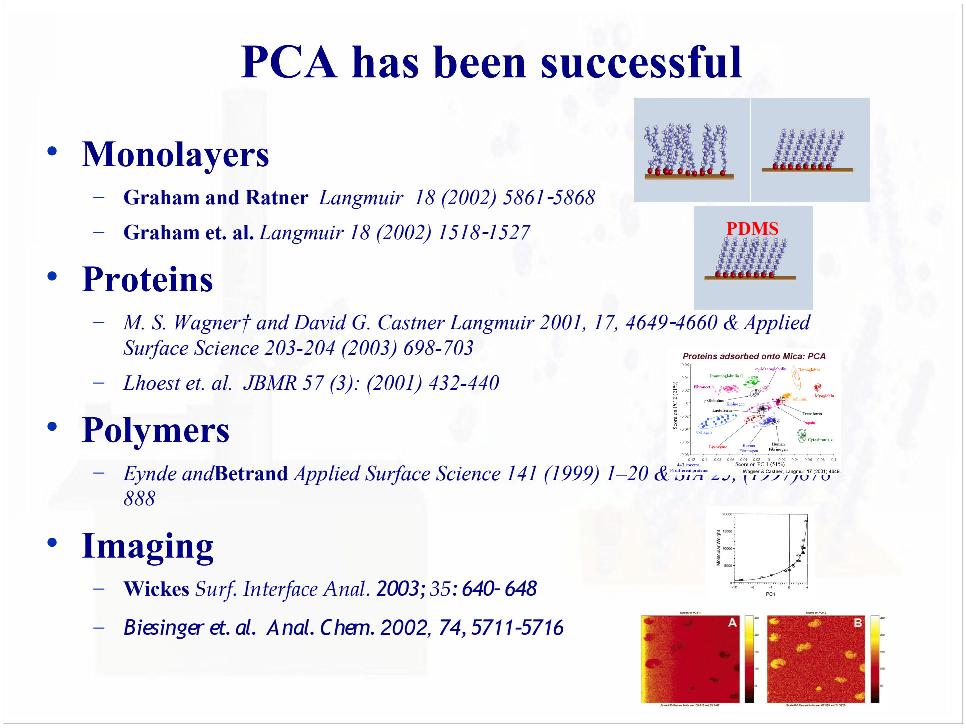
UNIVERSITY OF WASHINGTON

The Successful application of PCA to TOF-SIMS data come with careful planning, execution and interpretation of experiments. PCA is not magic. It cannot plan experiments for you or interpret your data for you. It is a tool that can enable and facilitate the interpretation of TOF-SIMS data.

Plans to use PCA should start in the initial stages of your experimental plan. The samples analyzed, number of samples, and number of spectra per sample should be carefully considered in order to maximize the possibility of getting good results from PCA. Since PCA captures the major directions of variation within a data set, it is important to plan your experiment so that the differences seen within your spectra can be attributed to the differences in the sample chemistry or structure. One good way to do this is to only allow 1 variable to change during each experiment. In the absence of contaminants or matrix affects, this type of single variable experiment will allow monitoring the changes in the fragmentation patter of the data due to this variable.

The use of proper controls is also important. PCA looks for differences in the data, so a well designed control can provide clear separation of treated and untreated samples.

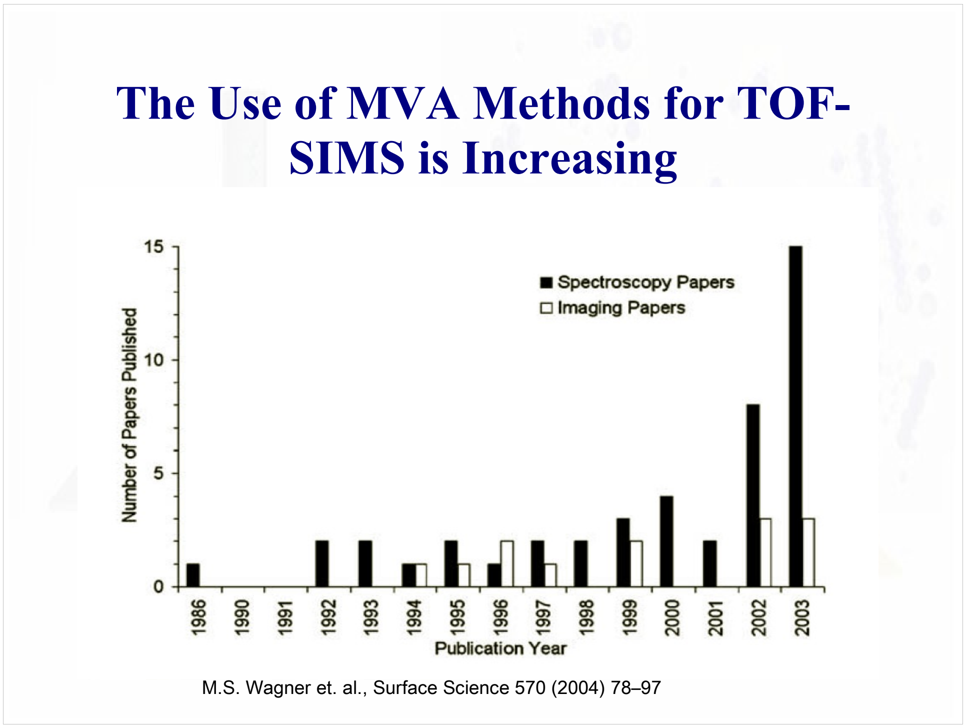
Finally, careful interpretation of the PCA results is also critical. It is important to remember that PCA can determine trends in the data, but it cannot predict causality.



PCA has been successful

- **Monolayers**
  - **Graham and Ratner** *Langmuir 18 (2002) 5861-5868*
  - **Graham et. al.** *Langmuir 18 (2002) 1518-1527*
- **Proteins**
  - *M. S. Wagner† and David G. Castner Langmuir 2001, 17, 4649-4660 & Applied Surface Science 203-204 (2003) 698-703*
  - *Lhoest et. al. JBMR 57 (3): (2001) 432-440*
- **Polymers**
  - *Eynde and***Betrand** *Applied Surface Science 141 (1999) 1–20 & SIA 25, (1997)878-888*
- **Imaging**
  - **Wickes** *Surf. Interface Anal.* *2003; 35: 640- 648*
  - *Biesinger et. al. Anal. Chem. 2002, 74, 5711-5716*

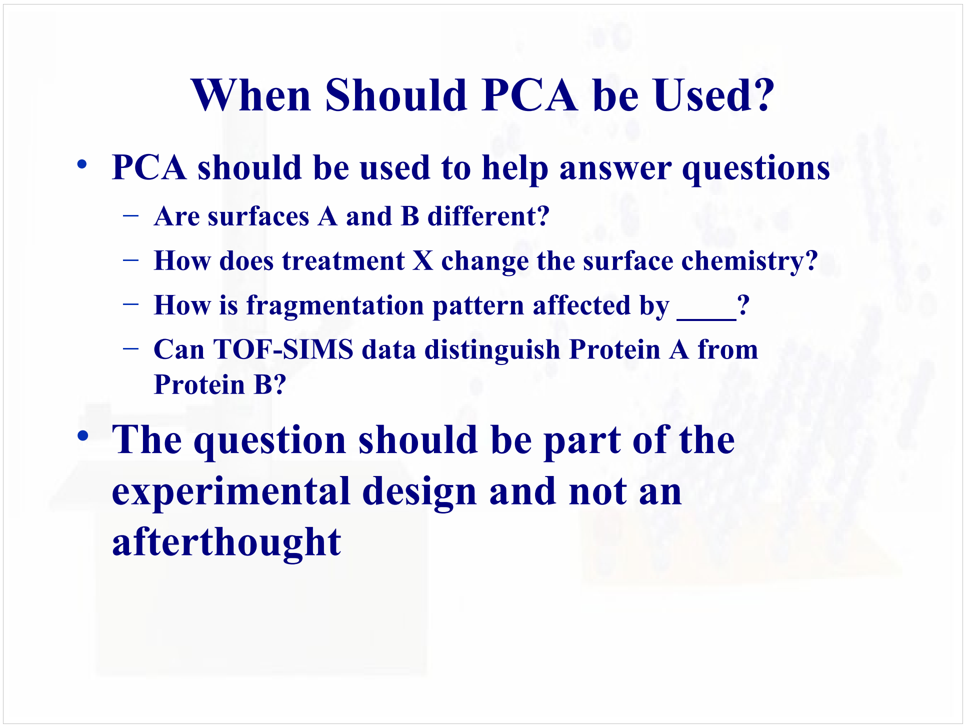
PCA has been successfully applied to many different types of samples. This slide gives a small sampling of the different sample systems and a few references. A more extensive list can be seen on the references page of this website.



The Use of MVA Methods for TOF-SIMS is Increasing

M.S. Wagner et. al., Surface Science 570 (2004) 78–97

The success of these initial studies in the application of multivariate methods to TOF-SIMS data has resulted in a significant increase in the number of publications where multivariate methods have been used. The number of publications has changed from 1 or two a year to more than 15 in only a few years. This number is bound to increase as more researchers realized the potential of these methods to aide in data interpretation. As this number rises it is also important that the users of multivariate methods understand how to properly apply there methods and interpret the results from their data.

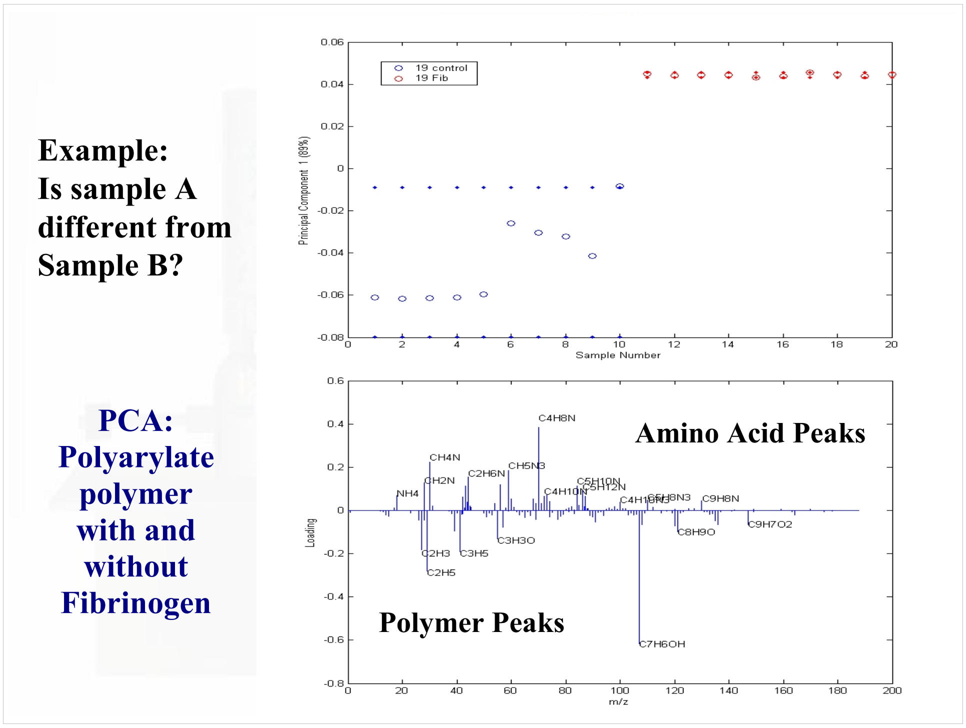


When Should PCA be Used?

- PCA should be used to help answer questions
  - Are surfaces A and B different?
  - How does treatment X change the surface chemistry?
  - How is fragmentation pattern affected by ____?
  - Can TOF-SIMS data distinguish Protein A from Protein B?
- The question should be part of the experimental design and not an afterthought

Many sample systems are ideal candidates for multivariate methods such as PCA. PCA is useful to answer questions such as, why is sample A different from sample B? Or how does treatment 'X' change the chemistry of a surface. These types of questions have well defined statement and lend themselves to hypothesis driven research. If treatment 'X' is an oxidative process, one could hypothesize that application of treatment 'X' to a surface will cause an increase of oxygen containing groups on a surface. PCA of the data from these samples could then be used to track the changes in these samples and verify or negate the hypothesis based on the differences seen in the data.

When using PCA the question to be answered should be used as part of the experimental design and not just an afterthought.

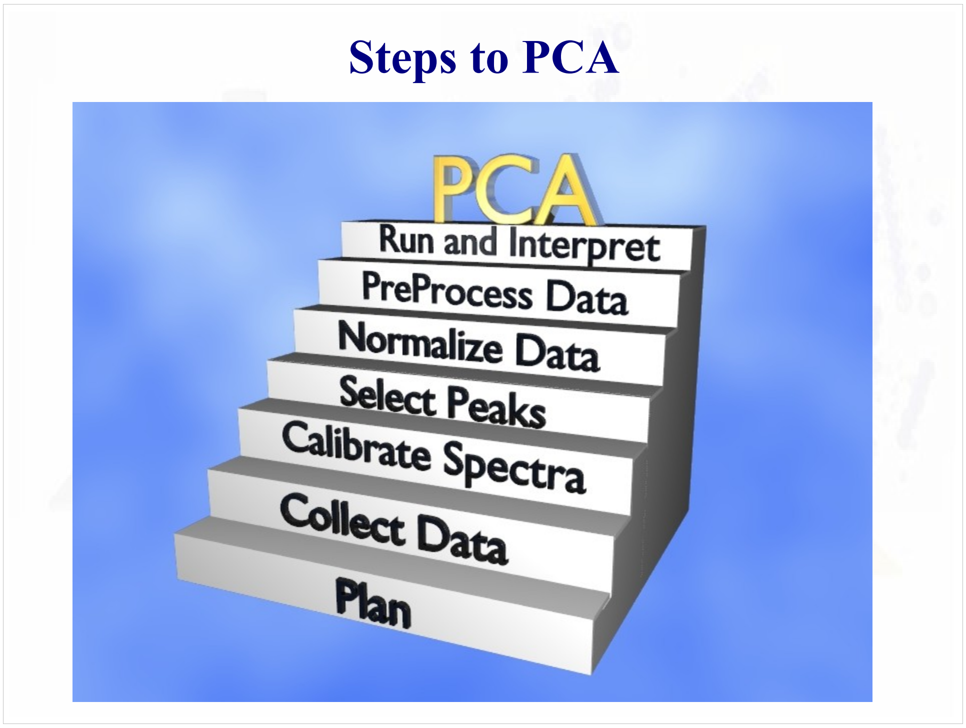


Example:
Is sample A different from Sample B?

PCA: Polyarylate polymer with and without Fibrinogen

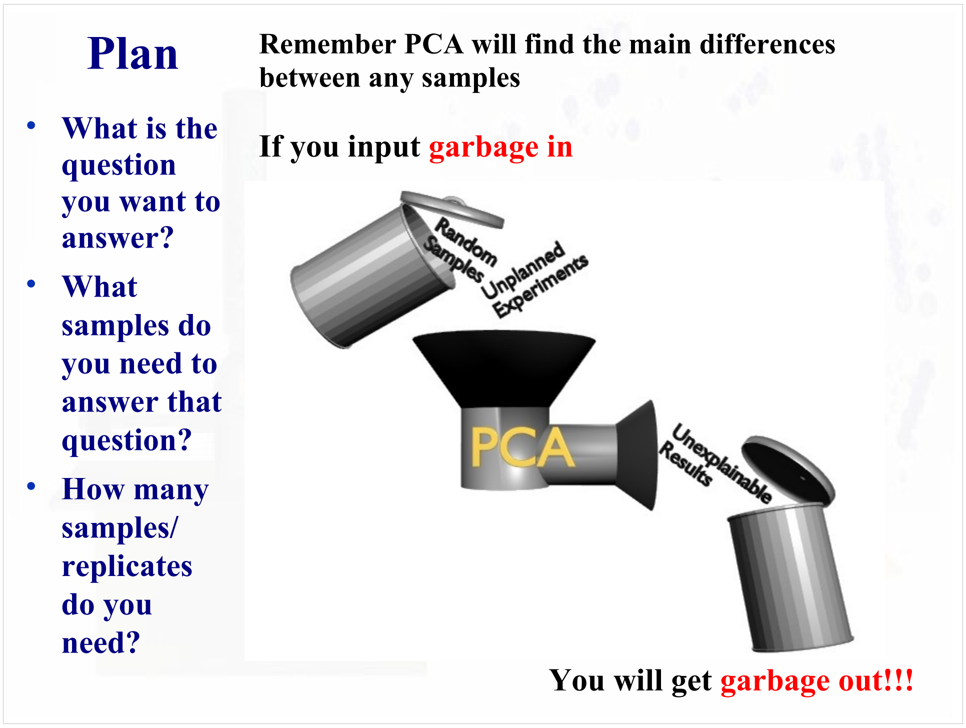
This slide illustrates a quick example of how one sample differs from another. The figures show the scores and loadings plots from a polyarylate polymer with and without adsorbed fibrinogen. The upper figure shows the PC1 scores plot with the scores plotted against the sample number. As seen in the figure the two samples are clearly separated as noted by the separation of the scores for the spectra of each sample type. The PC1 loadings show that the separation of these samples is due to the present of amino acid peaks on the sample with fibrinogen and polymer peaks for the bare polymer surface.

Though this is a simple example, it shows how PCA can be used to quickly determine sample differences between a set of samples. This can be done without a priori knowledge of the sample set.



Steps to PCA

There are several steps that should be taken in order to successfully use PCA with TOF-SIMS data. These steps are outlined in the figure above. Each of these steps are important. Each one will be explained in more detail in the slides that follow.



Plan

- What is the question you want to answer?
- What samples do you need to answer that question?
- How many samples/ replicates do you need?

Remember PCA will find the main differences between any samples

If you input garbage in

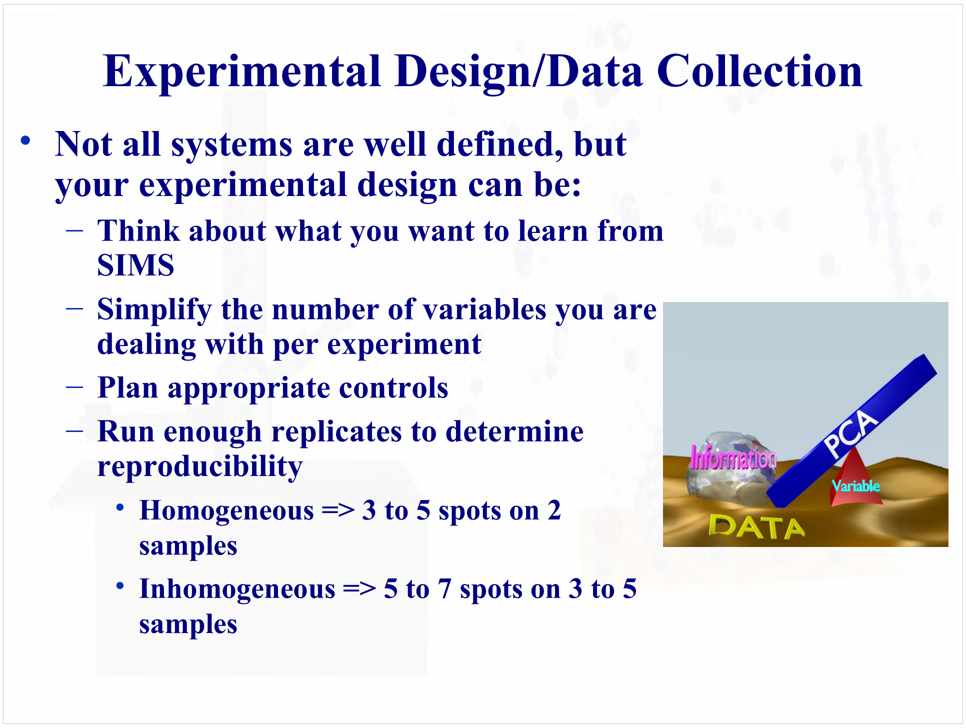
Random Samples
Unplanned Experiments
PCA
Unexplainable Results

You will get garbage out!!!

As has been mentioned, planning is one of the most important parts of the PCA process. Before collecting any data you should know what the question is that you are trying to answer. Once the question has been determined and your hypothesis about this question has been formed, you should then determine what samples are necessary in order to answer that question. You should also consider how many replicates of each sample you will need.

It is important to note that since PCA looks for the largest directions of variance in a data set if you do not plan well, or if your samples contain contaminants, then the results obtained from PCA may not make any sense or will be very difficult to interpret.

PCA definitely holds to the adage of “garbage in, garbage out”.



Experimental Design/Data Collection

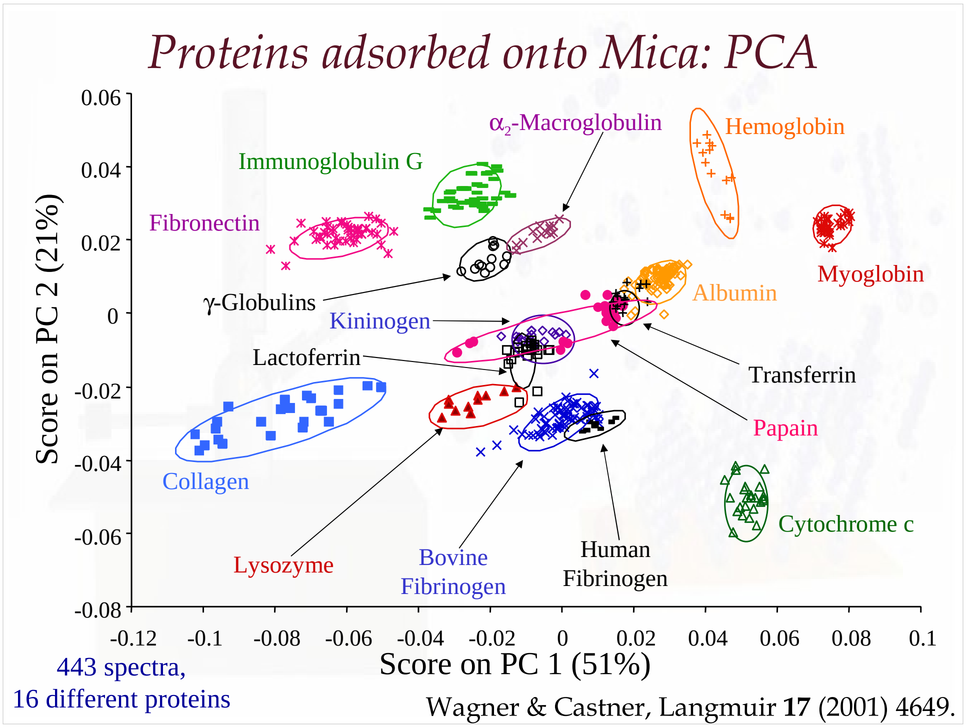
- Not all systems are well defined, but your experimental design can be:
  - Think about what you want to learn from SIMS
  - Simplify the number of variables you are dealing with per experiment
  - Plan appropriate controls
  - Run enough replicates to determine reproducibility
    - Homogeneous => 3 to 5 spots on 2 samples
    - Inhomogeneous => 5 to 7 spots on 3 to 5 samples

Good experimental design can make the difference between confusion and enlightenment. As mentioned before ideal experimental designs are those that only allow 1 variable to change at a time. When only 1 variable is changed across a data set, you can then use PCA to 'lever' out the gems of information from within the sea of TOF-SIMS data (this is illustrated graphically in the figure).

One important consideration in using PCA for TOF-SIMS data is the number of replicates necessary to determine the reproducibility of the samples and any differences between sample sets. The slide above illustrates some general guidelines for the number of replicates and spectra needed.

For homogeneous samples it is typically adequate to take 3 to 5 spectra across 2 samples of a given chemistry for PCA (total of 6 to 10 spectra per sample type). This should give enough data to determine sample differences with statistical significance.

For samples that are not as homogeneous or that tend to have a lot of spot to spot variability, it is recommended to take 5 to 7 spectra on 3 to 5 replicates (a total of 15 to 21 spectra per sample type).



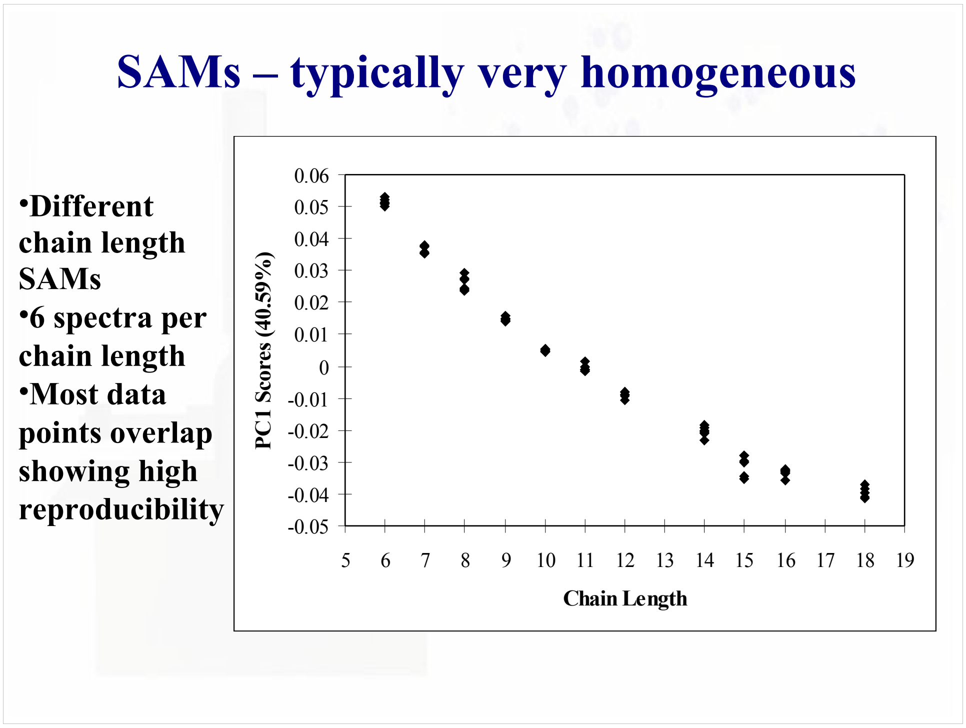
Proteins adsorbed onto Mica: PCA

443 spectra,
16 different proteins

Wagner & Castner, Langmuir **17** (2001) 4649.

This slides illustrates why non-homogeneous samples require more spectra per sample. The figure shows the PC1 vs PC2 scores for a series of 16 different proteins adsorbed onto mica. The ellipses shown are the 95% confidence intervals for each protein. As seen in this scores plot, there is significant scatter among the different protein samples. If only a few spectra were acquired for each protein, the clustering of the protein types and significance of the separation would not have been as clear.

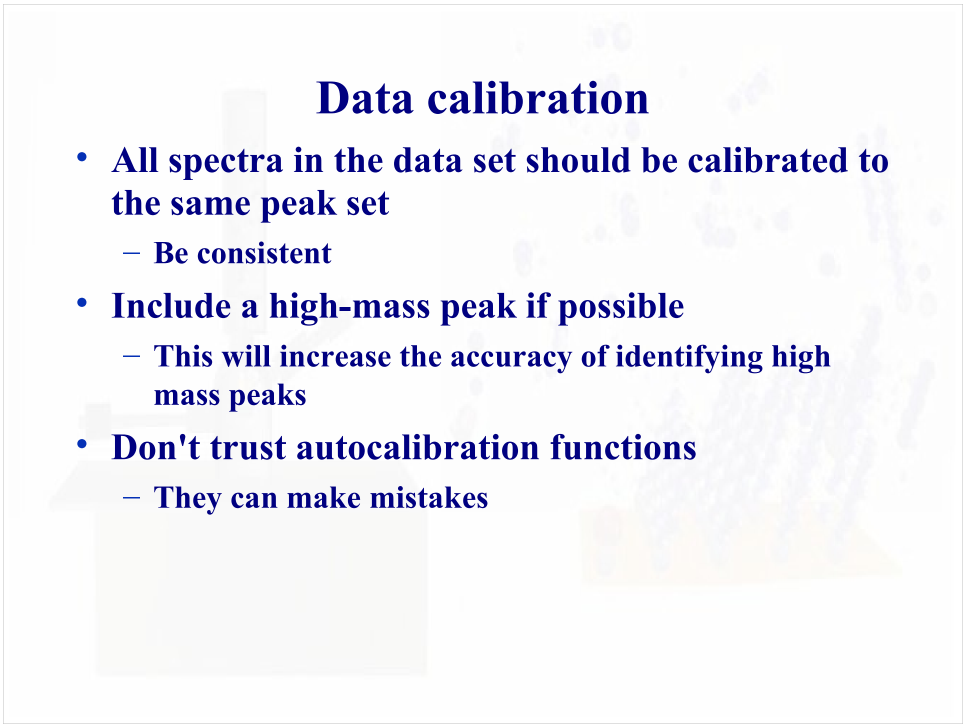
With todays computers and instrumentation taking this volume of data is well within reason.



SAMs – typically very homogeneous

- •Different chain length SAMs
- •6 spectra per chain length
- •Most data points overlap showing high reproducibility

This slides shows an example of PCA from a set of homogeneous samples. The figure shows the PC1 scores plotted against the chain length for a set of different length alkane thiols self-assembled monolayers on gold. 6 spectra were taken for each chain length across 2 samples. As seen in the figure for many of the chain lengths the diamonds representative of each spectra overlap significantly. This means they have very similar scores values (low variance), which also means the spectra were very reproducible spot to spot.



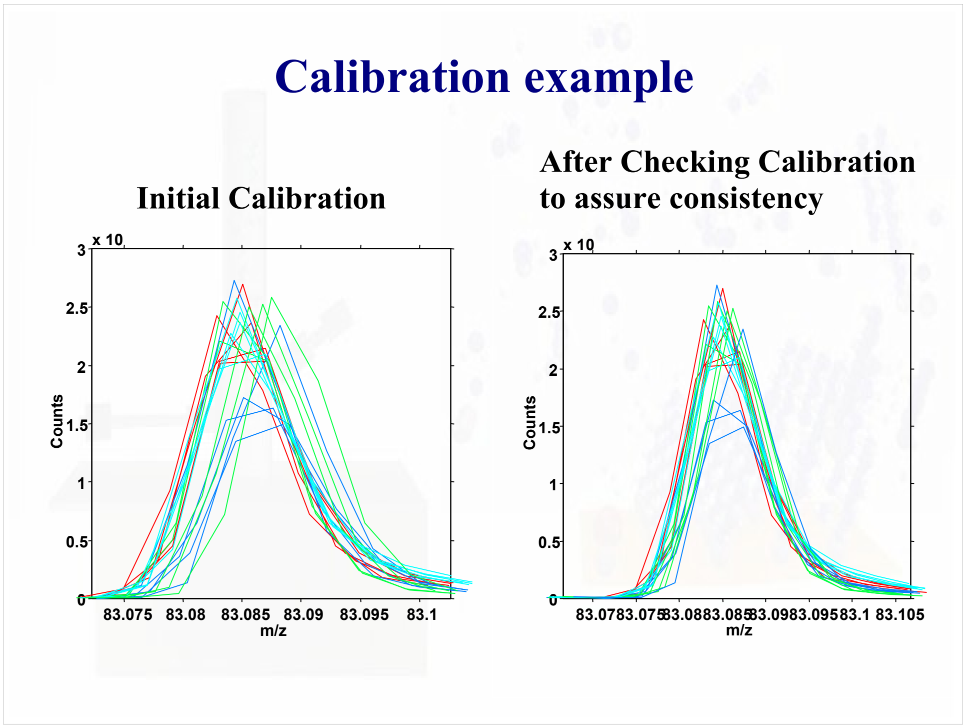
Data calibration

- All spectra in the data set should be calibrated to the same peak set
  - Be consistent
- Include a high-mass peak if possible
  - This will increase the accuracy of identifying high mass peaks
- Don't trust autocalibration functions
  - They can make mistakes

Once the data has been collected, it must first be calibrated before applying PCA. Calibration is included as a step to successful PCA because it is important that all the spectra within a sample set are calibrated properly and in the same way.

To aide in the accuracy of high-mass peak identification it is important to include a high mass peak in the calibration. Of course it is important to know the identity of any peak used in a calibration set. You cannot just guess.

To be most accurate, calibration should be done by hand. Autocalibration routines often do not work very well. Calibration should be verified by checking the spectra. This is illustrated on the next slide.



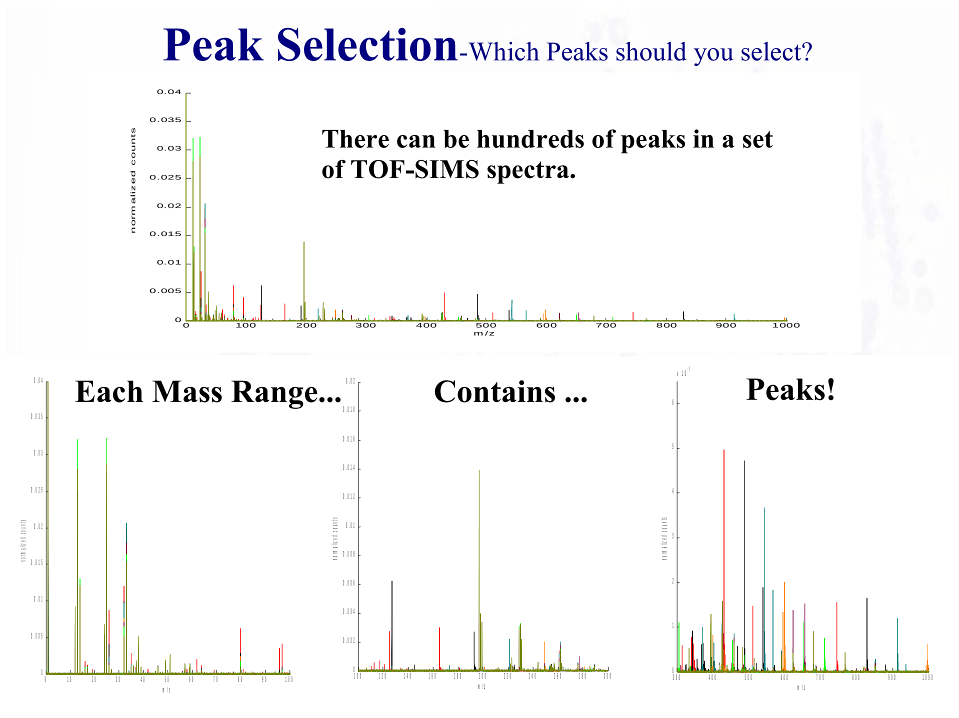
Calibration example

Initial Calibration

After Checking Calibration to assure consistency

This slide shows an overlay plot of several spectra that have all been calibrated with the same peak set, using the same criterion of keeping the error in the calibration below 10ppm. As seen in the figure on the left, even though the spectra were all calibrated in the same way, there is significant scatter in the peak positions. After rechecking the calibration it was noted that some spectra were not properly calibrated. The figure on the left shows the same spectra after rechecking the calibration for all spectra. It can be seen that all the spectra overlap as would be expected for this mass region.

If this were not corrected, errors could be made in placing the integration limits for the peaks in the data set, and variance could be introduced into the peak areas that is not due to real sample differences.



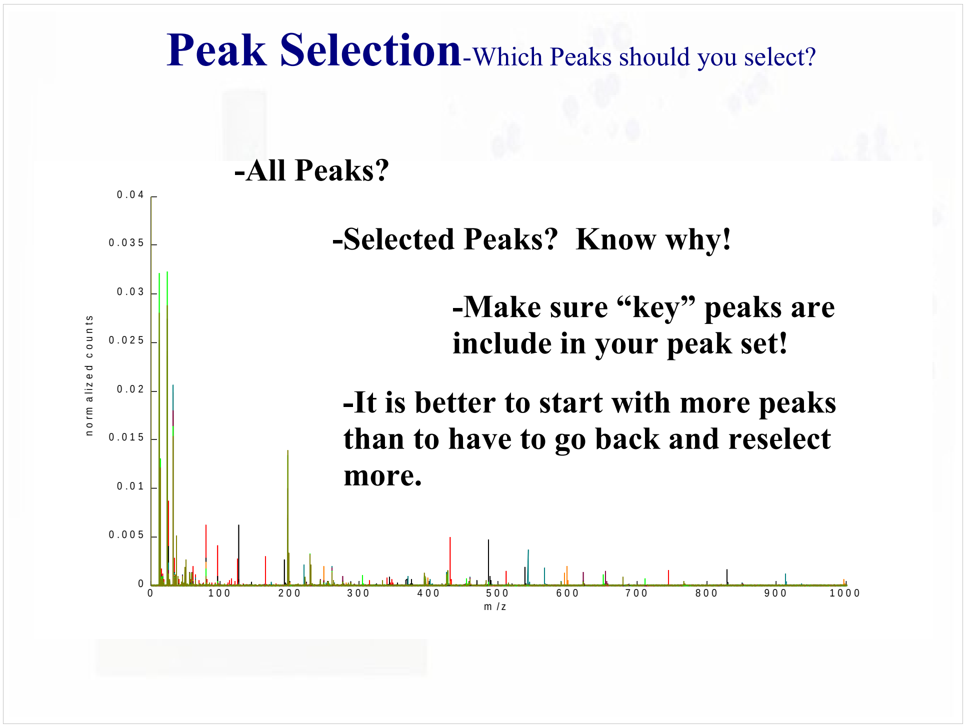
Peak Selection-Which Peaks should you select?

There can be hundreds of peaks in a set of TOF-SIMS spectra.

Each Mass Range... **Contains ...** **Peaks!**

Once the data has been calibrated one has to decide which peaks to include in the data matrix. There are some programs that can read in an entire spectrum for PCA. In this case the entire data set is considered by PCA. Yet, there are cases where including all the peaks in a set of spectra can confound the PCA results and mask sample differences that are overwhelmed by substrate or matrix affects.

There can be hundreds of peaks within any given spectrum. The figures above show an overlay plot of several spectra from different chain length self assembled monolayers. As seen in the figures there are a lot of peaks throughout the entire spectrum Many of these peaks can be seen to be unique to on sample type (different colors).



Peak Selection-Which Peaks should you select?

-All Peaks?

-Selected Peaks? Know why!

-Make sure "key" peaks are include in your peak set!

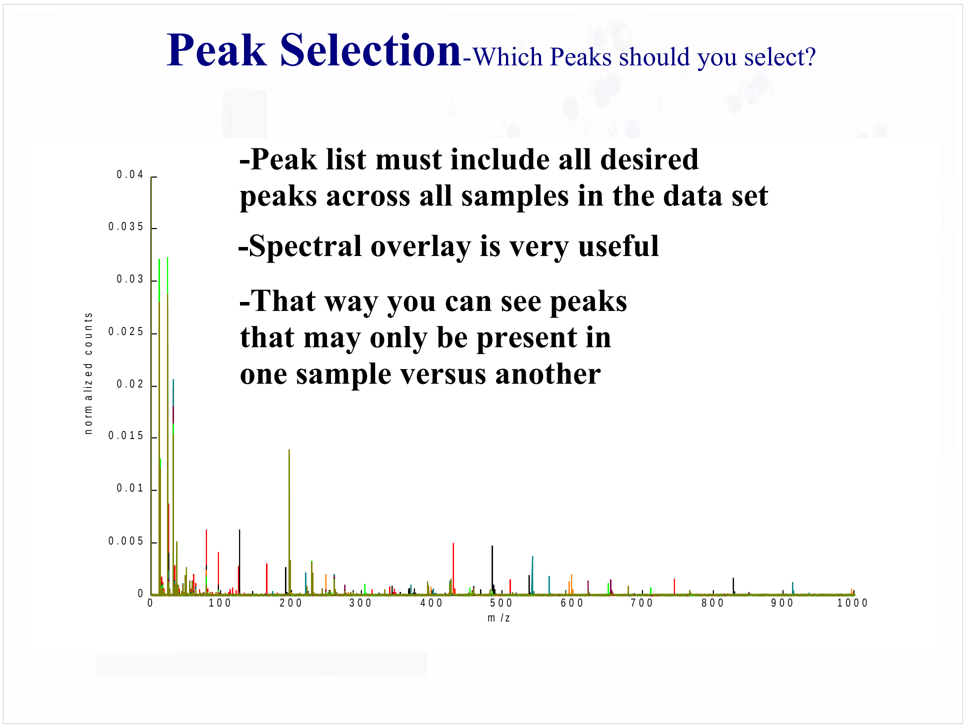
-It is better to start with more peaks than to have to go back and reselect more.

When starting with a given set of data, how many peaks should be included in the data matrix? All? Only some?

When starting with a data set it is often best to start by selecting all the peaks within a given set of criteria. For example all the peaks above a given intensity or background level could be selected. Selecting more peaks from the beginning can save time in the long run since selecting peaks and adjusting integration limits can be time consuming. If later in the analysis it is determined that some peaks are not necessary, they can always be removed from the data matrix. Whereas if the peaks were not selected in the original data set, one would have to go back to the original data to get the peak areas.

If you do select only a few peaks from a given set of spectra, the reason for the peak selection should be understood and stated when reporting the results.

Make sure you include "key" peaks in your peak set. For example if your sample set contains surfaces that produce unique peak signatures, make sure these peaks are included in your selected peaks. This may seem obvious, but can be easily overlooked.



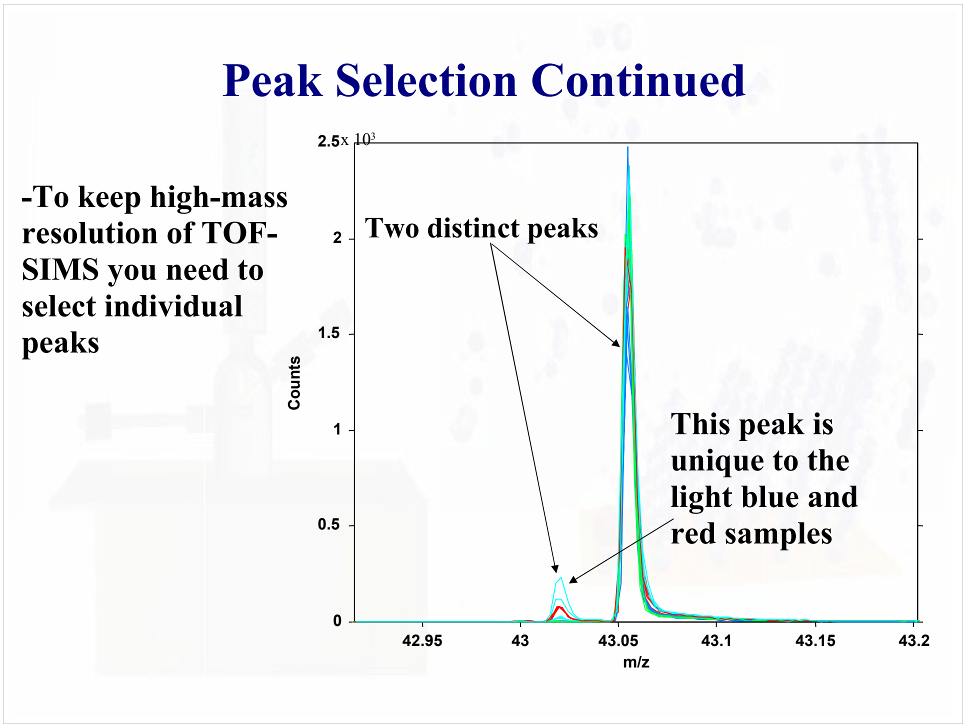
Peak Selection-Which Peaks should you select?

-Peak list must include all desired peaks across all samples in the data set

-Spectral overlay is very useful

-That way you can see peaks that may only be present in one sample versus another

Since the same peak set must be used for all spectra that are to be used in PCA, it is useful to do peak selection from overlaid spectra. This allows the user to see peaks from all spectra on the same axis and helps avoid missing peaks that only show up in the spectra from 1 sample type within the set.



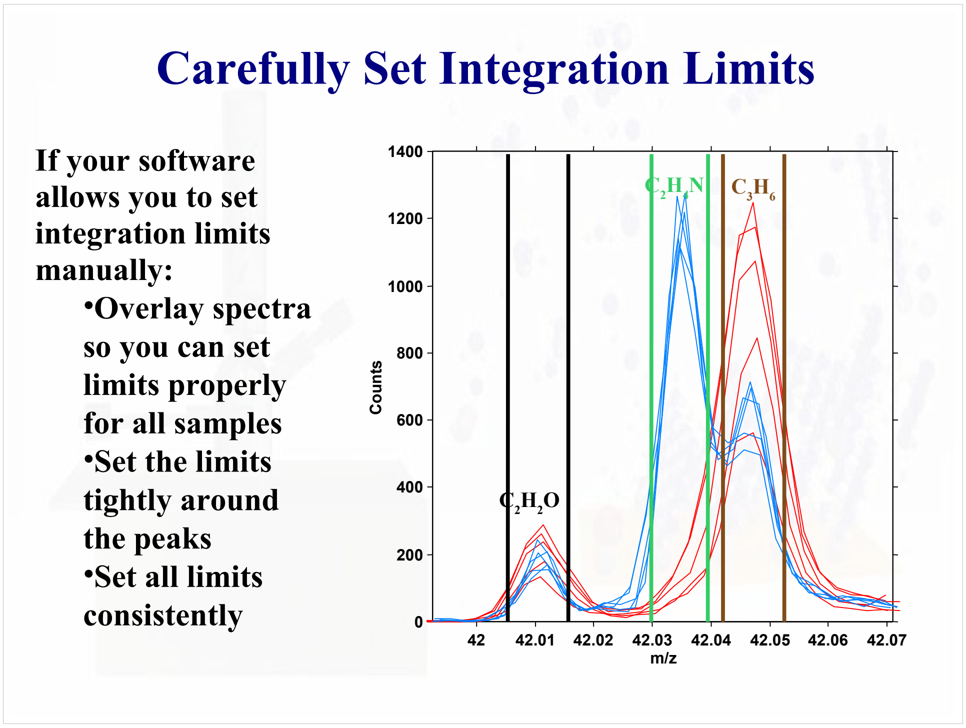
Peak Selection Continued

-To keep high-mass resolution of TOF-SIMS you need to select individual peaks

Two distinct peaks

This peak is unique to the light blue and red samples

Though some programs contain routines to automatically select peaks from a spectrum, it is recommended to do peak selected manually. This will make sure that all the necessary peaks are properly chosen. Also most automatic selection routines are not able to set proper integration limits for the peaks. This can cause problems with PCA since improper peak integration limits mean that the data input into PCA is not an accurate representation of the spectra differences.



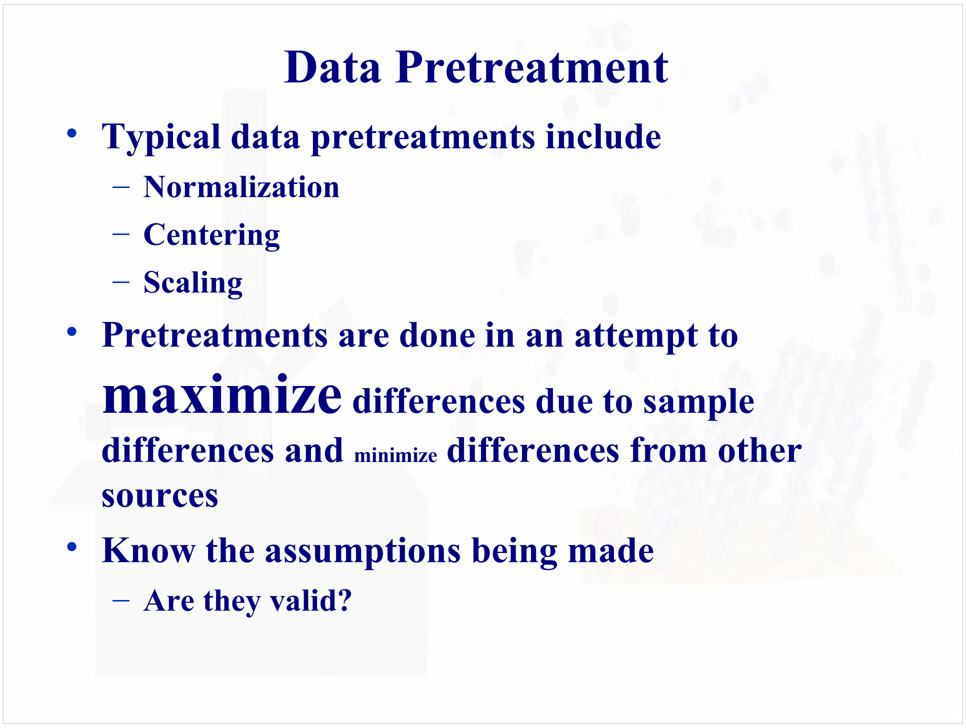
Carefully Set Integration Limits

If your software allows you to set integration limits manually:

- **•Overlay spectra so you can set limits properly for all samples**
- **•Set the limits tightly around the peaks**
- **•Set all limits consistently**

It is important to carefully set all peak integration limits. As seen in the figure above, there are clearly 3 peaks in this mass region. The two peaks on the right side of the figure overlap partially. To minimize integration of the overlapping regions it is necessary to set the peak integration limits in tightly around each peak. Since this is necessary for overlapping peaks, it should be done for all peaks. This will assure consistent, accurate measurement of all peak areas.

Checking peak integration limits can be time consuming, but is necessary for accurate measurement of peak areas.



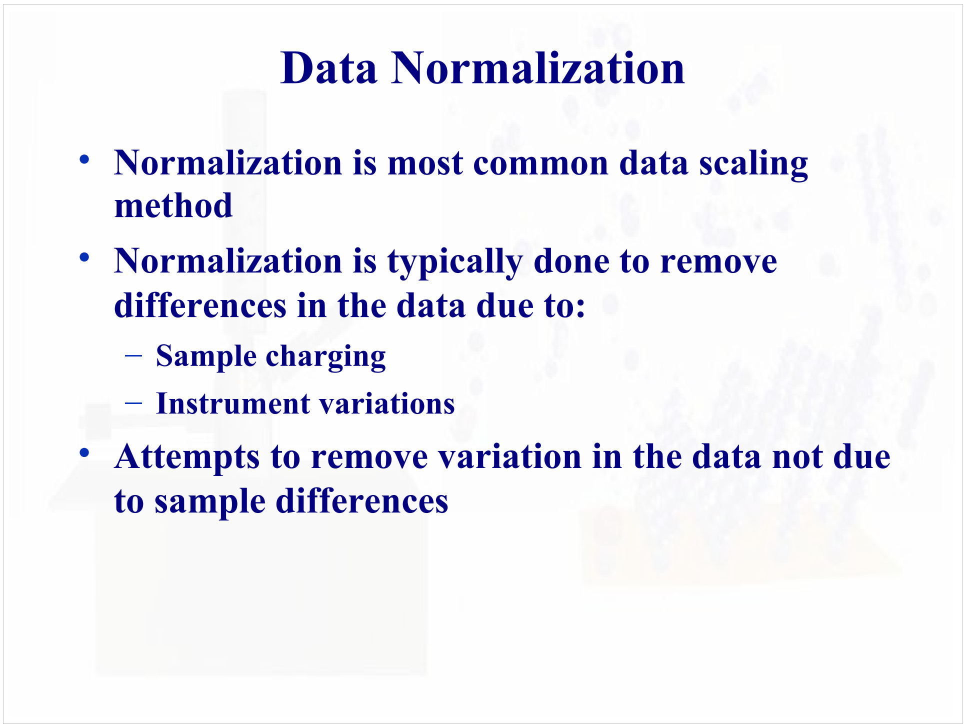
Data Pretreatment

- Typical data pretreatments include
  - Normalization
  - Centering
  - Scaling
- Pretreatments are done in an attempt to **maximize** differences due to sample differences and minimize differences from other sources
- Know the assumptions being made
  - Are they valid?

Before applying MVA methods such as PCA to a data set, it is common to preprocess the data. This is done in order to assure that the differences found in the data set are from true sample differences, and not simply due to differences in the scale or means of the variables included in the data set.

All data preprocessing methods carry with them a set of assumptions. Even by doing no preproccessing you are assuming that the raw data intensities are the best representation of the sample set variation.

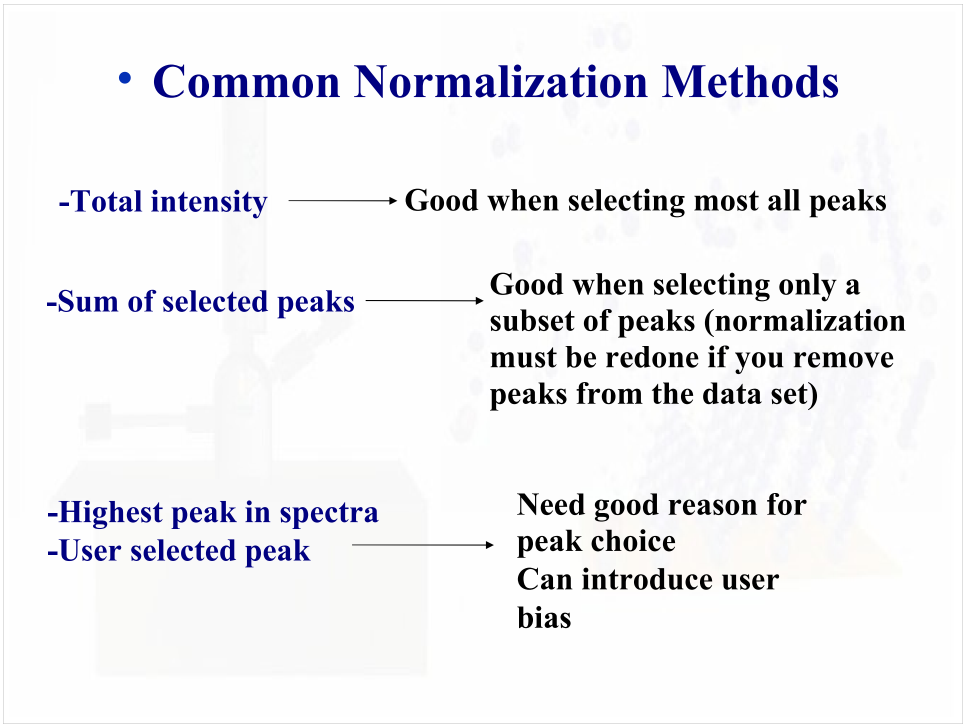
Whichever method of data preprocessing is chosen, it is important to understand the assumptions being made with the method, and to know whether the assumptions made are valid.



Data Normalization

- Normalization is most common data scaling method
- Normalization is typically done to remove differences in the data due to:
  - Sample charging
  - Instrument variations
- Attempts to remove variation in the data not due to sample differences

Data normalization is probably one of the most common preprocessing methods. Normalization is done to account for differences in the data that are due to topography, sample charging, and instrumental conditions. There are many different ways to normalize a set of data. These include normalizing to the total intensity, to the sum of the intensities of the selected peaks, to the highest peak in the spectrum, to a user selected peak, or to a given combination of peaks. Each of these methods brings with it a set of assumptions. For example if you normalize a set of data to the total intensity of each respective spectrum, you are assuming that the total intensity of the spectra does not contain useful chemical information about the samples. This may or may not be true for a given set of data. No matter what normalization method is used, normalization removes information from the data set.



• Common Normalization Methods

-Total intensity → Good when selecting most all peaks

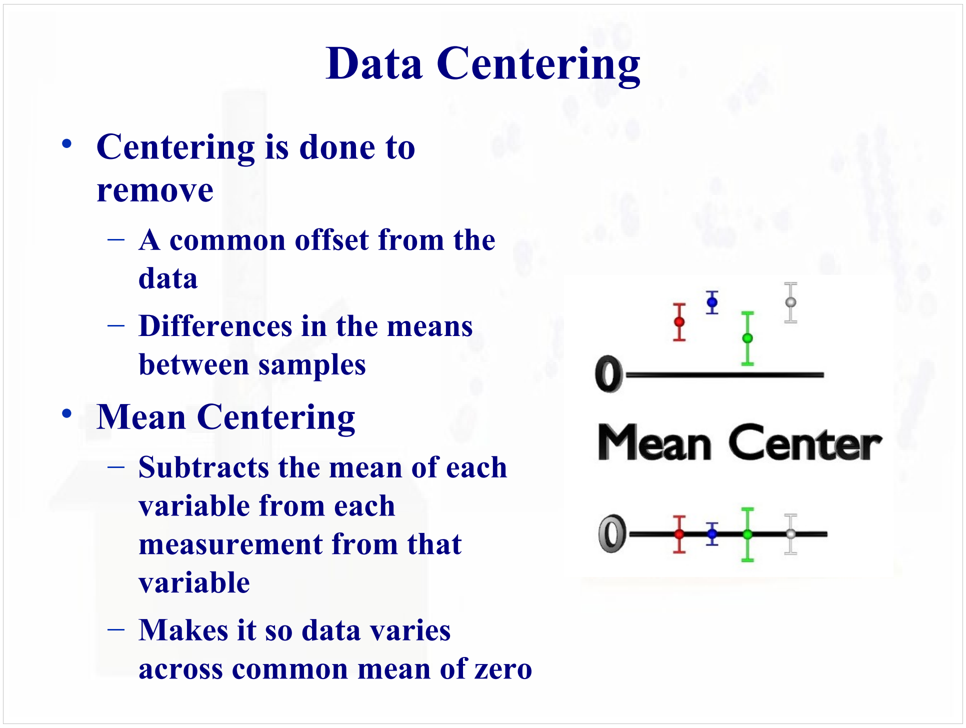
-Sum of selected peaks → Good when selecting only a subset of peaks (normalization must be redone if you remove peaks from the data set)

-Highest peak in spectra
-User selected peak → Need good reason for peak choice
Can introduce user bias

There are many different ways of normalizing a data set. Normalization is typically done by dividing or multiplying the values in the data matrix by a given number or set of numbers. Some typical ways of normalizing TOF-SIMS data include dividing by the total intensity, the sum of selected peaks, the highest peak, or to a user selected peak for the given spectrum.

When using a single peak for spectrum normalization, care should be taken in the selection of the peak. It is possible that by choosing the wrong peak, one may introduce random variation into the data set that would be undesirable.

The choice of data normalization is likely to depend on the data set. If most all peaks have been selected from a given set of spectra, then dividing by the total intensity may be the best choice for normalization. If for some reason, only a few selected peaks have been chosen for the data set, then normalizing by the sum of selected peaks may be the best choice to accentuate the differences between the selected variables.



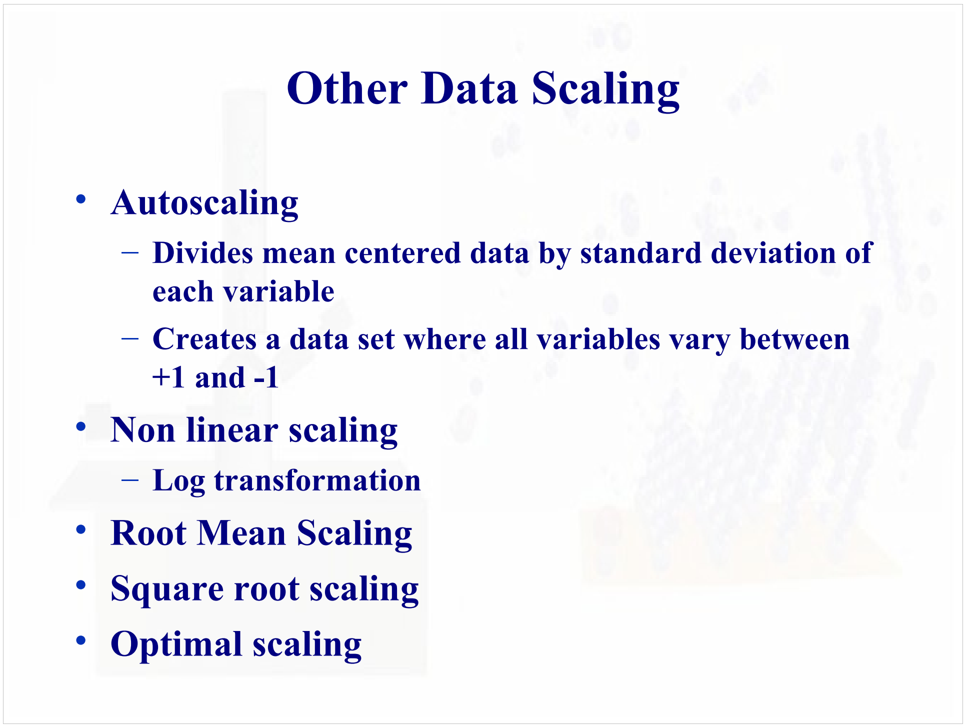
Data Centering

- Centering is done to remove
  - A common offset from the data
  - Differences in the means between samples
- Mean Centering
  - Subtracts the mean of each variable from each measurement from that variable
  - Makes it so data varies across common mean of zero

0
Mean Center
0

The many peaks across a set of TOF-SIMS spectra have a wide range of different intensities. This means that the mean value for any given peak intensity will likely be different for each peak. If one were to use PCA to analyze the data from these types of peaks(variables), PCA would then likely find differences across the data set due to the means of the variables and not the relative variation between the variables.

Mean centering helps avoid this problem by subtracting the mean of each variable from each measurement for the given variable. This results in a data set where all the variables vary across a common mean of zero. This allows looking a the relative intensity differences of the peaks and not just differences in the means.



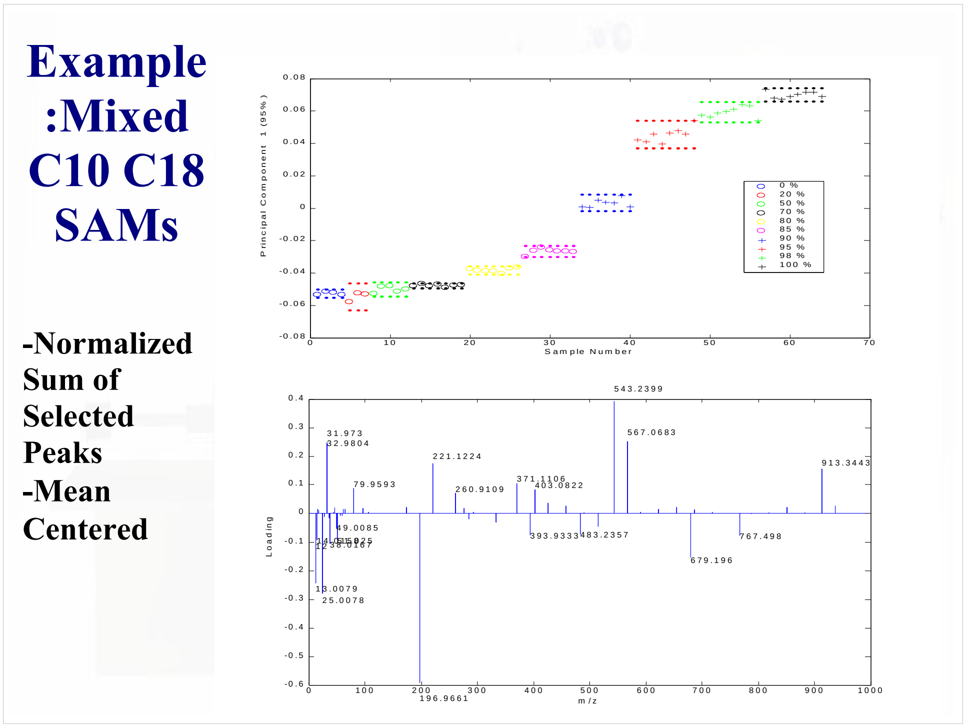
Other Data Scaling

- Autoscaling
  - Divides mean centered data by standard deviation of each variable
  - Creates a data set where all variables vary between +1 and -1
- Non linear scaling
  - Log transformation
- Root Mean Scaling
- Square root scaling
- Optimal scaling

Data scaling is done to account for differences in the variance scales between variables Normalization can be considered a scaling operation since with normalization we are dividing or multiplying by some value to adjust for unwanted variances in the data. One common scaling method used with PCA is autoscaling. Autoscaling is done by dividing a mean centered data set by the standard deviation of each column. This results in a data set where all variables vary between +1 and -1. Autoscaling is commonly used when data from different measurement methods are combined into one data set and one wants to correct for differences in the absolute variance scales of the different methods.

There is still some debate on whether or not SIMS data should be scaled and what method is best to use. Some argue that since the intensity of peaks in a TOF-SIMS spectrum decreases with increasing mass, simply due to the characteristics of the SIMS process and instrumentation, that the data has built in differences in variance scales and should be autoscaled or log scaled. Others argue that regardless of the differences in intensity across a spectrum, all the data comes from the same instrument and therefore does not need autoscaling.

For TOF-SIMS images there is evidence that accounting for the Poisson nature of the noise in the data gives better results from PCA processing.



Example :Mixed C10 C18 SAMs

-Normalized Sum of Selected Peaks
-Mean Centered

The following few slides illustrate the affects that data preprocessing can have on PCA processing of TOF-SIMS data.

The data presented here comes from a set of spectra taken from mixed monolayers of decanethiol (C10) and octadecanethiol (C18). The monolayers were assembled from different solution percentages of the two thiols for >24 hours. This long assembly time assured that the resulting monolayers were most likely in a completely assembled, ordered layer.

Typical peaks for these monolayers include (M= HS(CH2)nCH3):

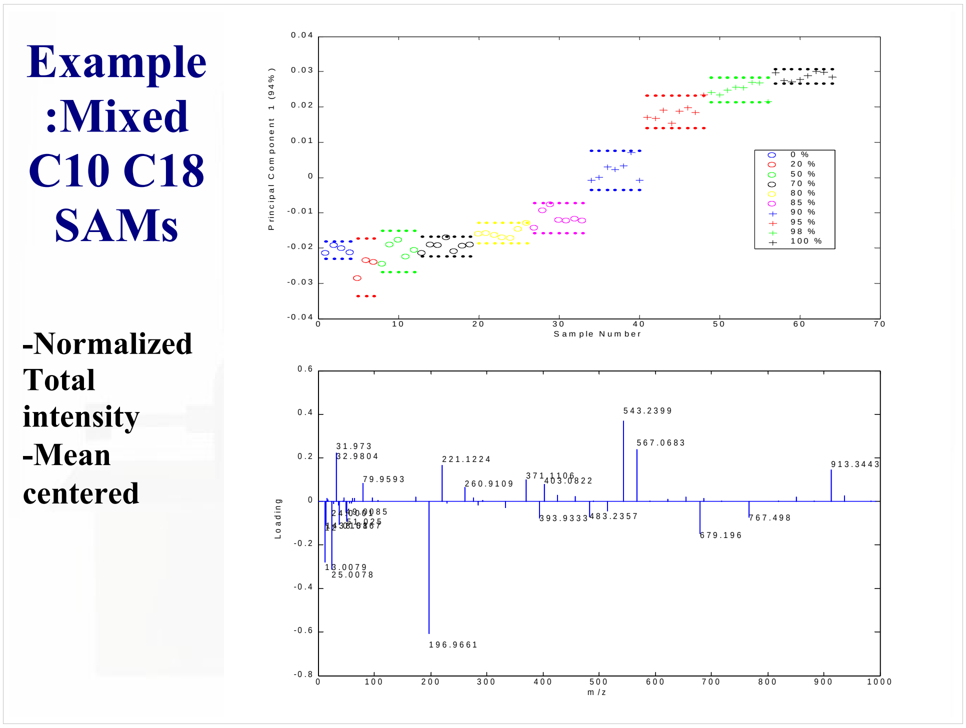
	C10	C18
M-H	173.14	285.26
AuM	371.11	483.24
Au2[M-H]	567.07	679.196
Au[M-H]2	543.24	767.5

The plots shown in these and the subsequent figures show the PC1 scores (upper plot) plotted against the sample number. Since the samples are organized with increasing percentage of C10 thiol, the x-axis can be considered as plotting the increase in the C10 percentage in solution.

The data shown on this slide was normalized to the sum of selected peaks of the respective spectrum and then mean centered.

As seen in the scores plot PCA is able to separate out some sample concentrations, but many of the samples still overlap.

As would be expected the molecular ion clusters for the C10 thiol correspond with higher percentages of C10 thiol (samples with positive scores), and the molecular ion clusters for the C18 thiol correspond with higher percentages of C18 thiol (lower percentages of C10, negative scores). It is noted however that many of the low mass peaks have higher loading values than the high mass molecular ion clusters.

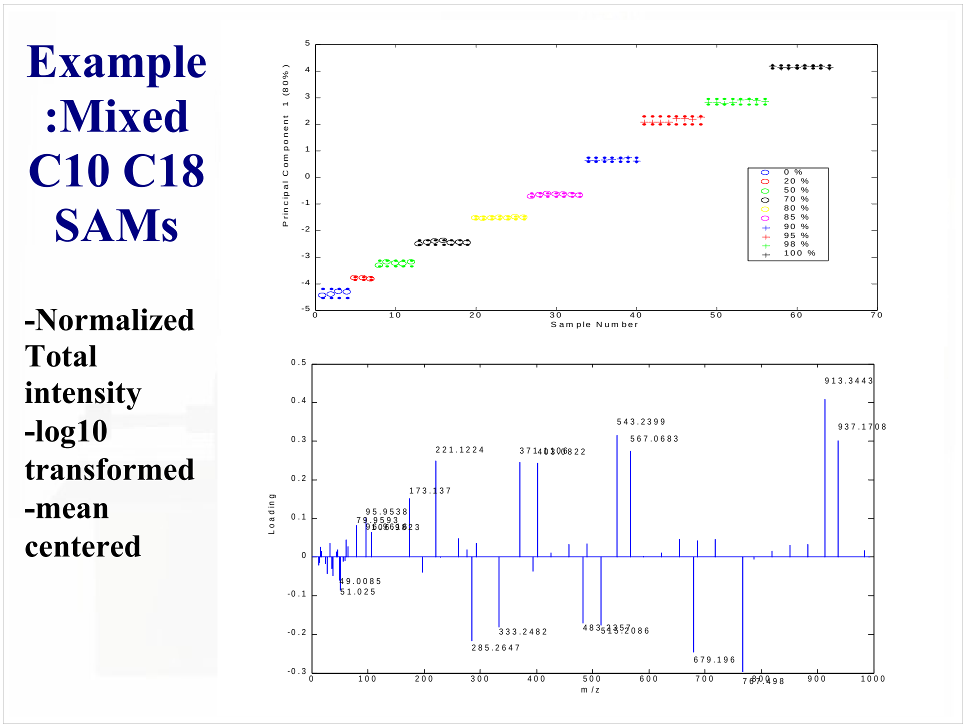


Example :Mixed C10 C18 SAMs

-Normalized Total intensity
-Mean centered

The data shown on this slide shows this same data set normalized to the total intensity of each spectrum and then mean centered. As seen in the scores plot, the sample separation is similar to that seen on the previous slide, but the scatter within each sample group has increased significantly.

The loadings only showed minor changes across the peak set.



Example :Mixed C10 C18 SAMs

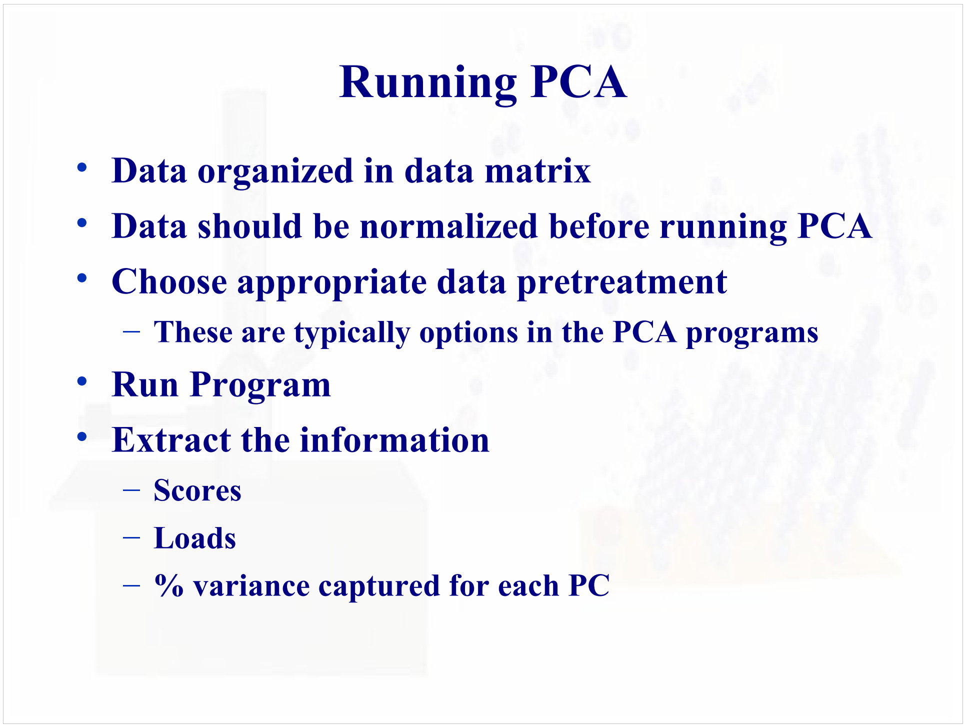
-Normalized Total intensity
-log10 transformed
-mean centered

This slide shows PCA of the C10/C18 data set after normalization to the total intensity, log10 transformation of the data and then mean centering. Log10 transformation was done to equalize the scale of the peaks across the spectrum.

The scores and loadings plots for this data set show distinct changes. As seen in the scores plot, the samples from each solution concentration can clearly be separated along the PC1 axis.

Looking at the loadings plot reveals that the direction of the peak loadings have not changes, but the absolute value of many of the peaks have changed. Most notably, the loadings values of the high mass peaks relative to the low mass peaks have increased significantly.

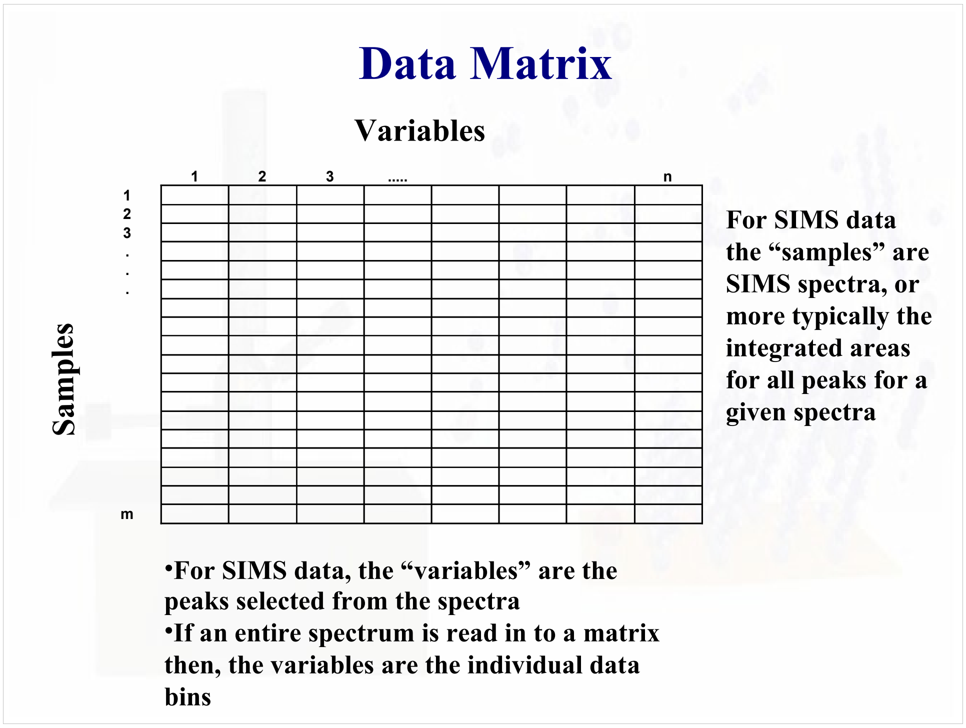
This shows that by reducing the dynamic range of the data, Log10 transformation accentuates the influence of the lower intensity, high mass peaks. In the case of this mixed monolayer data, this allowed better separation of the samples since most of the high mass peaks are highly characteristic of the C10 and C18 thiols.



Running PCA

- Data organized in data matrix
- Data should be normalized before running PCA
- Choose appropriate data pretreatment
  - These are typically options in the PCA programs
- Run Program
- Extract the information
  - Scores
  - Loads
  - % variance captured for each PC

The process of running PCA is outlined on this slide. Firs the data needs to be organized in a matrix. The data is then typically normalized. Data preprocessing such as mean centering and autoscaling are often options provided within packages that perform PCA. After the data set is ready, PCA is run using a software package. Once PCA has been executed, the scores, loadings, and percent variance captured information should be extracted and analyzed.



Data Matrix

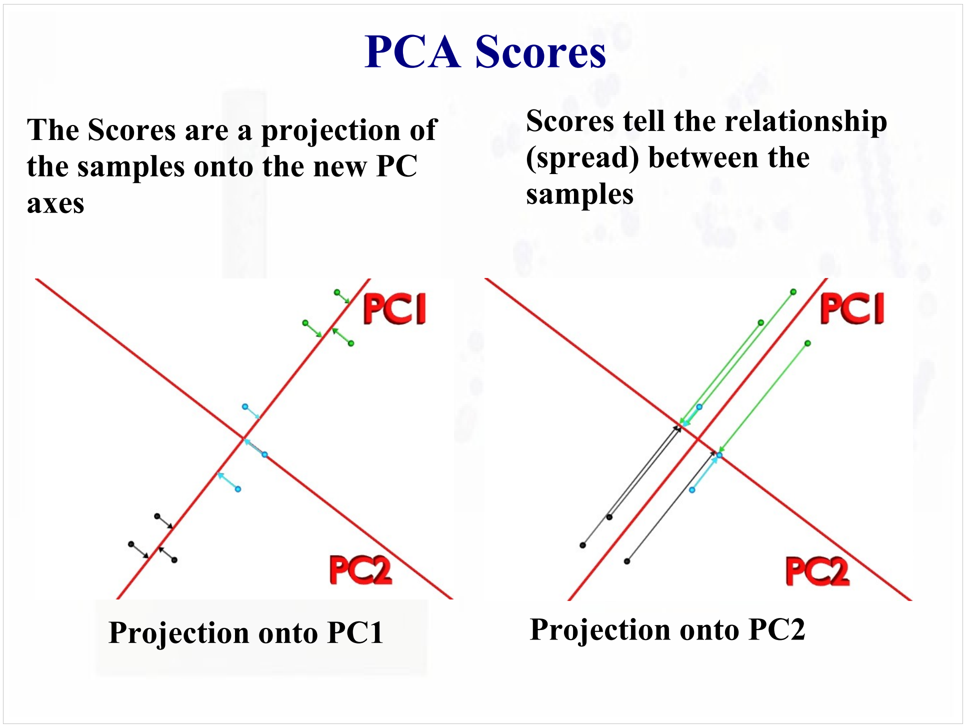
Variables

Samples

For SIMS data the "samples" are SIMS spectra, or more typically the integrated areas for all peaks for a given spectra

- •For SIMS data, the "variables" are the peaks selected from the spectra
- •If an entire spectrum is read in to a matrix then, the variables are the individual data bins

For PCA the data should be organized into a matrix where the samples are organized in rows and the variables are in columns. With TOF-SIMS data, the samples are the spectra from where the peak areas were measured. The variables are the individual peaks (or data bins) that were integrated from the spectra.



PCA Scores

The Scores are a projection of the samples onto the new PC axes

Scores tell the relationship (spread) between the samples

Projection onto PC1

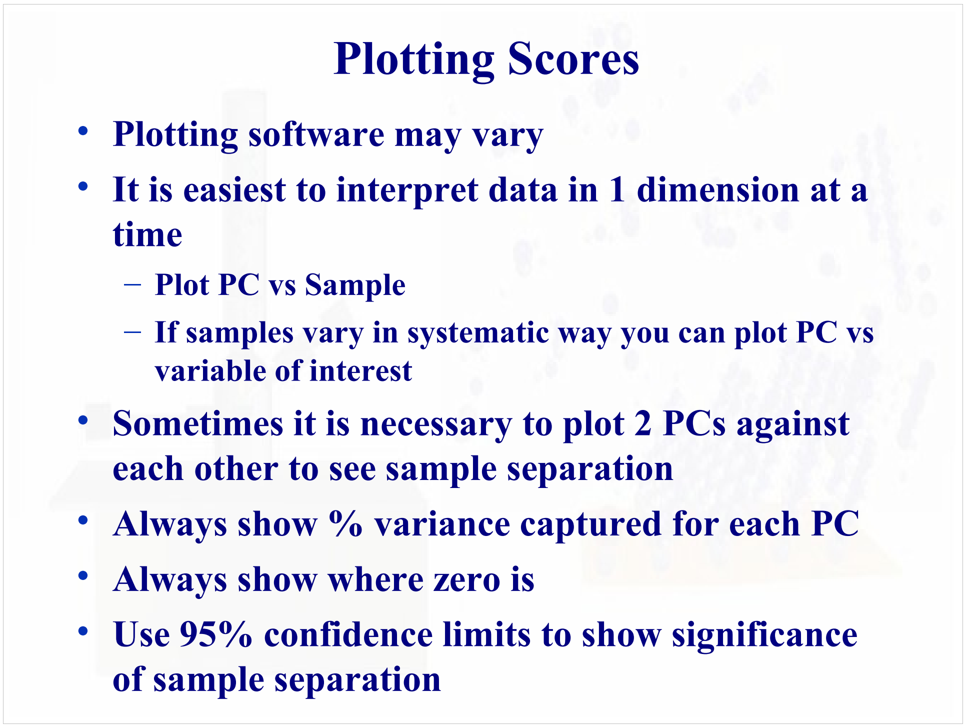
Projection onto PC2

Sometimes it helps to look at PCA graphically to understand better what the scores and loadings represent. As noted in the tutorial "Introduction to PCA", PCA is an axis rotation that creates a new set of axes that best capture the major directions of variation within the data.

The scores are the projection of the samples onto the new PC axes. The scores show the relationship between the samples along these new axes.

As seen in the figure above, the projection of a sample onto a given axis is determined by drawing a perpendicular line from the sample onto the axis. The score value for a sample is the value along the PC where the projection intersects the axis.

In the example shown in the figures on this slide, it can be seen that the green, blue and black sample groups are fairly well separated along PC1, while they completely overlap on PC2.



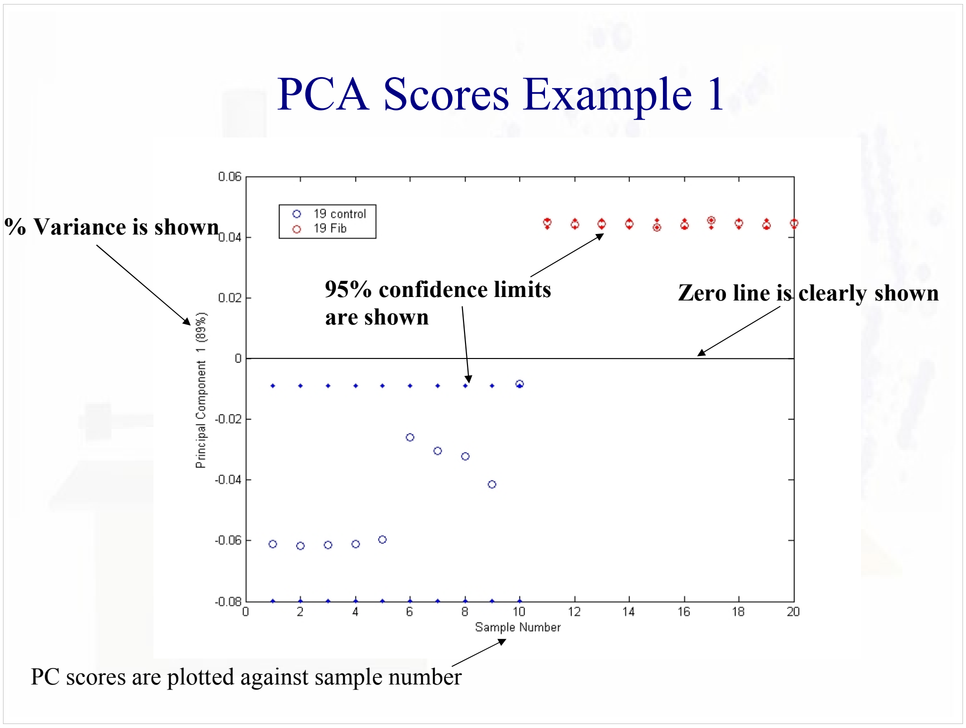
Plotting Scores

- Plotting software may vary
- It is easiest to interpret data in 1 dimension at a time
  - Plot PC vs Sample
  - If samples vary in systematic way you can plot PC vs variable of interest
- Sometimes it is necessary to plot 2 PCs against each other to see sample separation
- Always show % variance captured for each PC
- Always show where zero is
- Use 95% confidence limits to show significance of sample separation

There are many different ways of plotting the scores from PCA. Each software package provides different options for how scores plots can be made. It is often common to plot score values from 2 PCs against each other. With these types of plots you will be looking at how the samples are similar or different along 2 PC axes. For some data sets, this type of plot is necessary to see any separation between samples.

For many data sets however, it is easier to look at 1 PC at a time. This can be useful for data sets where the sample set is organized in order of increasing treatment time or concentration. Plotting a given PC against the sample number, or the variable of interest allows monitoring how the samples change due to this variable.

Whichever format is used for creating the PCA scores plots, it is important that the axes are labeled properly. Always include the PC number and percent variance captured for a given axis. It is also useful to clearly show where the zero line is along the axes. If possible it is useful to show the 95% confidence limits for each of the sample groups in the scores plots (see Wagner, M. S.; Castner, D. G. Langmuir 2001, 17, 4649-4660).



PCA Scores Example 1

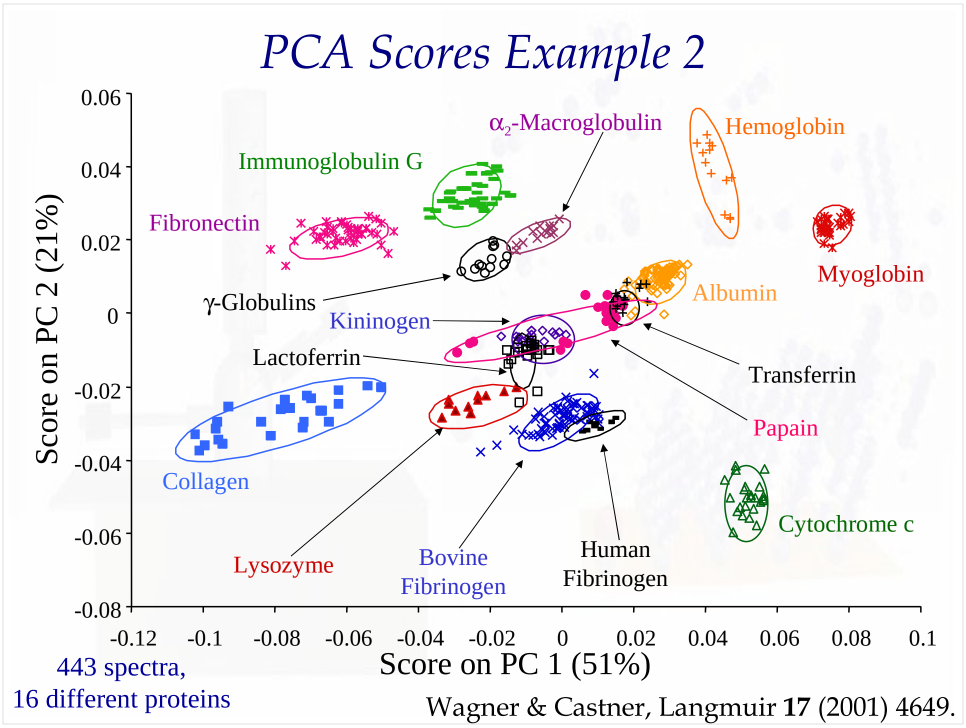
% Variance is shown

95% confidence limits are shown

Zero line is clearly shown

PC scores are plotted against sample number

The figure above shows an example of a scores plot. As seen in the figure, the scores in this example are plotted against the sample number. The PC axis is clearly labeled with the PC number and the percent variance captured. The score values are plotted along with the 95% confidence limits, and the zero line is clearly drawn on the figure. Plotting the scores in this way makes it easy to see that the samples in the plot are clearly separated along the PC1 axis.

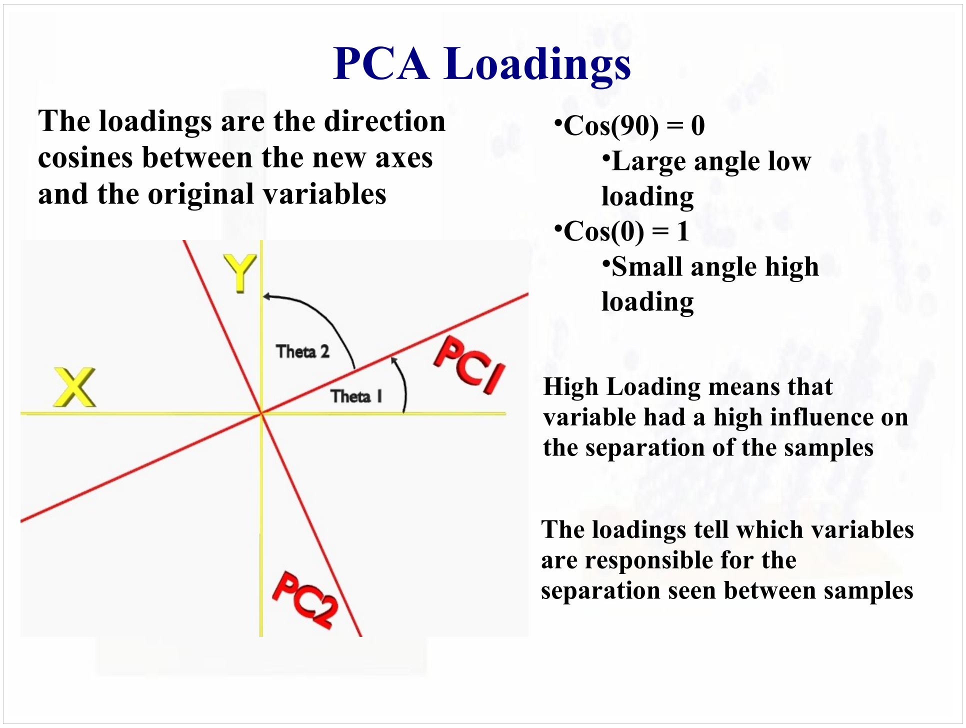


PCA Scores Example 2

443 spectra,
16 different proteins

Wagner & Castner, Langmuir **17** (2001) 4649.

This slide shows another example of a PCA scores plot. In this example, the scores for PC1 and PC2 are plotted against each other. Labels are given for each of the sample groups and the 95% confidence limits are also shown. The zero lines have been left out to avoid too much clutter in the figure. Plotting PC1 versus PC2 was necessary for this data set since it can be seen that along either of the individual PC axes, the samples show significant overlap, while the sample grouping is clearly seen in the cross plot



PCA Loadings

The loadings are the direction cosines between the new axes and the original variables

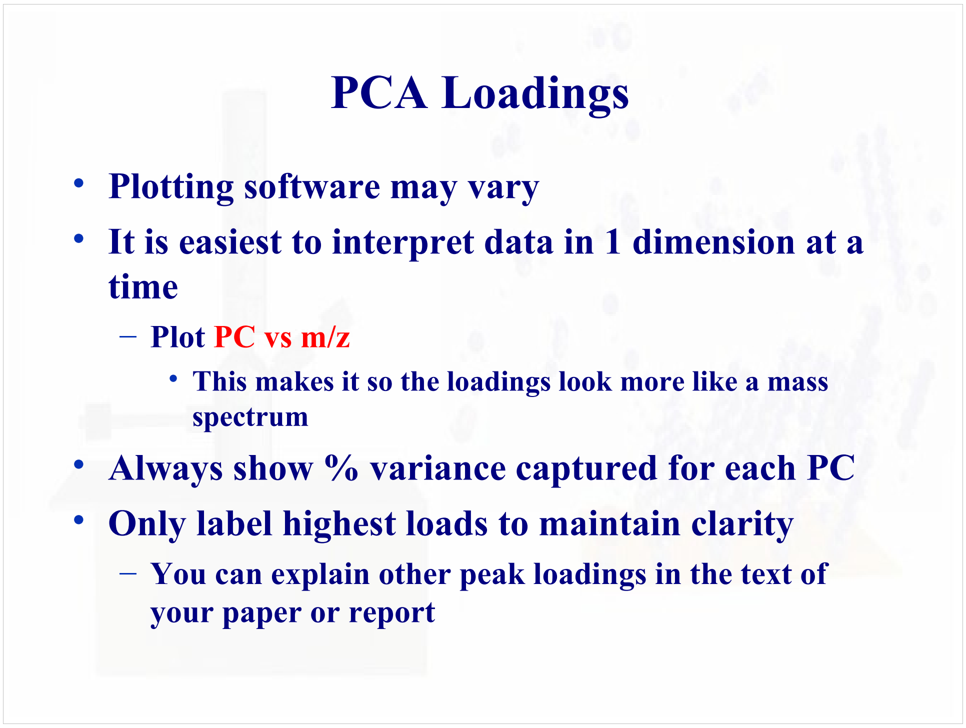
Y
X
Theta 2
Theta 1
PC1
PC2

- •Cos(90) = 0
  - •Large angle low loading
- •Cos(0) = 1
  - •Small angle high loading

High Loading means that variable had a high influence on the separation of the samples

The loadings tell which variables are responsible for the separation seen between samples

The PCA loadings are the direction cosine between the new axes and the original variables. Loadings are the weighting factors used for the original variables to get the new PC axes. A variable with a high loading suggests that the variable had a high influence on the separation of the samples. This means that the original variable and the new PC axes are more highly correlated (smaller angle between the two axes).



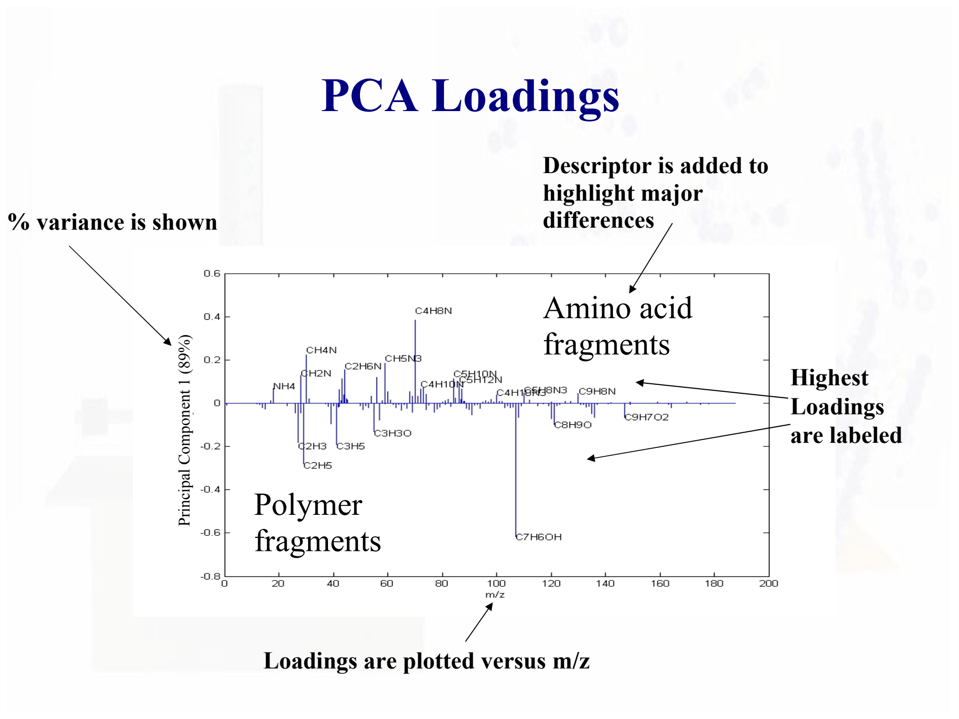
PCA Loadings

- Plotting software may vary
- It is easiest to interpret data in 1 dimension at a time
  - Plot PC vs m/z
    - This makes it so the loadings look more like a mass spectrum
- Always show % variance captured for each PC
- Only label highest loads to maintain clarity
  - You can explain other peak loadings in the text of your paper or report

When plotting PCA loadings for TOF-SIMS data it is useful to plot the loadings versus m/z. This makes the loadings look more like a mass spectrum and allows easier interpretation (since a peak at m/z 196 will appear at m/z 196). Many PCA software packages do not plot the loadings in this way. Most of them plot the loading versus variable number. These types of plots are not visually pleasing and can lead to confusion if not labeled properly.

Loadings plots are most easily interpreted by plotting 1 PC at a time regardless of how you plotted your scores plots.

As with the scores plots it is important to label all loadings plot clearly including the PC number and the percent variance captured. To maintain clarity in the plot it is best to only label the peaks with the highest loadings in the plot. You can explain other trends in the data in the text of your paper or report.



PCA Loadings

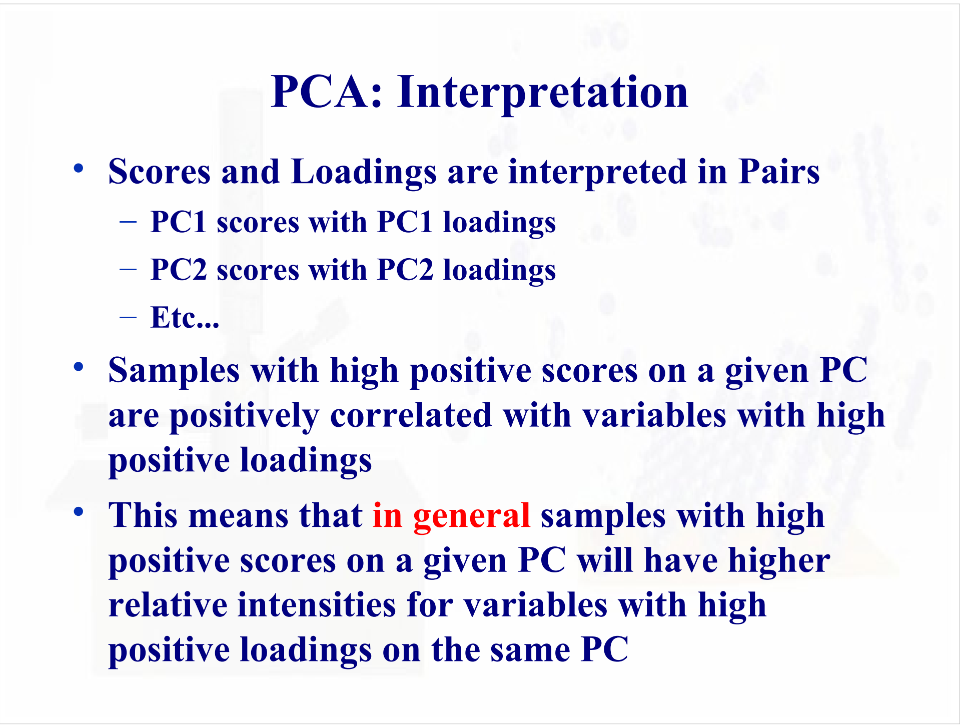
Descriptor is added to highlight major differences

% variance is shown

Highest Loadings are labeled

Loadings are plotted versus m/z

This slide shows an example PCA loadings plot. As seen in the figure, the loadings are plotted against m/z. The PC axis is clearly labeled. Peak labels are provided for the peaks with higher loadings values. It can also be seen that a general descriptor has been placed on either side of the PC1 axis to summarize the types of peaks seen on each side.



PCA: Interpretation

- Scores and Loadings are interpreted in Pairs
  - PC1 scores with PC1 loadings
  - PC2 scores with PC2 loadings
  - Etc...
- Samples with high positive scores on a given PC are positively correlated with variables with high positive loadings
- This means that in general samples with high positive scores on a given PC will have higher relative intensities for variables with high positive loadings on the same PC

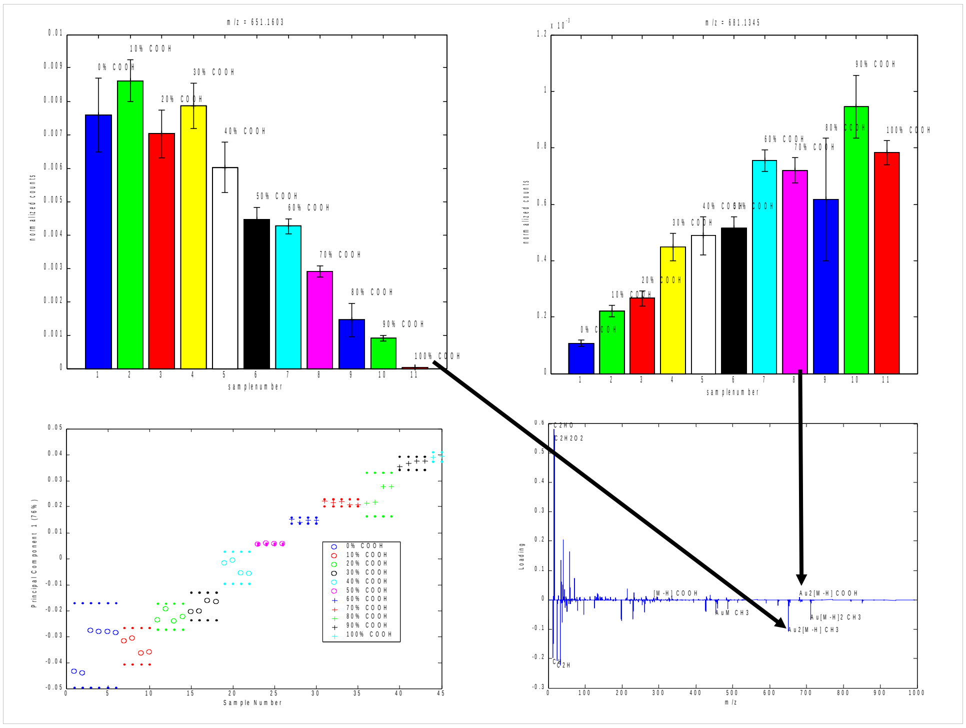
Interpretation of PCA results is done using the scores and loadings plots together. You cannot interpret PCA results by looking at either one alone. This is because the two plots contain complimentary information and either one without the other is incomplete. For example you can see clear separation between samples on a scores plot, but without looking at the loadings you will not know which peaks are responsible for the separation seen. When looking at the scores and loadings it is important to make sure you have the two plots properly matched up (PC1 scores with PC1 loadings, etc).

The rules for interpreting PCA scores and loadings plots can be summarized as follows:

Samples with positive scores on a given PC axis are positively correlated with variables with positive loadings on the same PC axis. Samples with negative scores are positively correlated with variables with negative loadings. This means that, in general, samples with positive scores will have higher relative intensities for peaks with positive loadings than samples with negative scores. The opposite is also true, samples with negative scores will, in general, have higher relative intensities for peaks with negative loadings.

It is also true that samples with positive scores are negatively correlated with variables with negative loadings and that samples with negative scores are negatively correlated with variables with positive loadings.

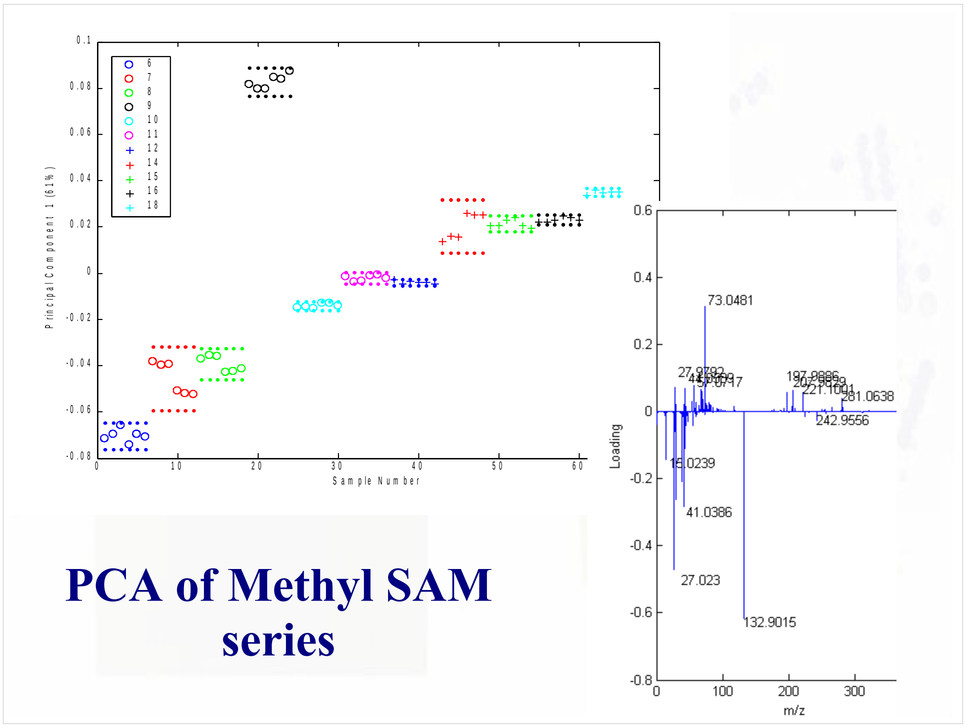
It is important to note that since PCA looks at differences in the relative intensity of variables, even if a variable is negatively correlated with a given set of samples, it does not mean that the value of that variable for those samples is necessarily zero. It just means that those samples have a lower relative intensity than samples that are positively correlated with the variable.



This slide shows an example of scores and loadings plots in the lower two figures. The upper figures are shown to illustrate how the trends shown in the scores and loadings are reflected in the original normalized data.

The bar charts on top of this slide show the original, normalized, data for one of the COOH thiol peaks (positive loading right chart), and one of the CH3 thiol peaks (negative loading left chart). As seen in the charts the original data follows the trend that would be expected based on the appearance of the scores plot. The COOH peak is seen to increase with increasing percentage of COOH and the CH3 peak is seen to decrease with increasing COOH percentage.

In this case the trends in the loading are pretty clear, but this is not always the case as will be illustrated in the next example.

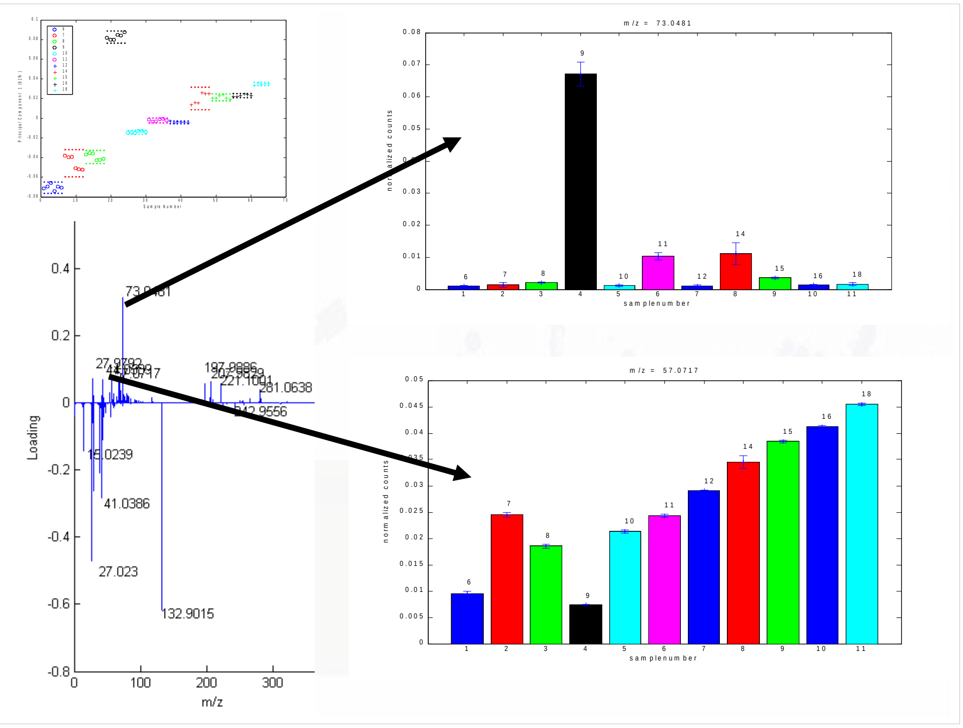


PCA of Methyl SAM series

This slide shows a more complex example where more than one trend is reflected in the scores and loadings plots. This example is shown to illustrate the importance of looking back at the original data to verify the trends seen in the PCA plots.

The slide above shows the PC1 scores and loadings plots from a set of methyl terminated self-assembled monolayers with varying chain lengths (from C6 to C18). It can be seen that the PC1 score values increase with increasing chain length with one clear outlier. The C9 thiol samples are seen to have significantly higher scores than the other samples and clearly do not follow the general trend. Looking at the loadings plot it is noted that the positive loadings are dominated by the peak at m/z = 73 (indicative of PDMS). There are also some low mass hydrocarbons that have positive loadings. Based solely on the trends seen in the scores plot, and what we know about interpreting scores and loadings, it would be logical to assume that if we looked at the original data for the peak at m/z = 73 we would see that the C9 samples would have the highest relative intensity followed by the C18, C16/C15, C14 and so forth. We might also expect this to be true if we plotted on of the hydrocarbon peaks (C9 would have the highest relative intensity followed by C18, C15/C16, etc).

On the next slide we will see that this is not the case.

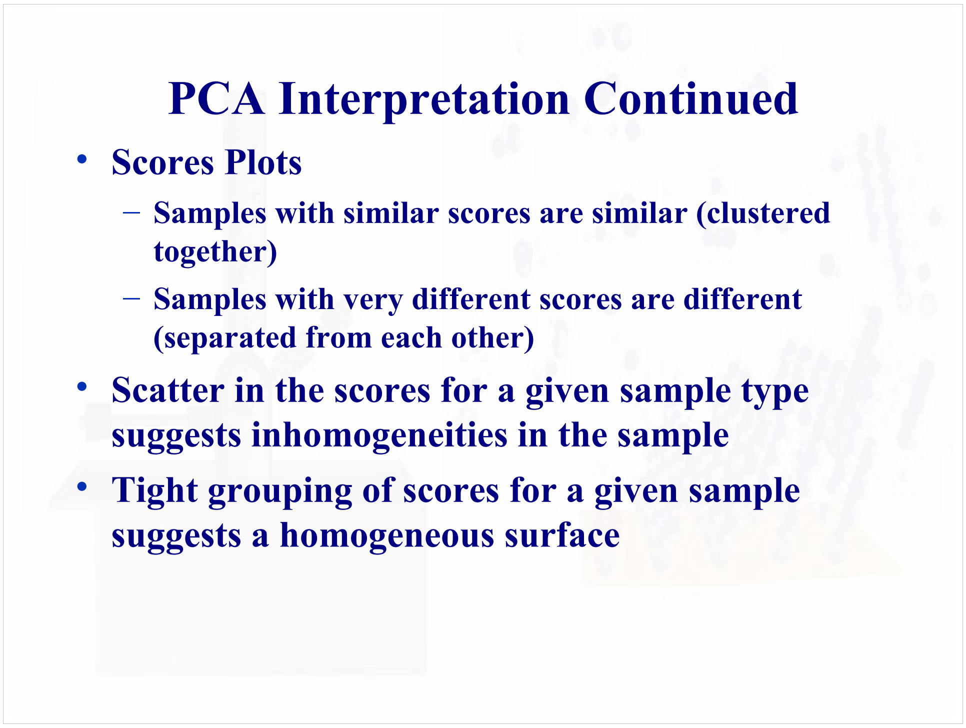


Here we seen the original, normalized data for the peak a m/z = 73 (top bar chart) and m/z = 57 (bottom bar chart). As seen in these bar charts, the peaks do not follow the assumed trends. Why is this? It appears that PC1 is tracking two major trends in the data. The first is the PDMS contamination on the C9 sample. The relative intensity of the peak at m/z = 73 is clearly orders of magnitude higher than the other samples. This is also true of other PDMS related peaks. PCA looks for variance in the data set and this is clearly a large source of variance. At the same time there is a large source of variance from the changes induced by the increasing chain length of the thiols. This is clearly seen in the bar chart for the hydrocarbon peak at m/z = 57. So PC1 is capturing a combination of the two sources of variation.

Hopefully this example has shown why it is important to actually check the original data and not just assume that the relative intensities of peaks highlighted in the loadings plots will follow the trends you expect to see. Most of the time they will, but you need to check!

One important thing to note is that checking the original data against the PCA results is not a simple straight forward task for PCs greater than PC1. For PC1 one can simply plot the data values from the matrix that was entered into the PCA. For PCs greater than PC1, you will need to subtract previous PCs from the data matrix before plotting the “original data” since this is what PCA does when calculating each PC.

For example the data analyzed to find PC2 is the original data matrix minus PC1. The data analyzed to find PC3 is the original data matrix minus PC1 and PC2 and so forth.



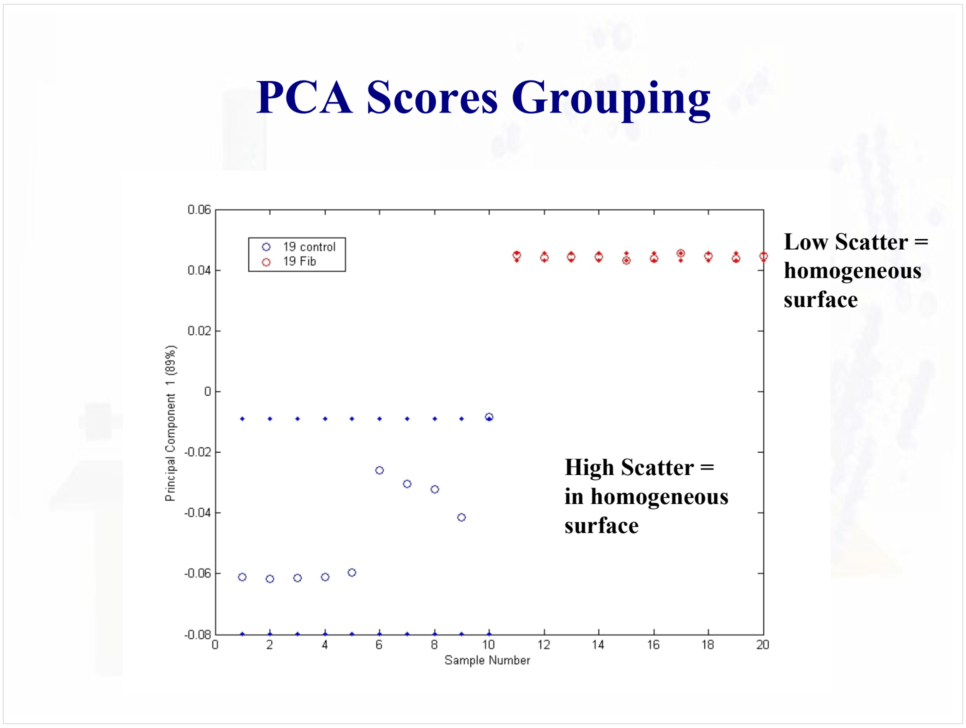
PCA Interpretation Continued

- Scores Plots
  - Samples with similar scores are similar (clustered together)
  - Samples with very different scores are different (separated from each other)
- Scatter in the scores for a given sample type suggests inhomogeneities in the sample
- Tight grouping of scores for a given sample suggests a homogeneous surface

PCA scores plots can provide several pieces of information about a sample set. First of all the scores can show the relationship between samples (are they similar or different). Samples with similar score values implies that the samples are similar based on the variables input into PCA. For TOF-SIMS data this means that samples that are clustered together in a scores plot are spectrally similar. Conversely, samples with very different scores values are spectrally different from each other.

Scores can also show the reproducibility of the spectra within a given sample group. For example, scatter in the scores values for a given set of spectra suggests there are inhomogeneities within th samples Tight clustering of spectra from a given sample group suggests the sample chemistry was homogeneous

Therefore the scores can be used to look for sample difference and to determine reproducibility within sample groups.



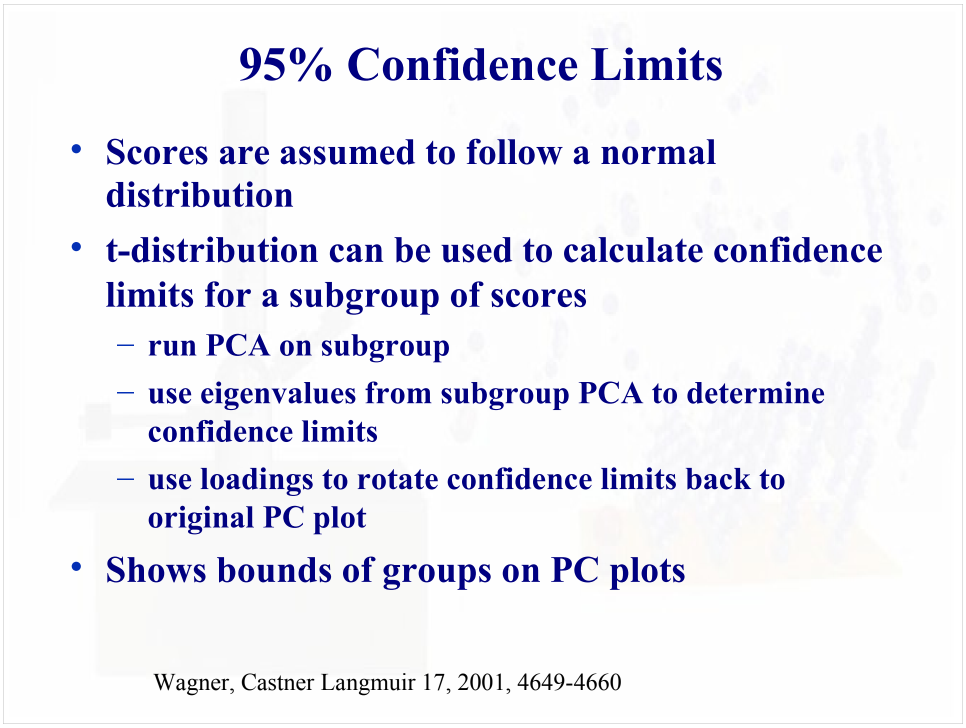
PCA Scores Grouping

Low Scatter =
homogeneous
surface

High Scatter =
in homogeneous
surface

The scores plot above shows the scores from a data from a polymer with and without adsorbed fibrinogen. As seen in the figure, the red dots representing the scores for the samples with adsorbed fibrinogen are all lined up with very low scatter, whereas the blue dots from the bare polymer show higher scatter suggesting inhomogeneities on the surface.

It should be noted that if there is a lot of scatter in the data, you will need to collect more data to be confident of any sample separation seen from PCA.

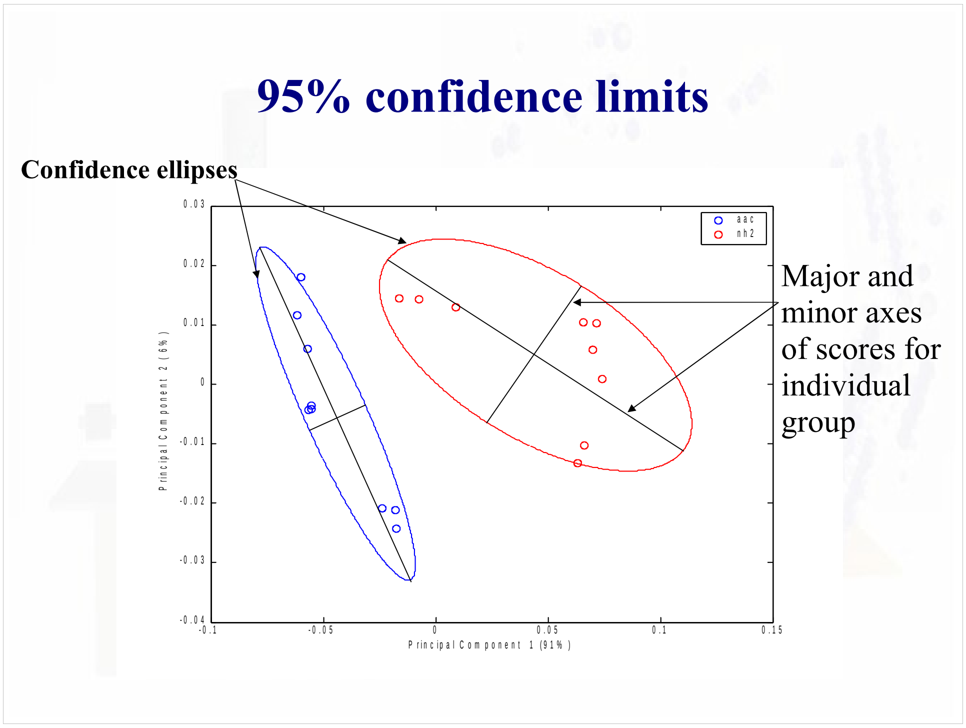


95% Confidence Limits

- Scores are assumed to follow a normal distribution
- t-distribution can be used to calculate confidence limits for a subgroup of scores
  - run PCA on subgroup
  - use eigenvalues from subgroup PCA to determine confidence limits
  - use loadings to rotate confidence limits back to original PC plot
- Shows bounds of groups on PC plots

Wagner, Castner Langmuir 17, 2001, 4649-4660

As shown in Wagner, M. S.; Castner, D. G. Langmuir 2001, 17, 4649-4660, 95% confidence limits can be calculated for each sample group. This gives a way of monitoring the significance of separation between samples. See the reference given for more information.

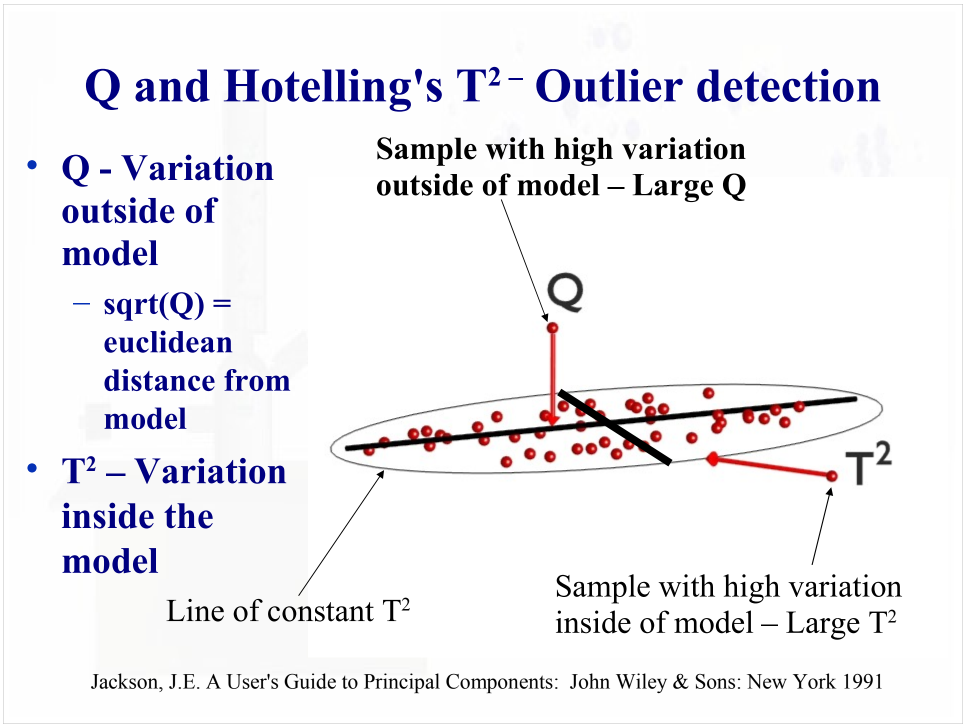


95% confidence limits

Confidence ellipses

Major and minor axes of scores for individual group

This plot shows an example of a PCA scores plot showing 95% confidence ellipses. The major and minor axes of the ellipses are shown to illustrate how the confidence ellipses capture the major variations within each subgroup within the plane of the scores plot. As outlined on the previous slide, this is done by running PCA on the subgroup alone and then projecting the data back into the original scores plot.



Q and Hotelling's T^2 – Outlier detection

- **Q - Variation outside of model**
  - **sqrt(Q) = euclidean distance from model**
- **T^2 – Variation inside the model**

Sample with high variation outside of model – Large Q

Line of constant T^2

Sample with high variation inside of model – Large T^2

Jackson, J.E. A User's Guide to Principal Components: John Wiley & Sons: New York 1991

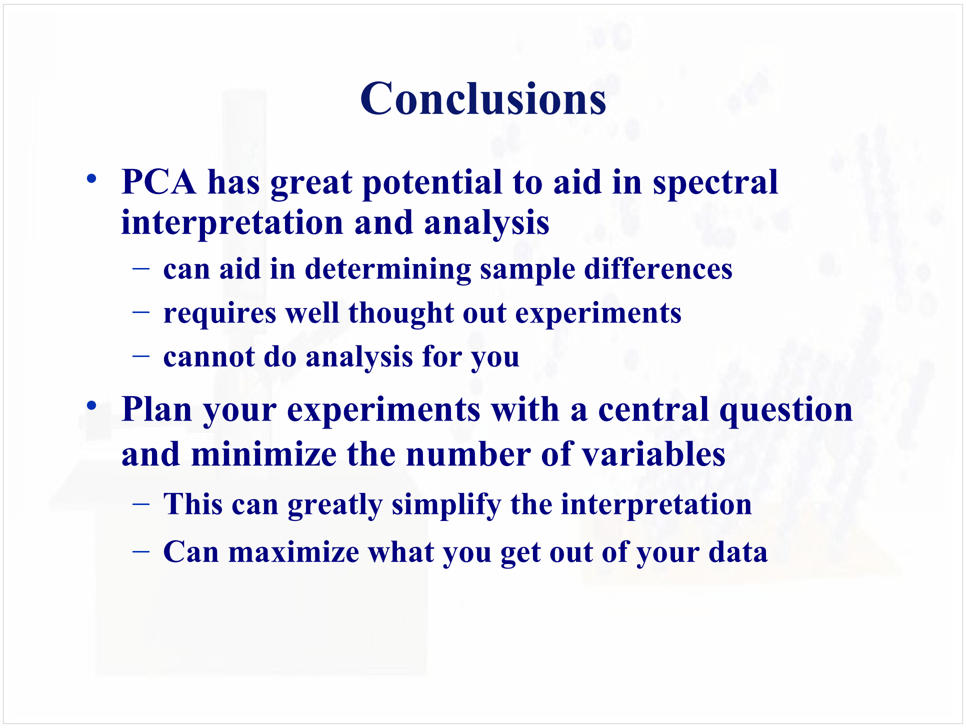
Q and the Hotellings T2 are two test statistics that can be calculated for PCA models to look for outliers.

Q highlights samples with large variation outside of a PCA model.

T2 highlights samples with large variation within a PCA model.

This is illustrated graphically in the figure above.

See the references given for more information about these statistics.



Conclusions

- **PCA has great potential to aid in spectral interpretation and analysis**
  - **can aid in determining sample differences**
  - **requires well thought out experiments**
  - **cannot do analysis for you**
- **Plan your experiments with a central question and minimize the number of variables**
  - **This can greatly simplify the interpretation**
  - **Can maximize what you get out of your data**

PCA is a powerful data analysis tool. Utilizing this power properly requires good experimental design, and understanding of how PCA works.

As with most experimentation, time and thought put into the planning of the experimental design will greatly increase the possibility of learning new and useful insights from the experiment results.