The Complexities of The Basics of PCA (Part I)

Dan Graham, Ph.D.
NESAC/BIO
University of Washington

An NIH National Center for Research Resources
(NIH grant EB-002027)
http://www.nb.uw.edu

NIH National Institute of Biomedical Imaging and Bioengineering

NESAC/BIO

Overview

- Why Use Multivariate Methods?
- What are Multivariate Statistics?
- How to use MVA with ToF-SIMS data
    - Basic complexities of running PCA - step by step
    - Basic complexities of displaying and interpreting results

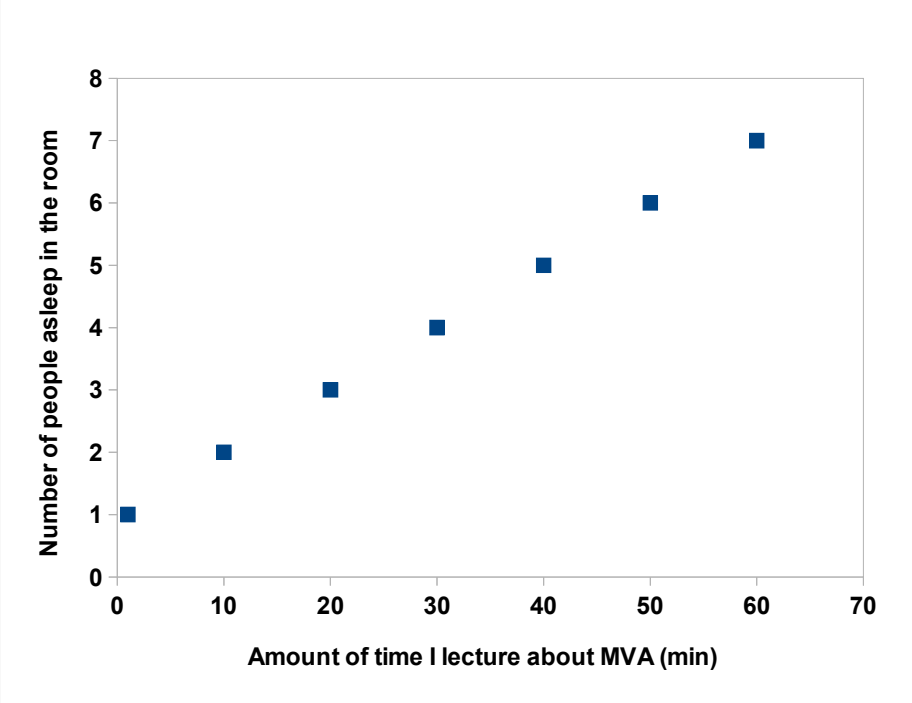
WHY?

Why Use Multivariate Statistics?

- Reduce Size of Huge Data Sets
  - Keep important information
- Remove User Bias
- Efficiently Use the Data
- Chemical Signatures are Multivariate
- Biological Problems are Multivariate
- Quantitative Analysis
- NOTE: MVA is a tool, not a replacement for knowing what you are doing

MVA Sounds Cool, Should I Always Use It?

- No. If you only have a few variables, it probably doesn't make sense to use MVA.
- If another method can answer your question, then use the other method.

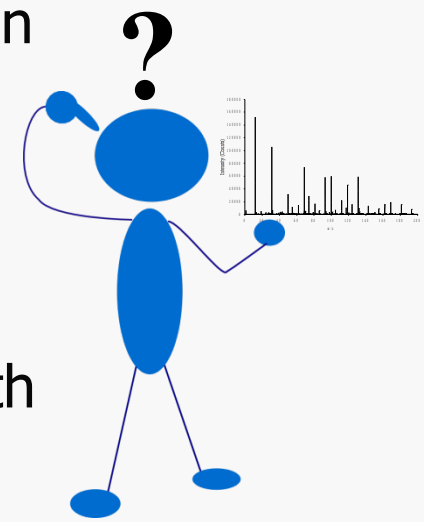


MVA Can Be Very Useful If:

- You have data sets with a large number of samples and variables
- You want to remove potential user bias
- You want to identify samples or predict a response based of experimental measurements

MVA Is Useful for ToF-SIMS Because SIMS Data Is Complicated

- Spectra contain hundreds of peaks
- Images contain thousands of spectra
- Peak intensities can be interrelated
- Matrix effects can cause non-linear changes in peak intensities
  - Due to sample composition
  - Due to the presence of oxides
  - Due to the presence of salts
- Peak intensities may or may not correlate with surface composition
- Peak intensities may vary due to differential sputtering
- Often heavy fragmentation and lack of molecular ion signals (assuming you know what signal to expect)



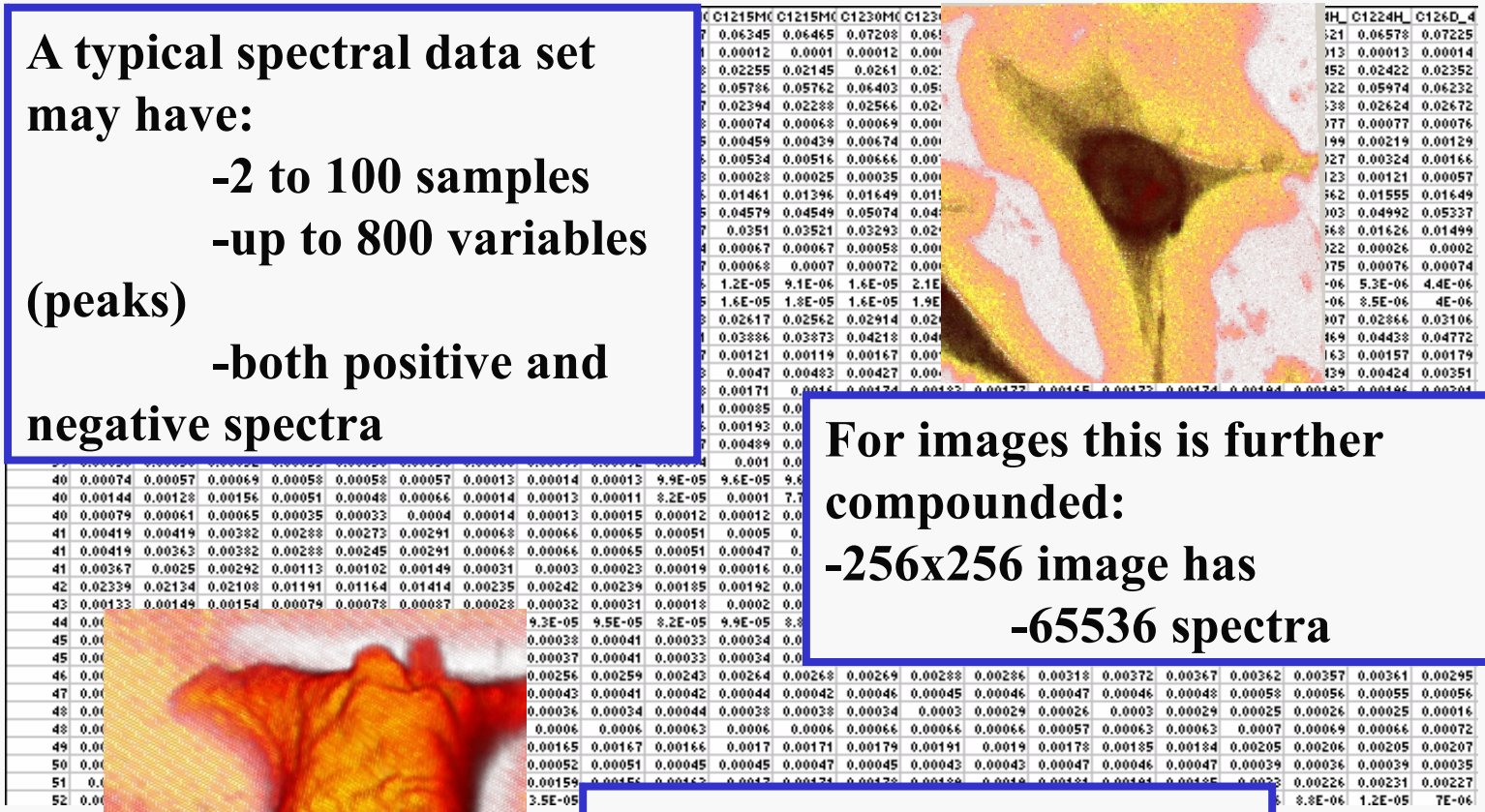
Data Overload

A typical spectral data set may have:

-2 to 100 samples

-up to 800 variables (peaks)

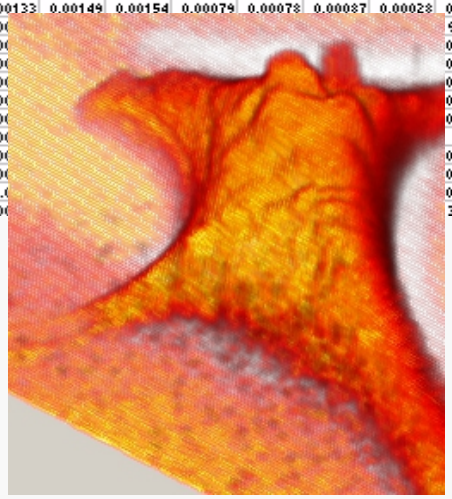
-both positive and negative spectra



For images this is further compounded:

-256x256 image has

-65536 spectra



This is multiplied by the number of slices in a 3D data stack – can quickly reach millions of spectra

WHAT?

“Multi” “Variate”

- Multivariate = more than 1 variable
- So multivariate analysis pertains to the analysis of multiple variables
  - Response from multiple measurements/instruments
    - XPS, ToF-SIMS, Contact Angle
  - Multiple responses from single measurement/instrument
    - Multiple ToF-SIMS peak intensities
- MVA – looks at the dependence (covariance) between different variables

Multivariate Analysis Methods

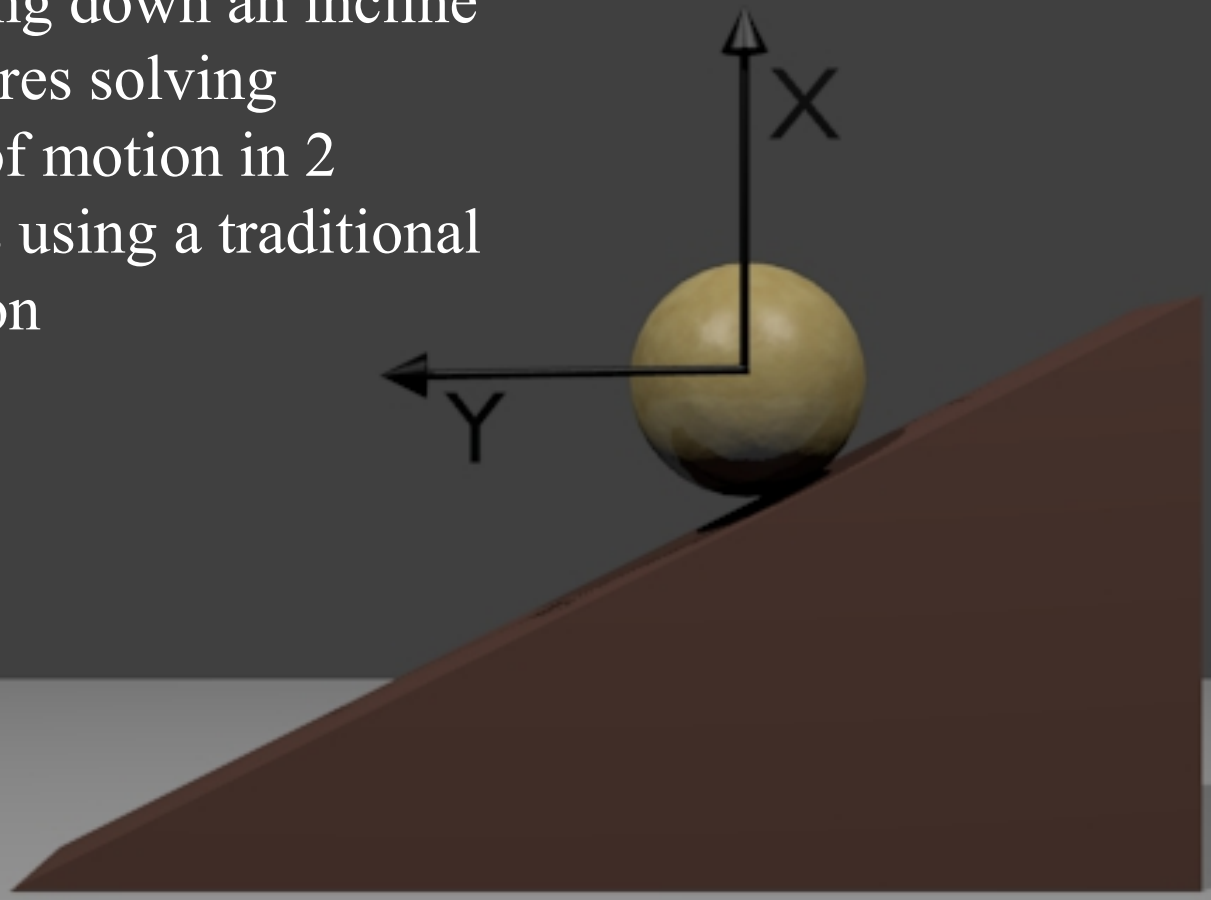
- Many different methods available
  - Principal component analysis (PCA)
  - Factor analysis (FA)
  - Discriminant analysis (DA)
  - Multivariate curve resolution (MCR)
  - Maximum Autocorrelation Factors (MAF)
  - Partial Least Squares (PLS)
  - PCA-DA, PLS-DA, CA
- We will focus on PCA
  - Most commonly used method
  - Successful with SIMS data
  - Forms the basis for many other methods

When Should MVA be Used?

- MVA 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

Multivariate Analysis
Reducing the Dimensionality of a Problem

A ball rolling down an incline plane requires solving equations of motion in 2 dimensions using a traditional axes rotation



Slide adapted from slide Bonnie Tyler

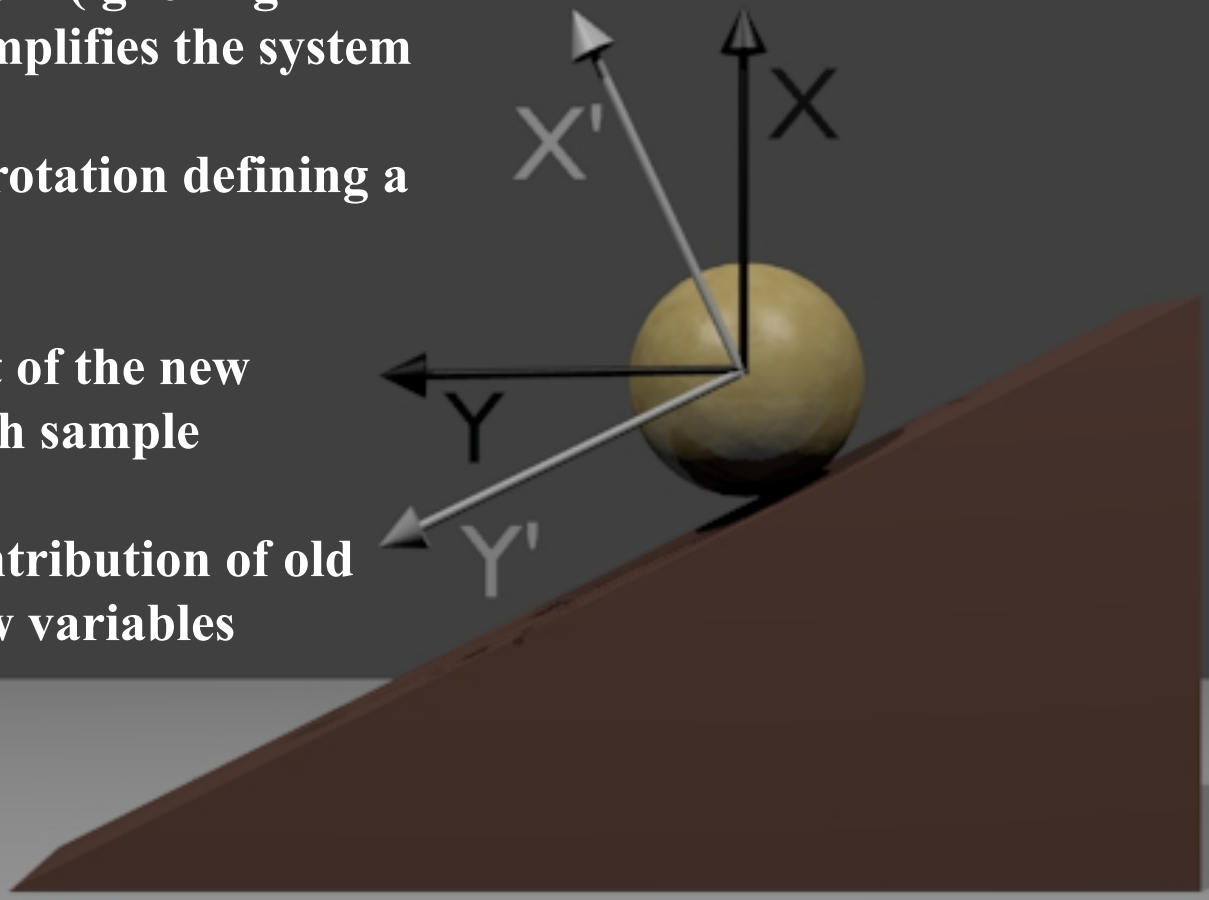
Multivariate Analysis
Reducing the Dimensionality of a Problem

Rotate the axes to creating a new coordinate system (ignoring gravity) that simplifies the system

PCA is an axis rotation defining a new set of axes

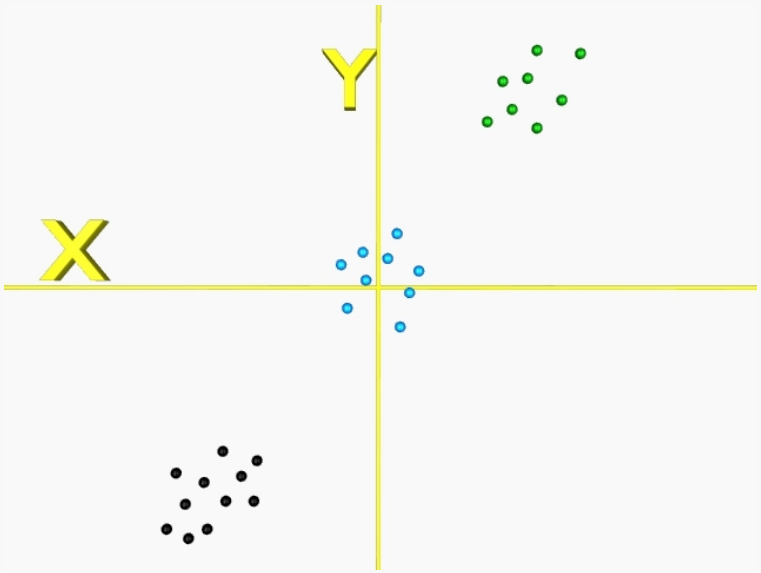
Score = amount of the new variables in each sample

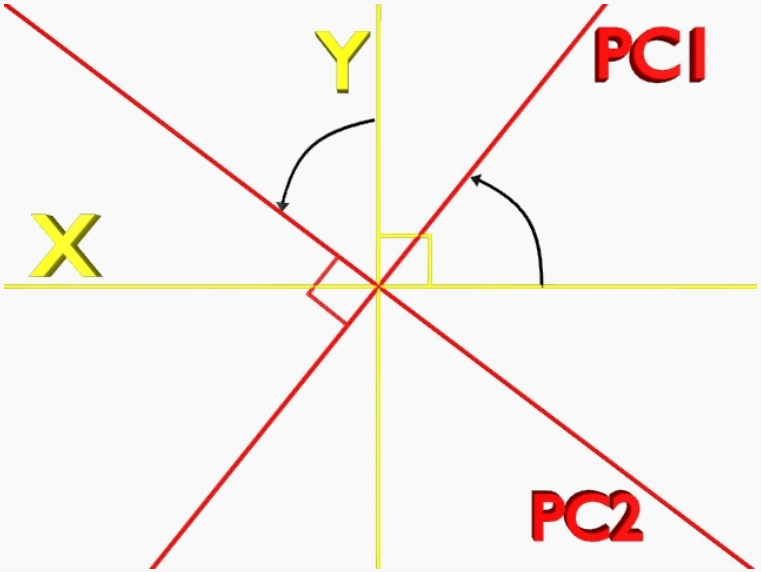
Loadings = Contribution of old variables to new variables

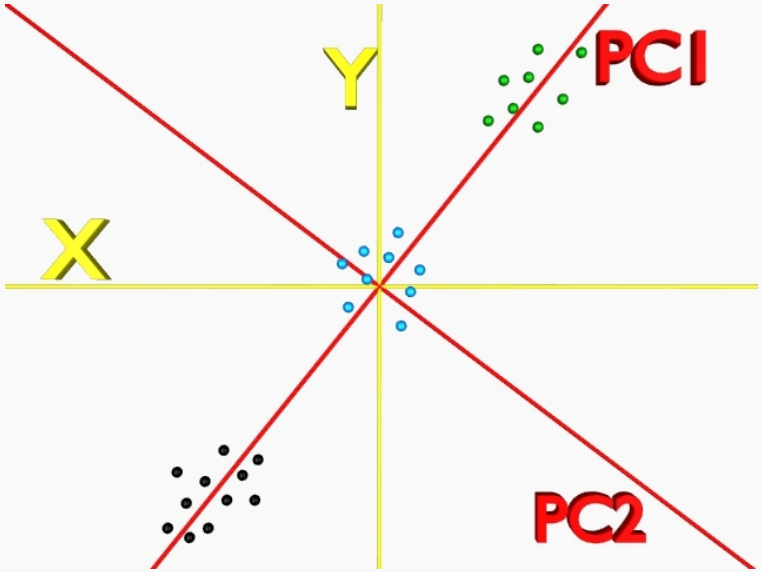


Slide adapted from slide Bonnie Tyler

MVA Geometrically



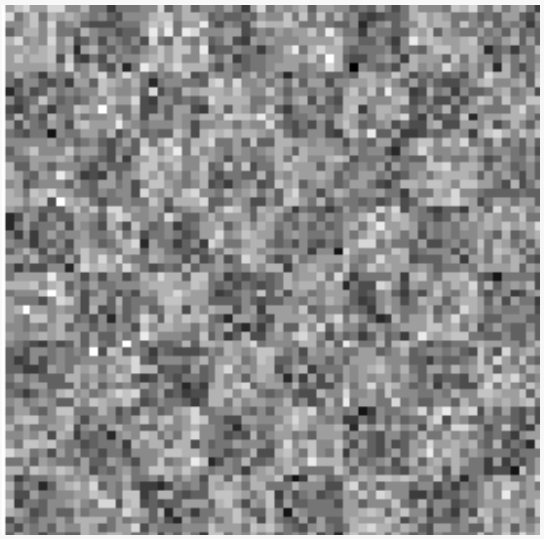


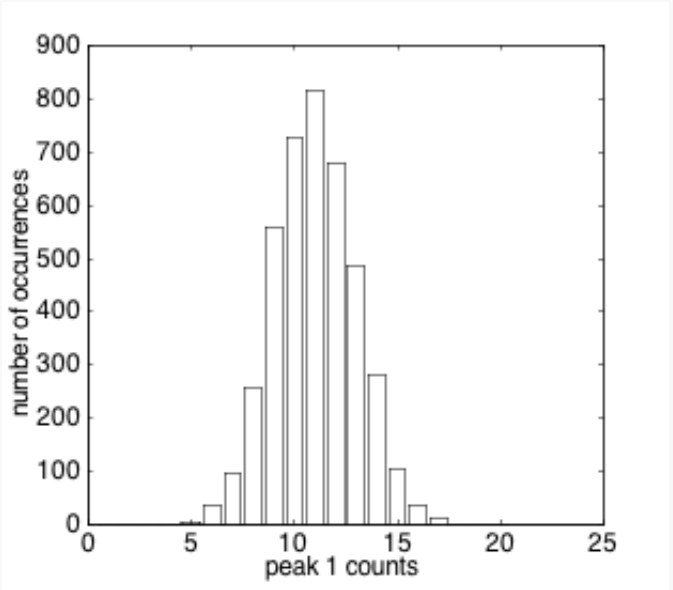


How Axis Rotation Can Help

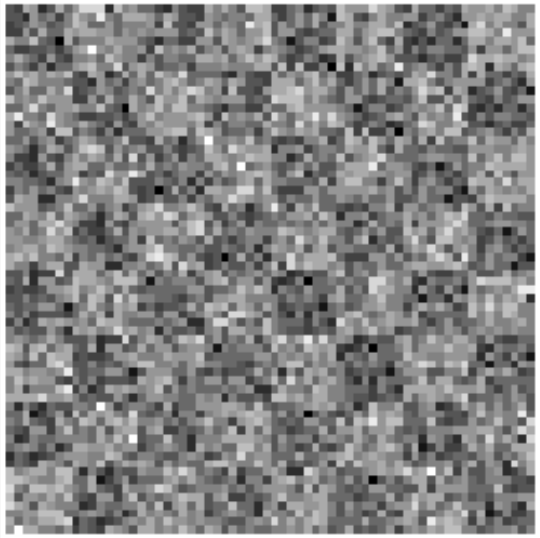
Computer Generated Image

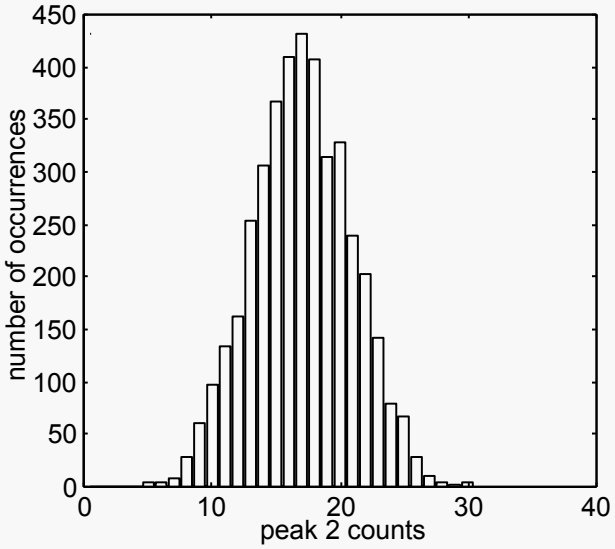
Peak 1 Image





Peak 2 Image

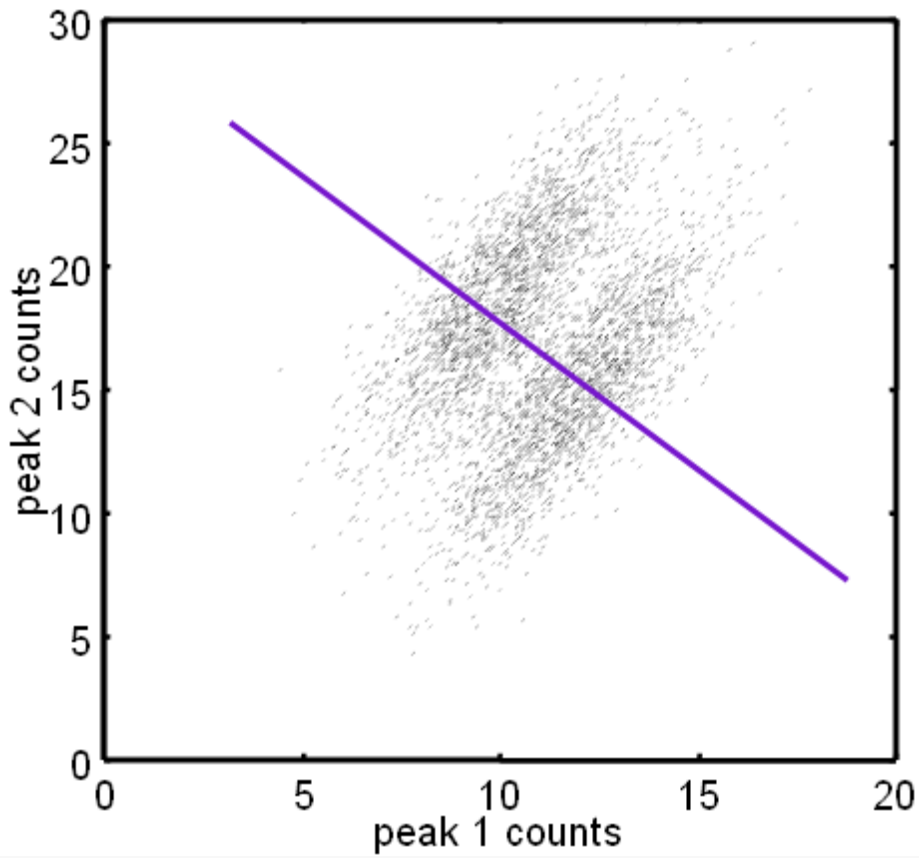




Slide courtesy of Bonnie Tyler

How Axis Rotation Can Help

Computer Generated Image

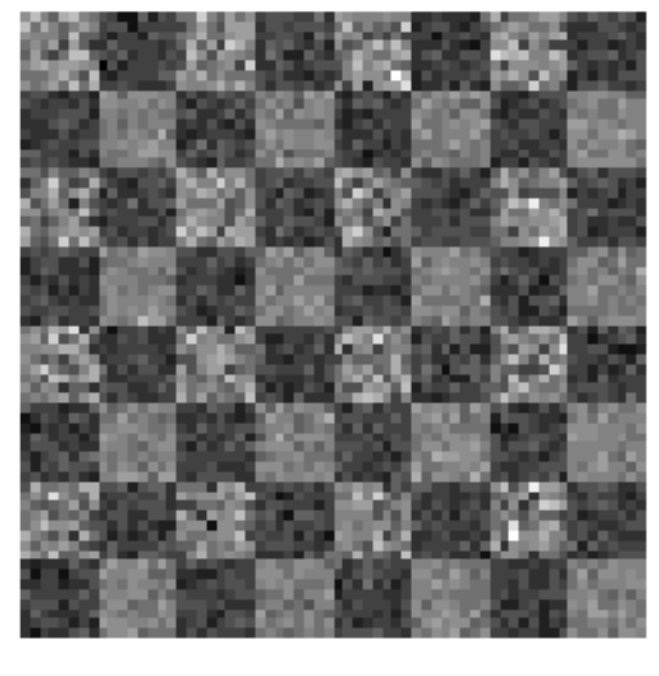


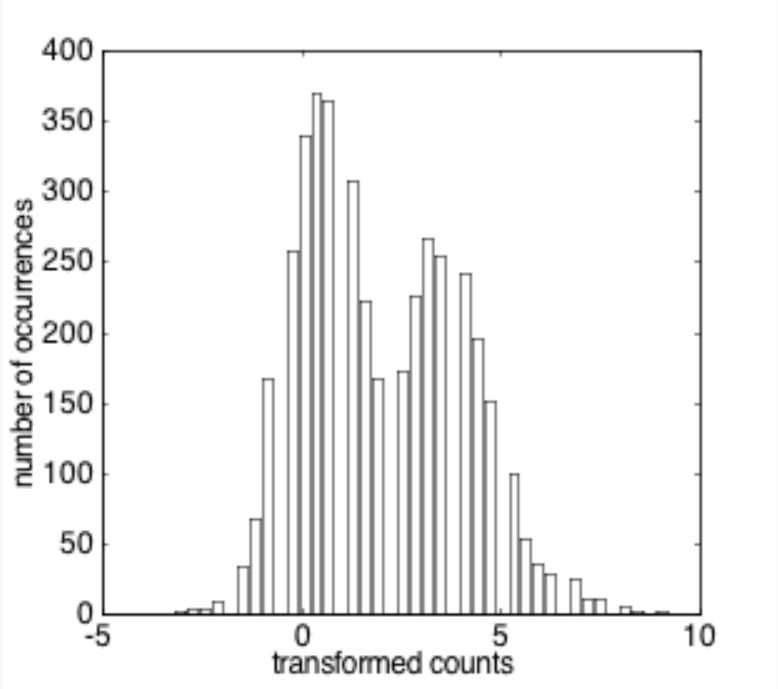
Slide courtesy of Bonnie Tyler

How Axis Rotation Can Help

Computer Generated Image

Transformed Variable Image and Histogram





Slide courtesy of Bonnie Tyler

The New Coordinate System

How do we decide how to rotate the axes?

Problem	Method
Sources of Variation in a Data Set	PCA, MCR, MAF
Prediction of Properties from Data Set	PLS, PCR
Differences Between Analysis,HCA, Groups or Classes	Discriminant PLS-DA
Find Features in Noisy Images	MAF

- Choose the method based off the goals of the analysis
  - No one method is "Better" than another
  - They are all just tools

Math

- MVA methods are based of linear algebra and matrix math
- Good reviews of PCA math can be found in the literature
  - Chemom. Intell Lab Sys, 1987, 2, 37-52
  - J Qual Technol, 1980, 12, 201-213
  - Jackson JE, 1991, User's guide to principal components, John Wiley & Sons, NY

Principal Components Analysis

- **Variance**
- **A measure of the spread in the data**

$$S^2 = \frac{\sum x - \bar{x}}{n - 1}$$

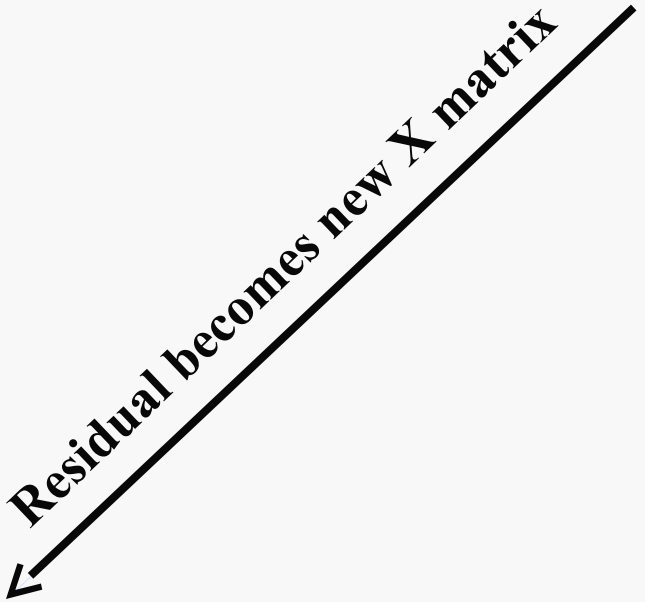
- **Covariance**
- **A measure of the degree that two variables vary together**
- **PCA is calculated from the covariance matrix**

$$\mathrm{cov}(X) = \frac{X^T X}{m - 1}$$

PCA Methodology

- **PCA determines sequential orthogonal axes that capture the greatest direction of variance within the data**

$$X = T1P1T + E$$



$$X = T2P2T + E$$

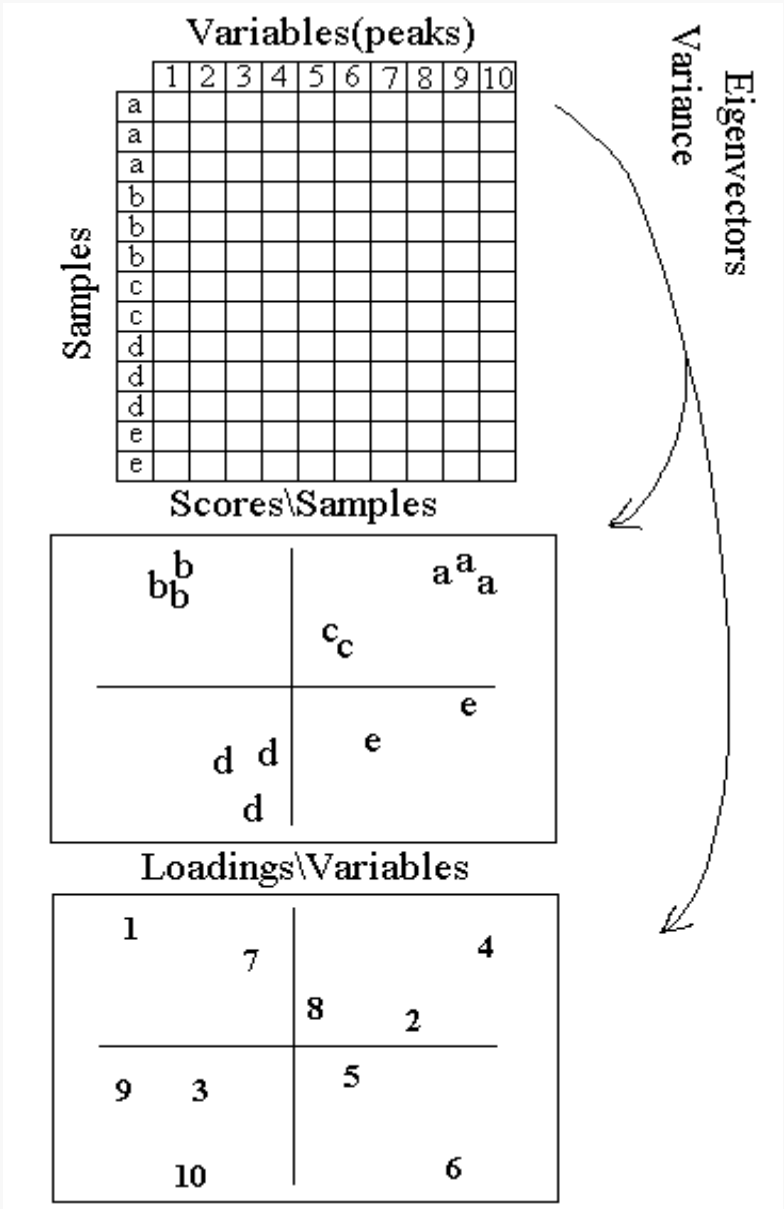
$$\text{var(PC1)} > \text{var(PC2)} > \text{var(PC3)} > \ldots > \text{var(PCk)}$$

What the Math Means

- Scores ---> “concentration”
- Loadings ---> “spectra”
- However these are not usually “concentrations” or “spectra” of a pure component or even an individual physical factor

PCA

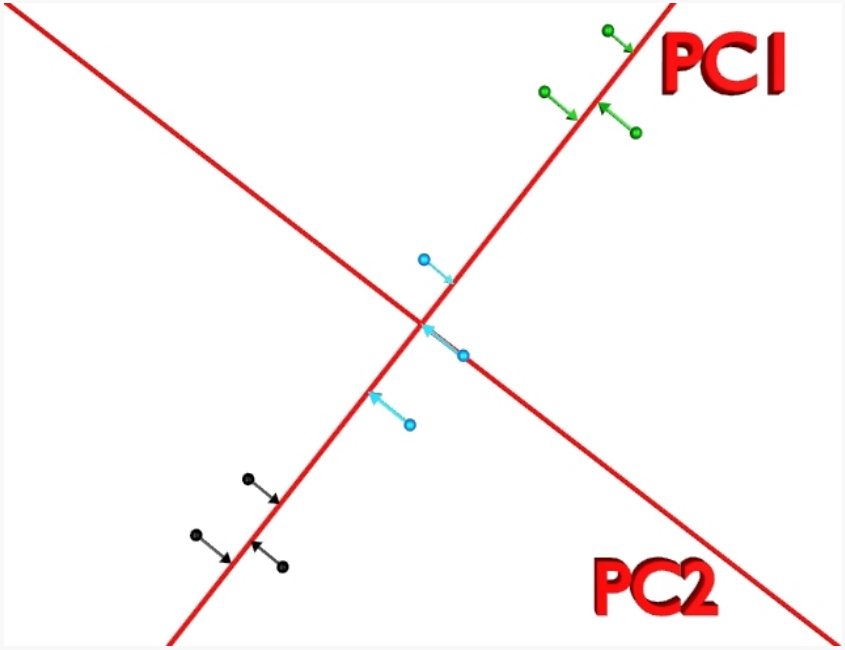
- **Looks at the variance patterns of a data matrix**
- **Reduces data dimensionality**
- **Gives simple graphical presentation of data**
- **Determines relationship of samples and variables based on the variance in the data**
- **No external constraints needed**
- **Original matrix is reconstructed into new matrices that define the major patterns of the data in multivariate space**
  - **SCORES -> Describe relationship between samples (spread) as described by PC's**
  - **LOADINGS -> Describe how the variables relate to the PC's**



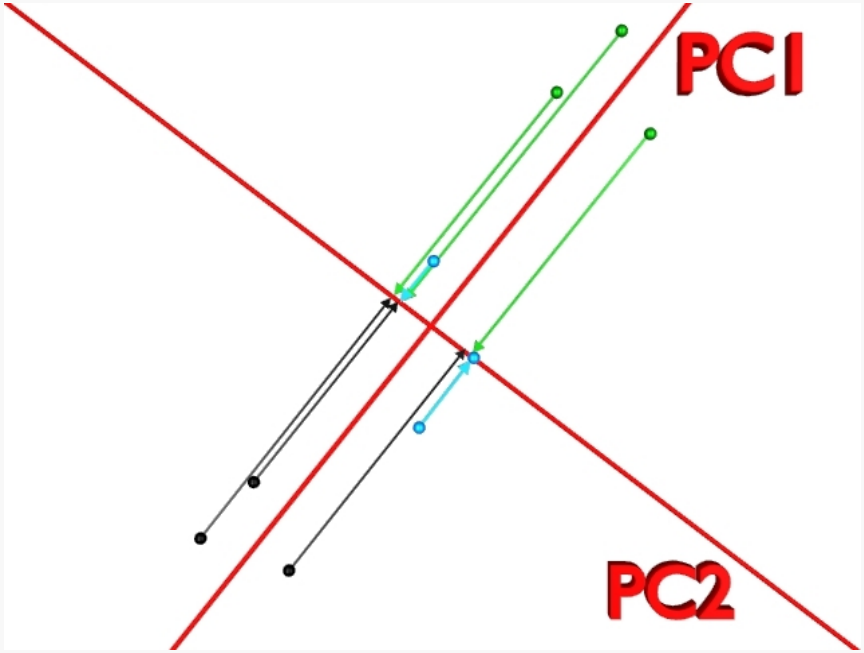
Scores

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

Scores tell the relationship (spread) between the samples



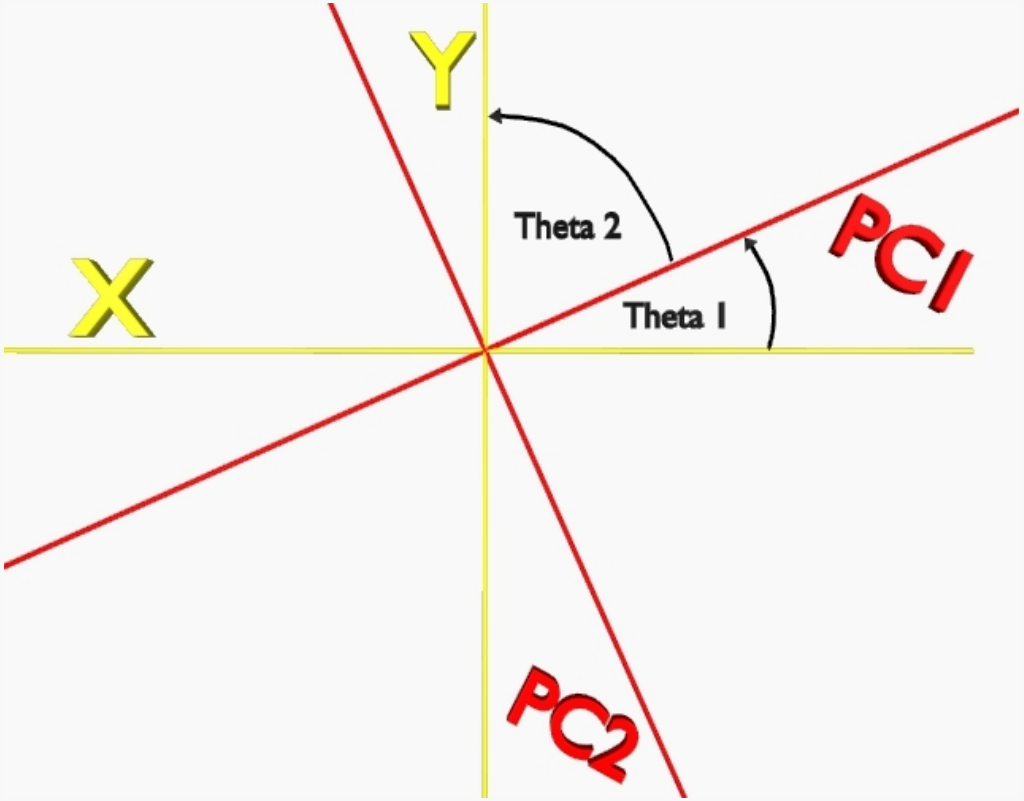
Projection onto PC1



Projection onto PC2

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**

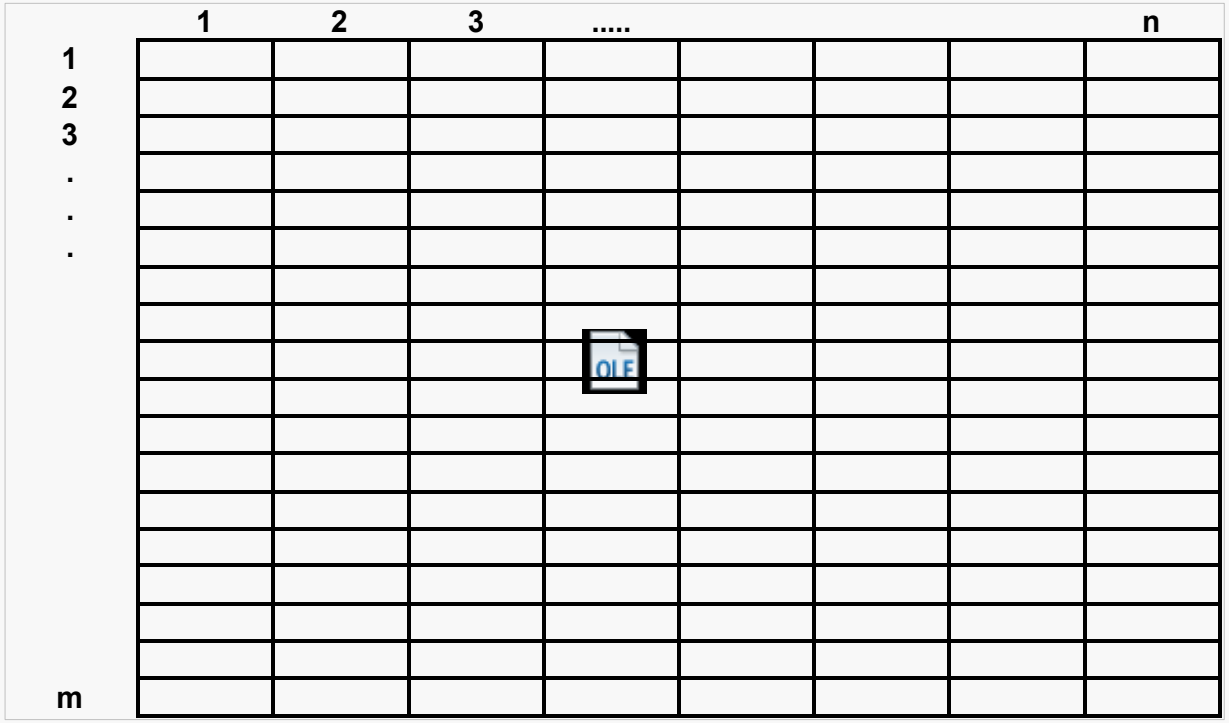
HOW?

Background Information

- Data is arranged in matrices
  - samples in rows
  - variables in columns
- m = number of samples
- n = numbers of variables
- k = number of PCs
- T = scores matrix
- P = loadings matrix

Data Matrix

Variables



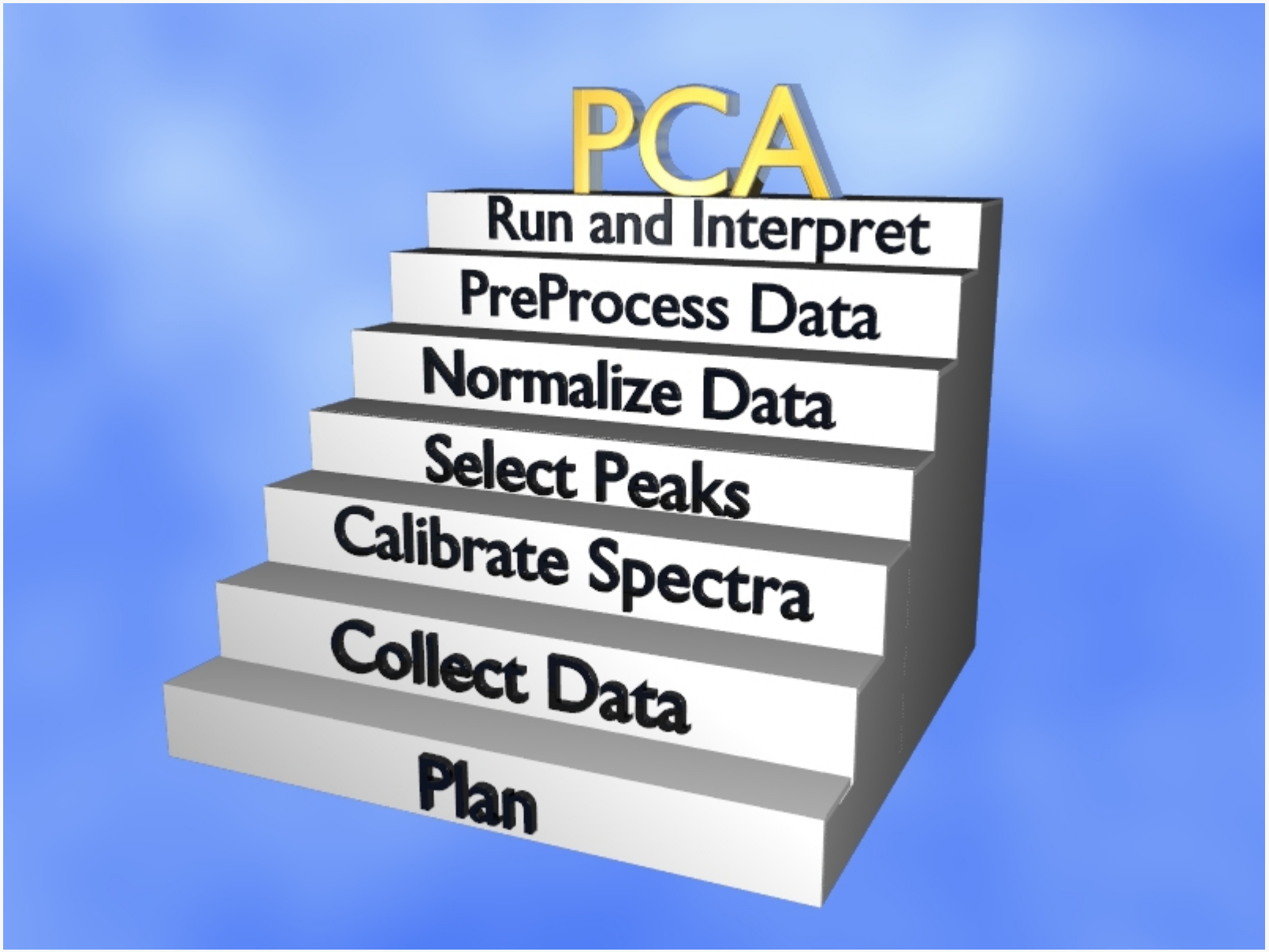
For SIMS data the "samples" are SIMS spectra, or more typically the integrated areas for all peaks for a given spectra

- **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: Things to know

- PCA assumes linear relationships between variables
- PCA is scale dependent
  - variables with larger values look more important
- PCA looks at variance in the data
  - It will highlight whatever the largest difference are
  - To make sure you are comparing things properly it is common to preprocess the data
    - Remove any instrument variation, or other non-related variance (normalization)
    - Make sure data is compared across a common mean (centering)
    - Make sure data is compared across common variance scale (autoscaling, variance scaling, etc)

Steps to PCA


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

Proper ToF-SIMS Analysis Requires:

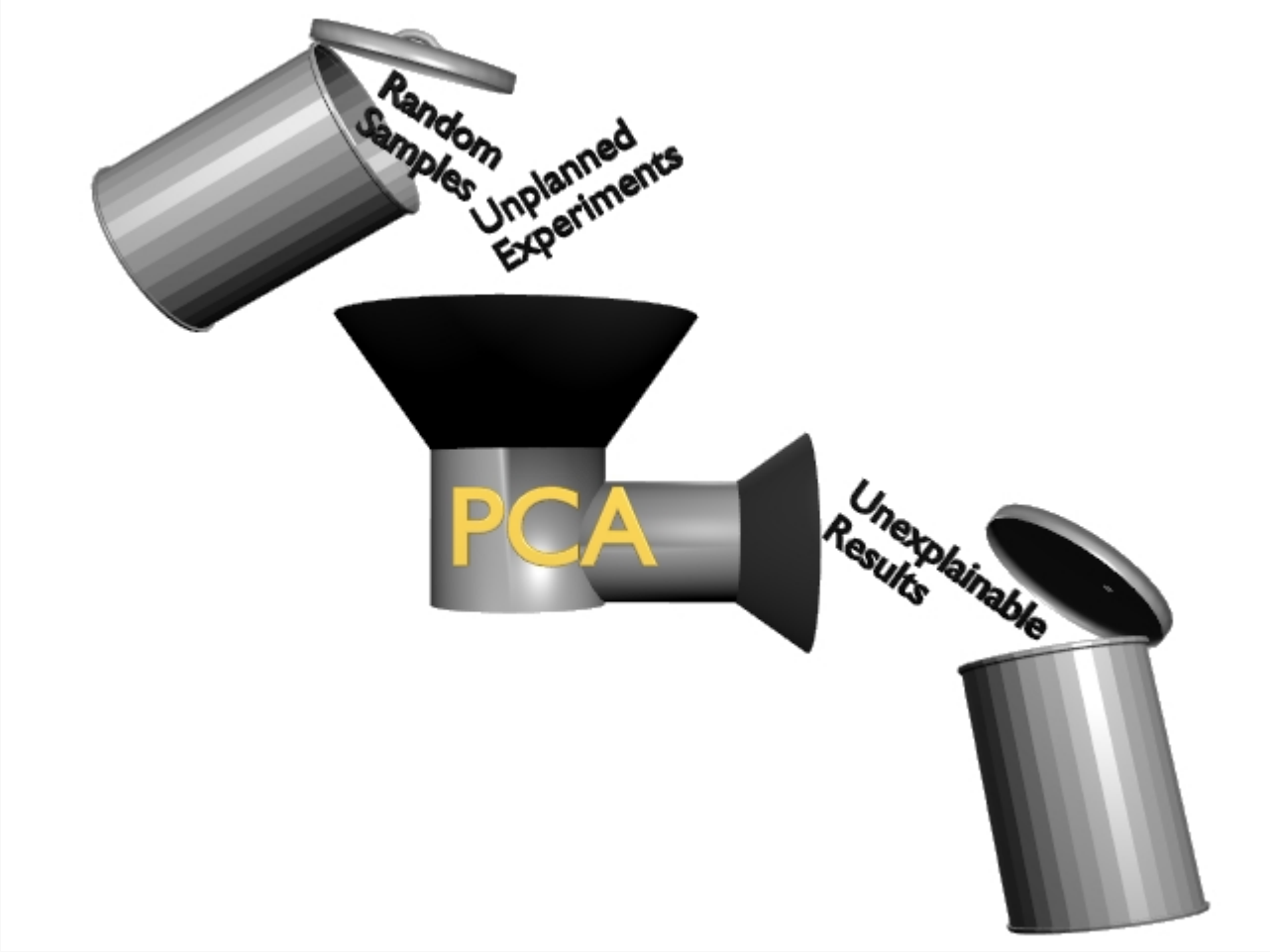
- Good experimental plans (controls)
- Proper sample preparation
- Careful data collection
- Consistent data calibration
- Sound understanding of the fundamentals of mass spectral analysis
- Knowledge of how to properly use the available tools to help with the analysis

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

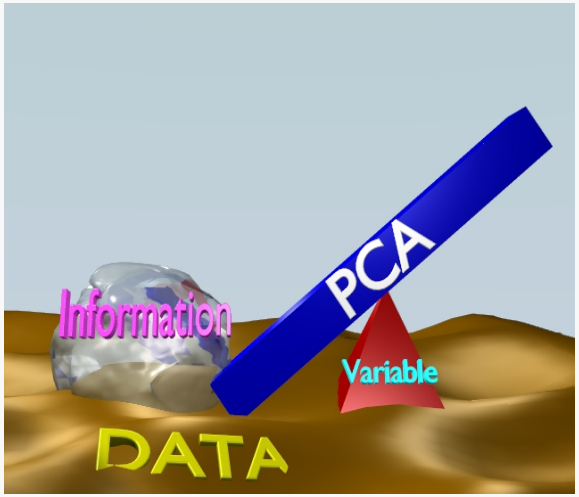
If you input garbage in



You will get garbage out!!!

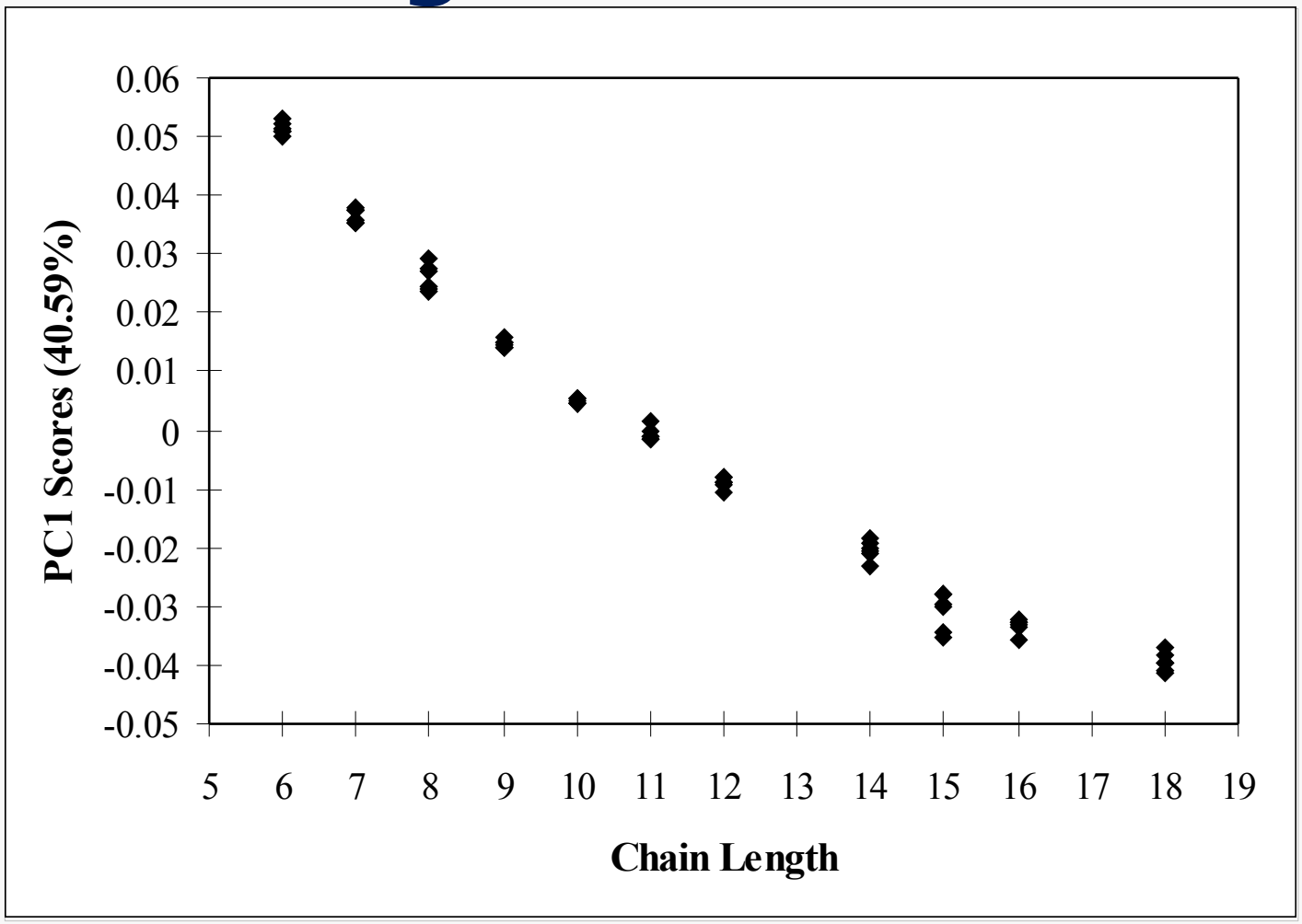
Experimental Design/Data Collection

- 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
    - Non-homogeneous => 5 to 7 spots on 3 to 5 samples

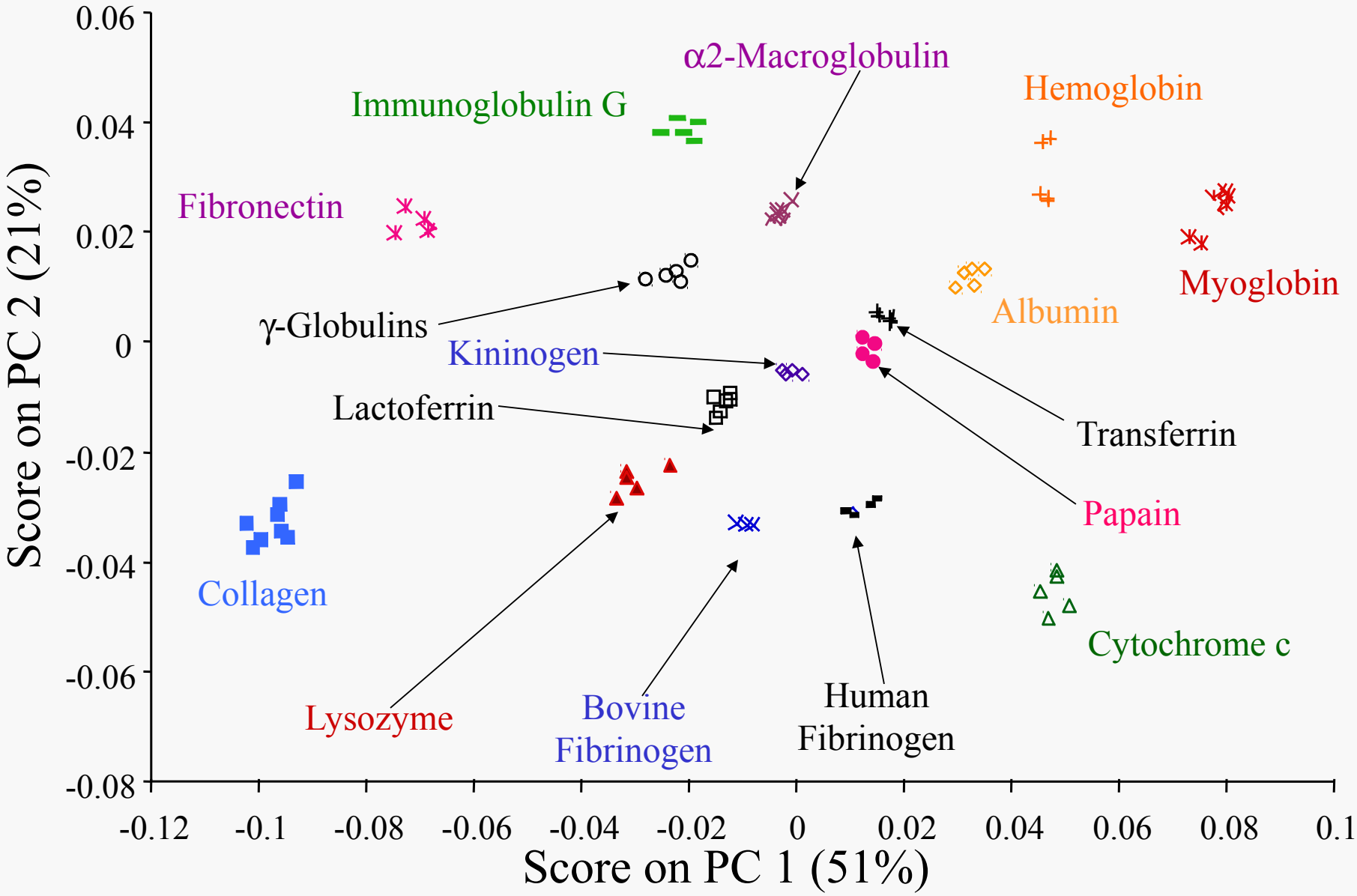


SAMs – typically very homogeneous

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

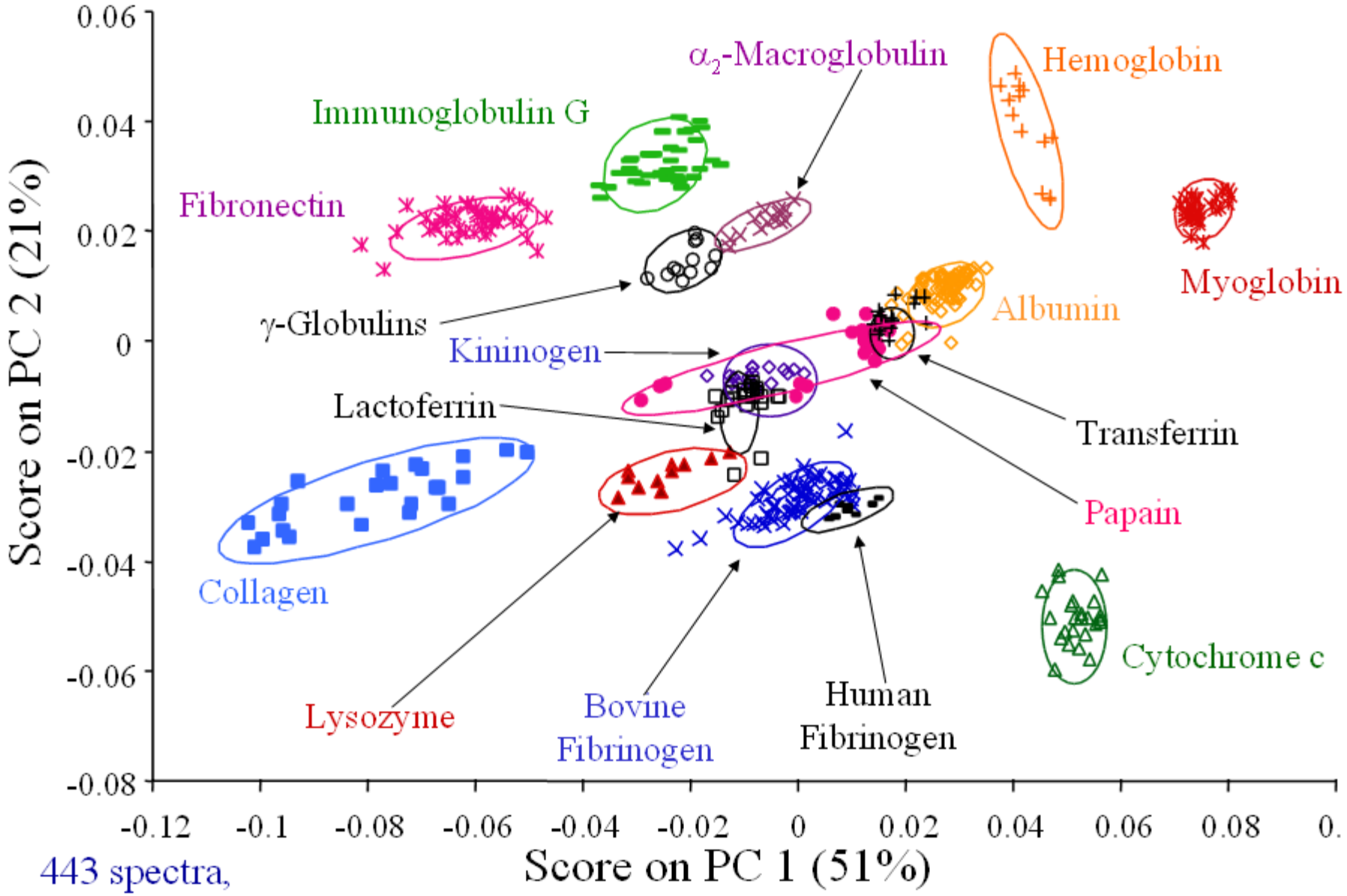


Proteins adsorbed onto Mica: PCA



Modified fromWagner & Castner, Langmuir **17** (2001) 4649.

Proteins adsorbed onto Mica: PCA



443 spectra,
16 different proteins

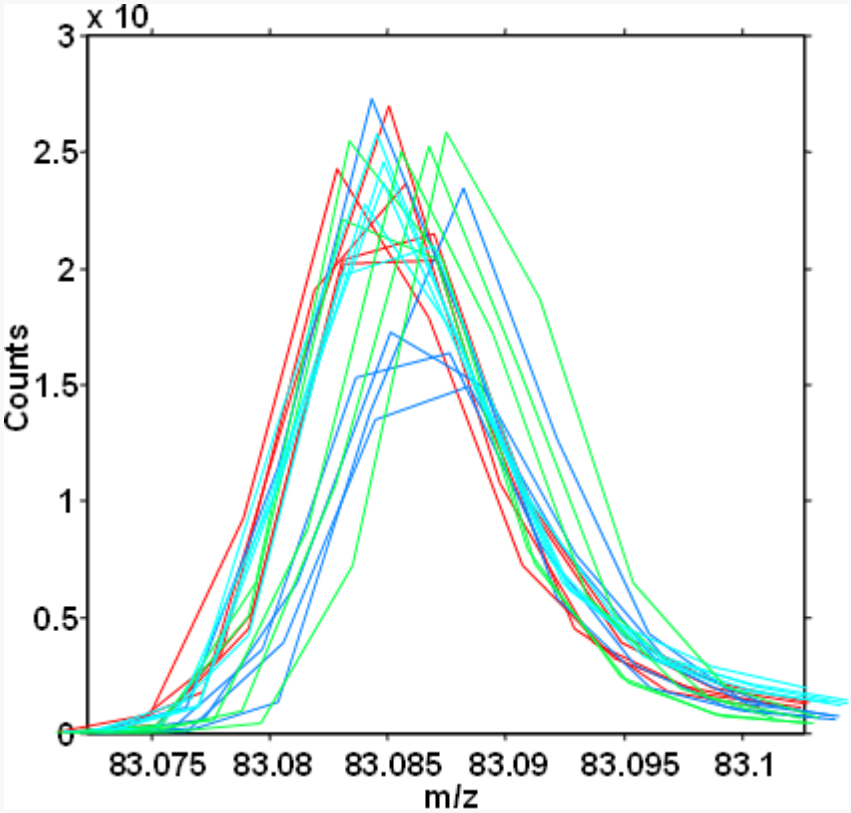
Wagner & Castner, Langmuir **17** (2001) 4649.

Data calibration

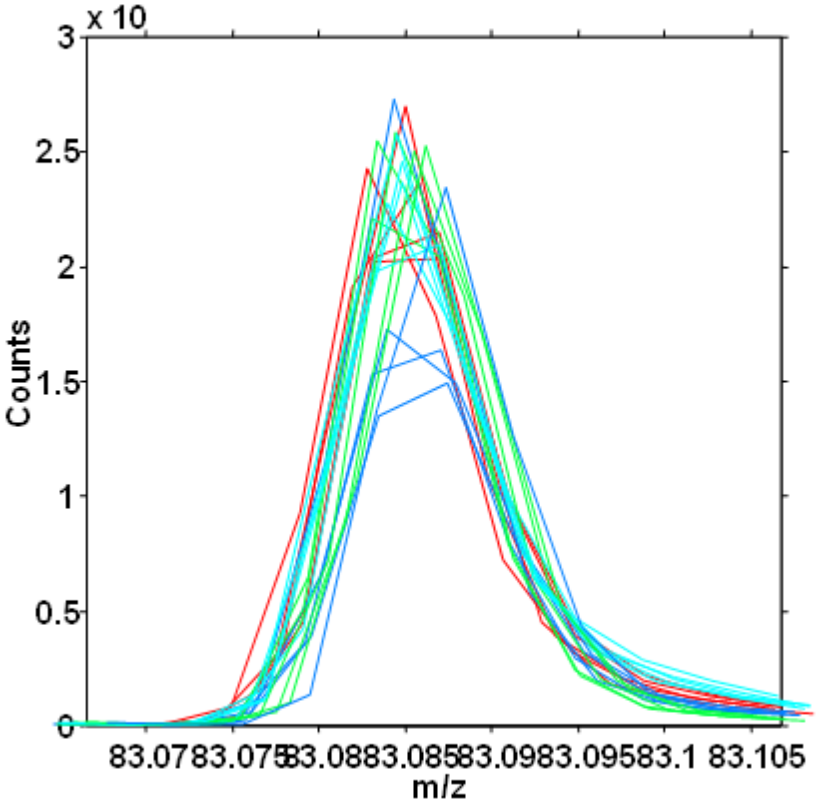
- All spectra in the data set should be calibrated to the same peak set
  - Be consistent
- Include a high-mass peak if possible
  - This will increase the accuracy of identifying high mass peaks
- Always double check autocalibration functions
  - They can make mistakes

Calibration example

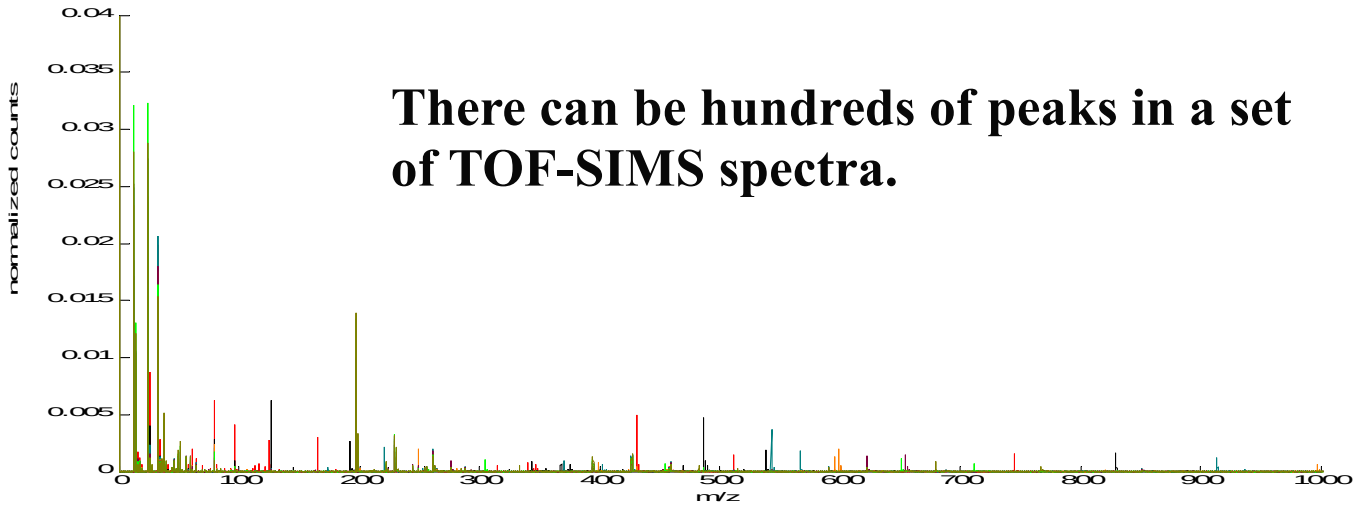
Initial Calibration



After Checking Calibration to assure consistency

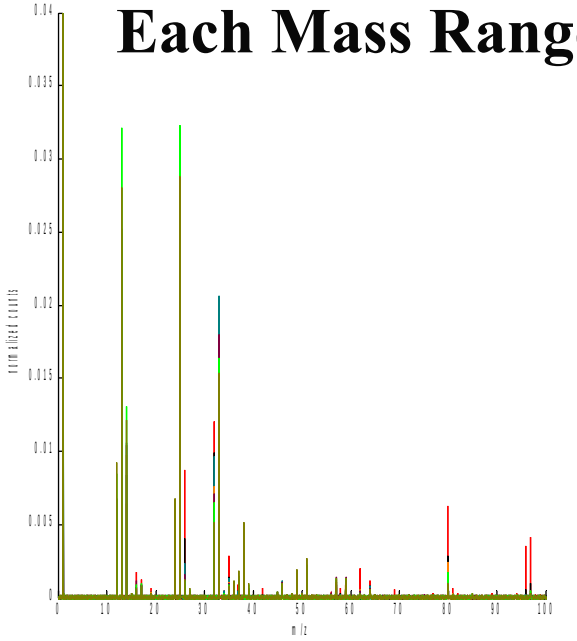


Peak Selection-Which Peaks should you select?

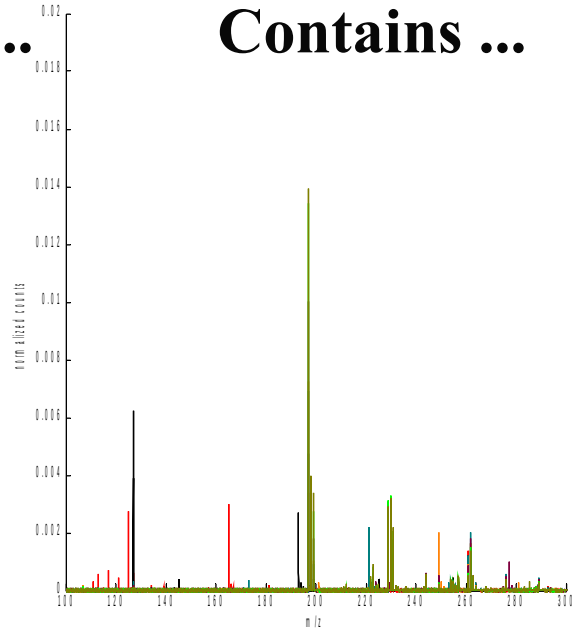


There can be hundreds of peaks in a set of TOF-SIMS spectra.

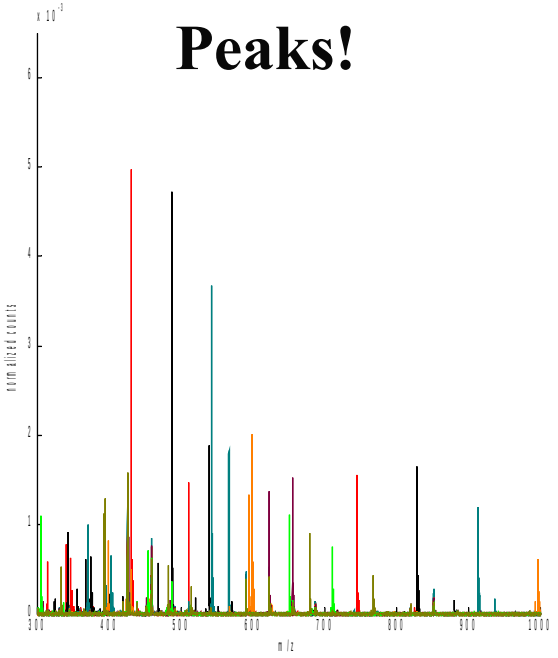
Each Mass Range...



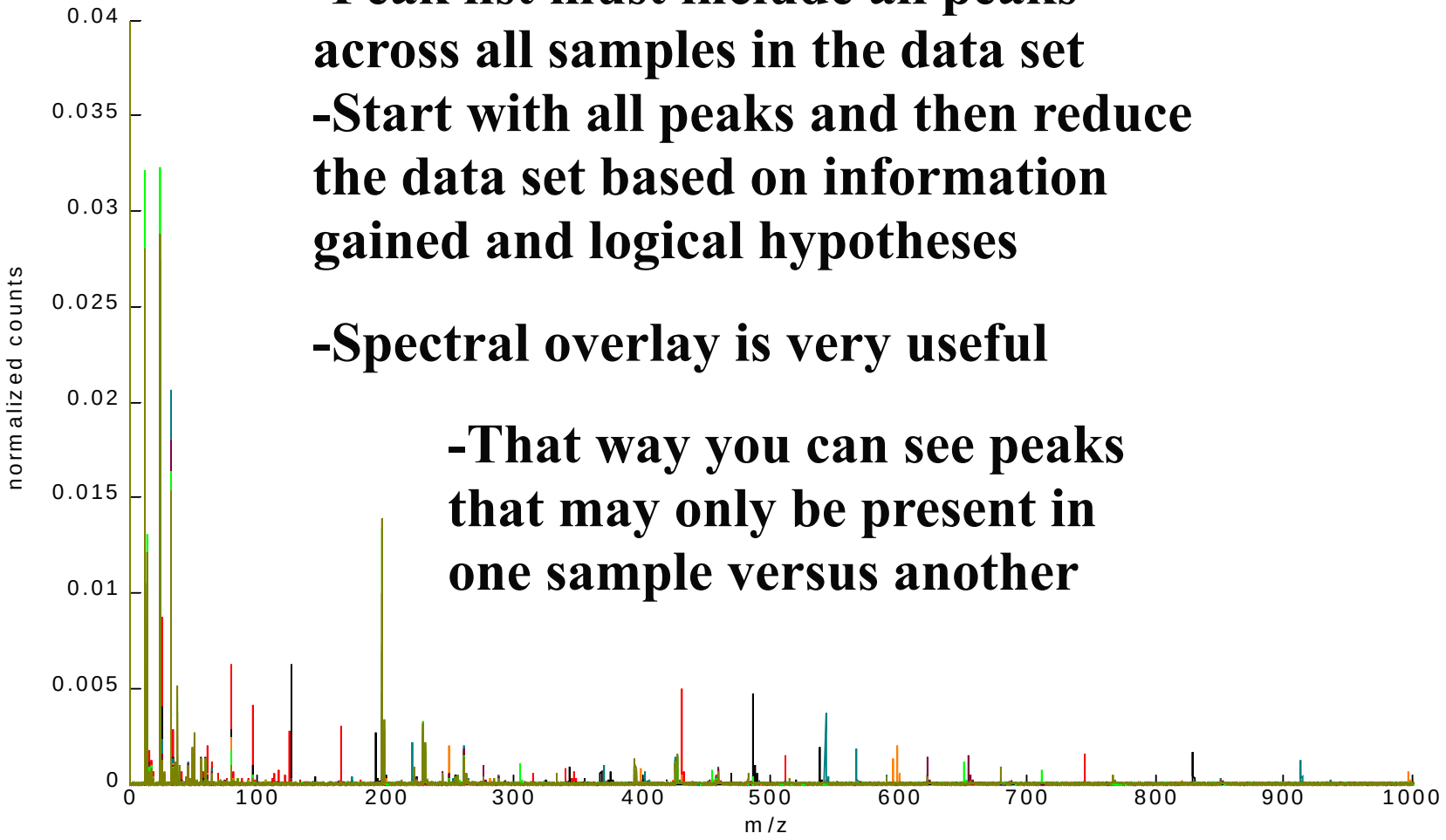
Contains ...



Peaks!



Peak Selection-Which Peaks should you select?



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

-Start with all peaks and then reduce the data set based on information gained and logical hypotheses

-Spectral overlay is very useful

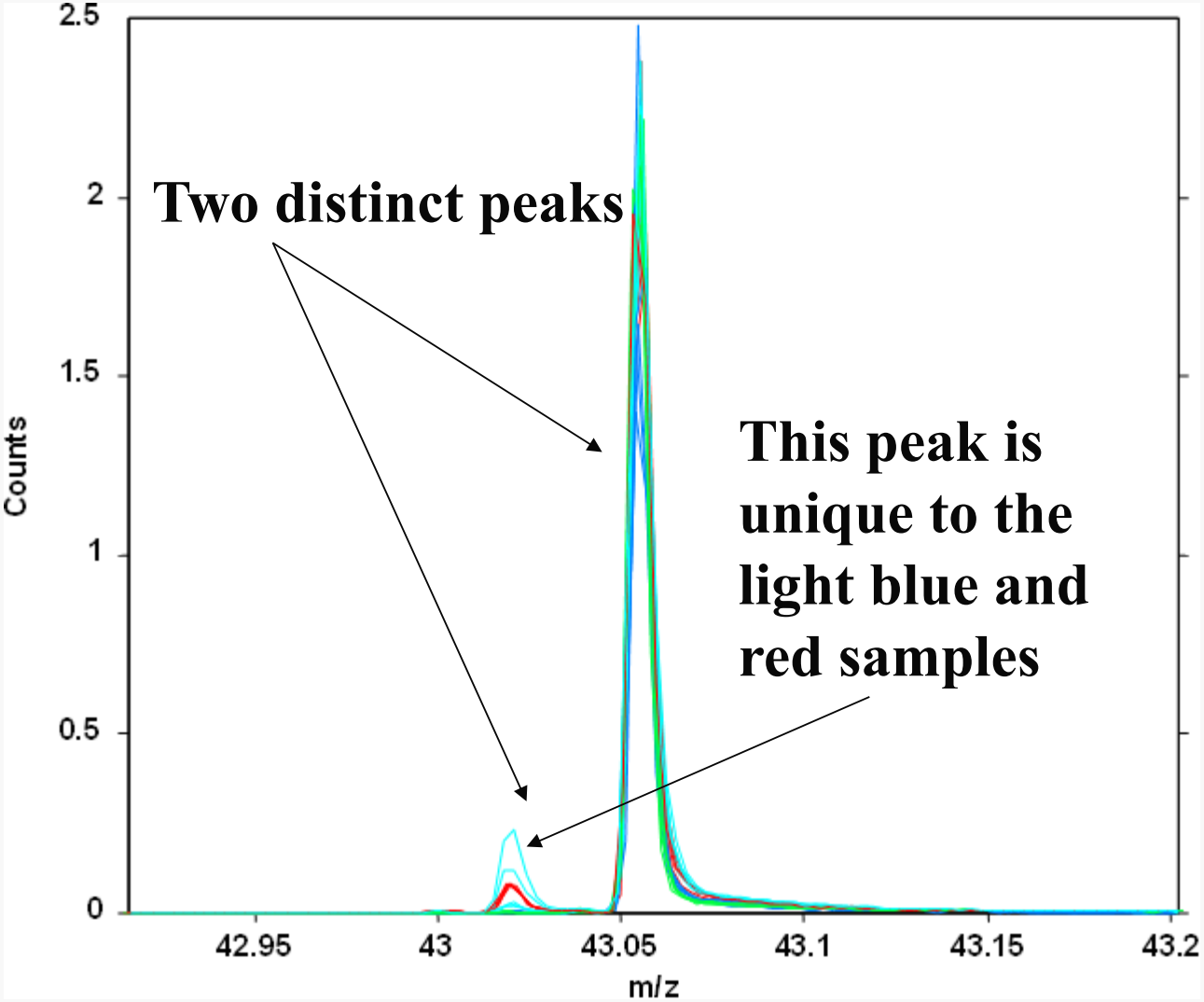
-That way you can see peaks that may only be present in one sample versus another

Peak Selection Continued

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

-Manual peak selection is recommended

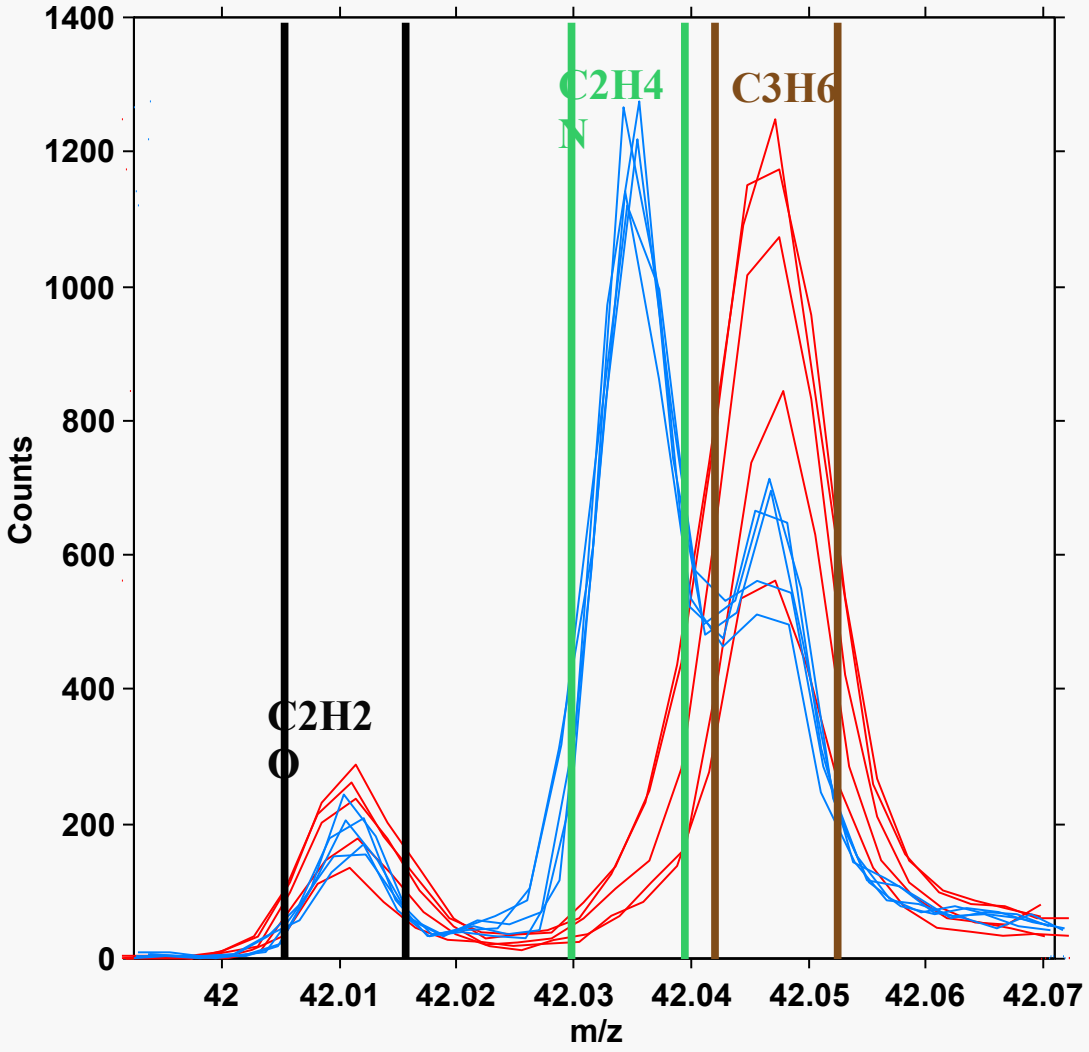
-It is time consuming, but not prohibitive



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

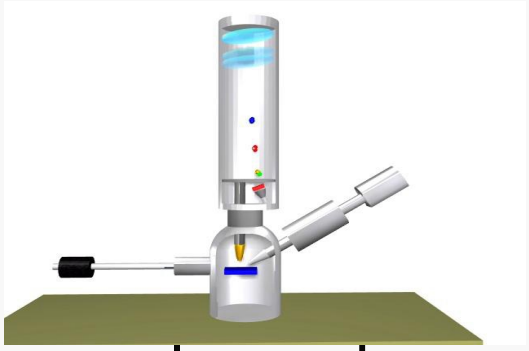
- Typical data pretreatments include
  - Normalization
  - Centering
  - Scaling
- Pretreatments are done in an attempt to maximize differences due to sample differences and minimize differences from other sources
- Know the assumptions being made
  - Are they valid?

PCA data Pretreatment

- No standards have been set for data pretreatment
- Some common trends include
  - normalizing the data (many different ways)
  - Square root transform or divide by the square root of the mean and then mean center for TOF-SIMS spectra
  - Poisson scaling or square root transformation for TOF-SIMS images

Normalization

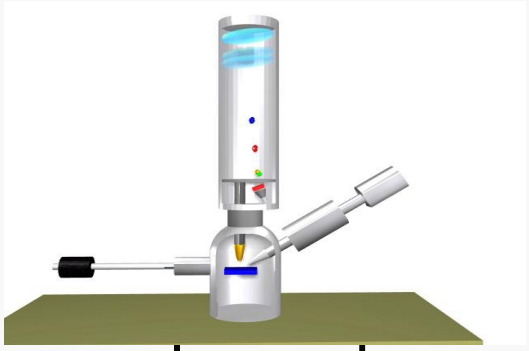
- Data normalization helps account for differences in the data due
  - *topography*
  - sample charging
  - instrumental conditions



- Many different methods are commonly used
  - Total intensity
  - Sum of selected peaks
  - Highest peak in spectrum
  - User selected peak
  - Total intensity minus H and contaminants
- Know assumptions being made
- Understand that normalization removes information from the data set

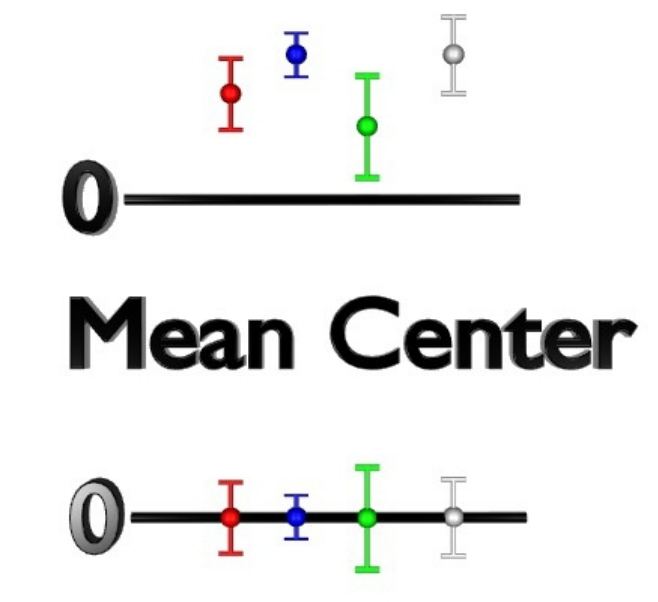
Normalization

- Data normalization helps account for differences in the data due
  - *topography*
  - sample charging
  - instrumental conditions



- Many different methods are commonly used
  - Total intensity
  - Sum of selected peaks
  - Highest peak in spectrum
  - User selected peak
  - Total intensity minus H and contaminants
- Know assumptions being made
- Understand that normalization removes information from the data set

Mean centering



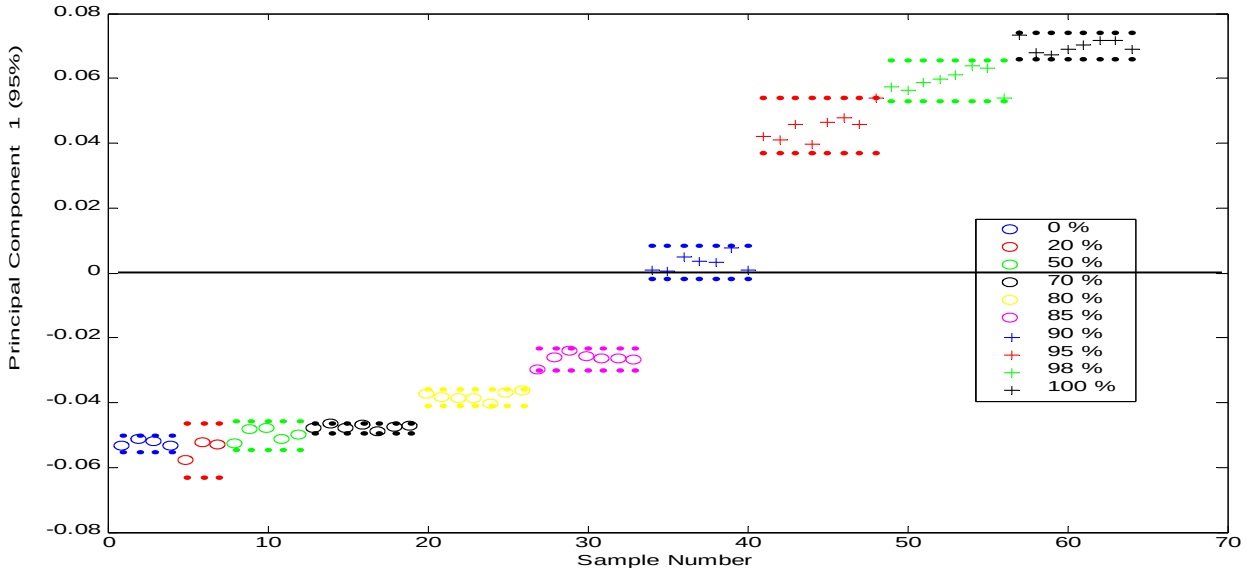
- Mean centering
  - Subtracts the mean of each column (variable) from each column element
  - Centers data so that all variables vary across a common mean of zero

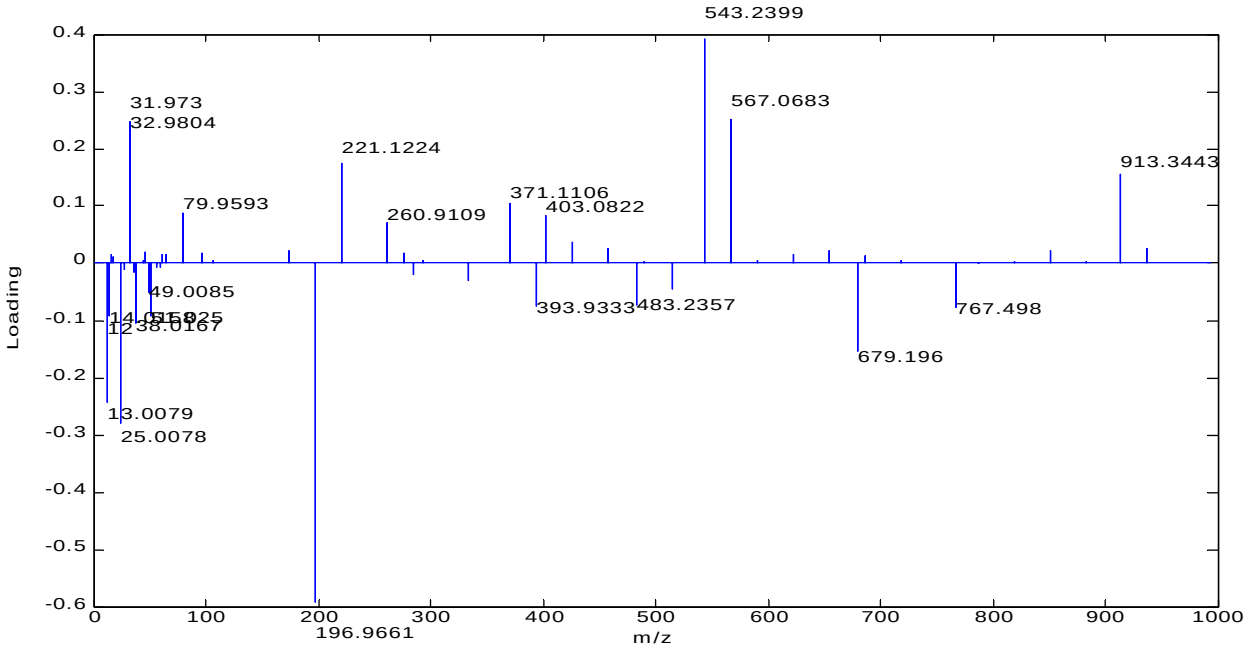
Scaling

- Scaling attempts to account for differences in variance scales between variables
  - Poisson scaling (takes into account Poisson noise structure that is often seen for SIMS data)
  - Dividing by the square root of the mean
  - Square root transform
  - Many others
- Need to know why you are using a given method
- Need to understand assumptions
  - Are the assumptions valid?

Example: Mixed C10 C18 SAMs

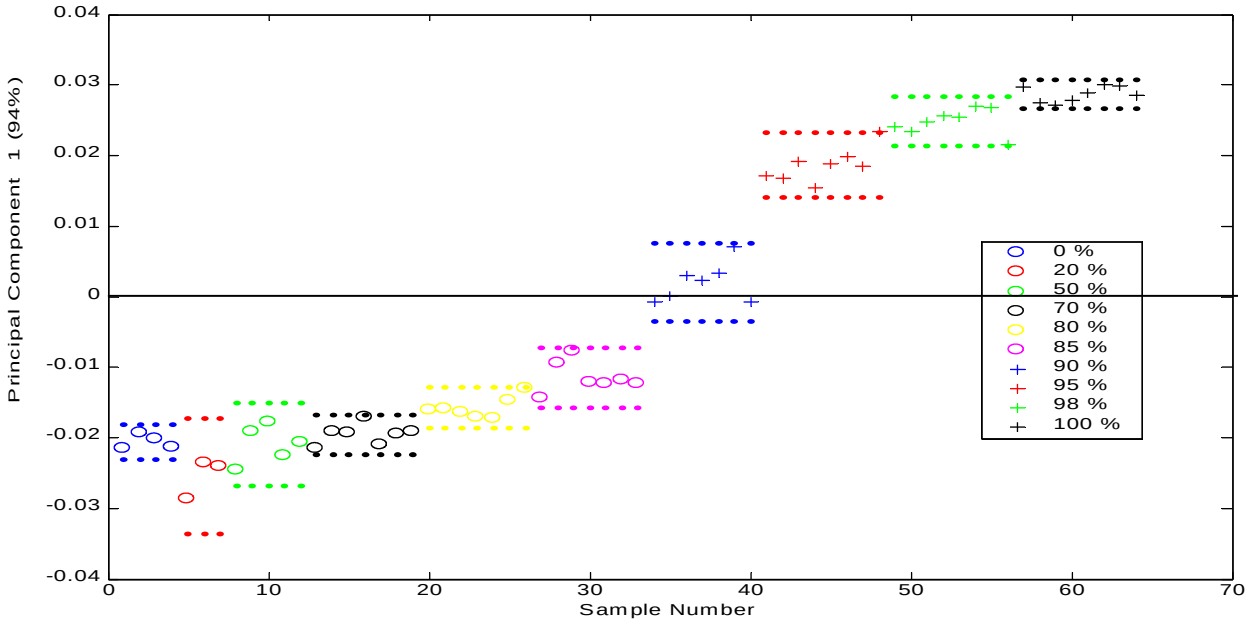
-Normalized Sum of Selected Peaks
-Mean Centered

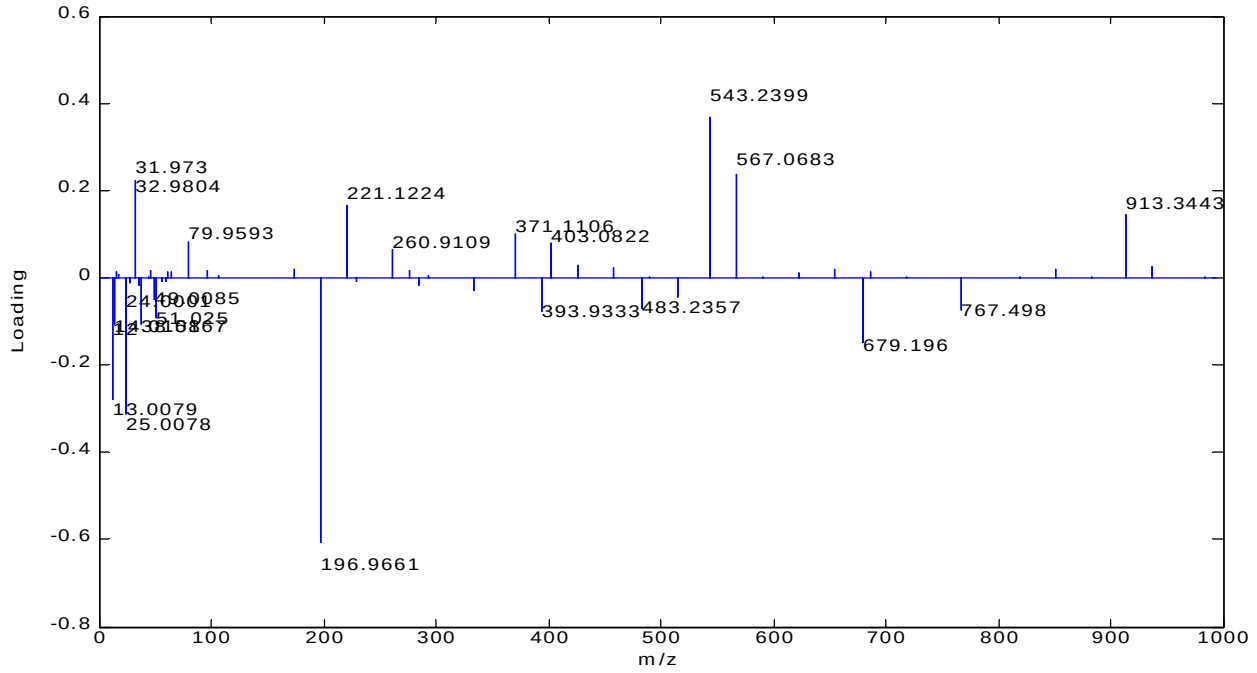




Example: Mixed C10 C18 SAMs

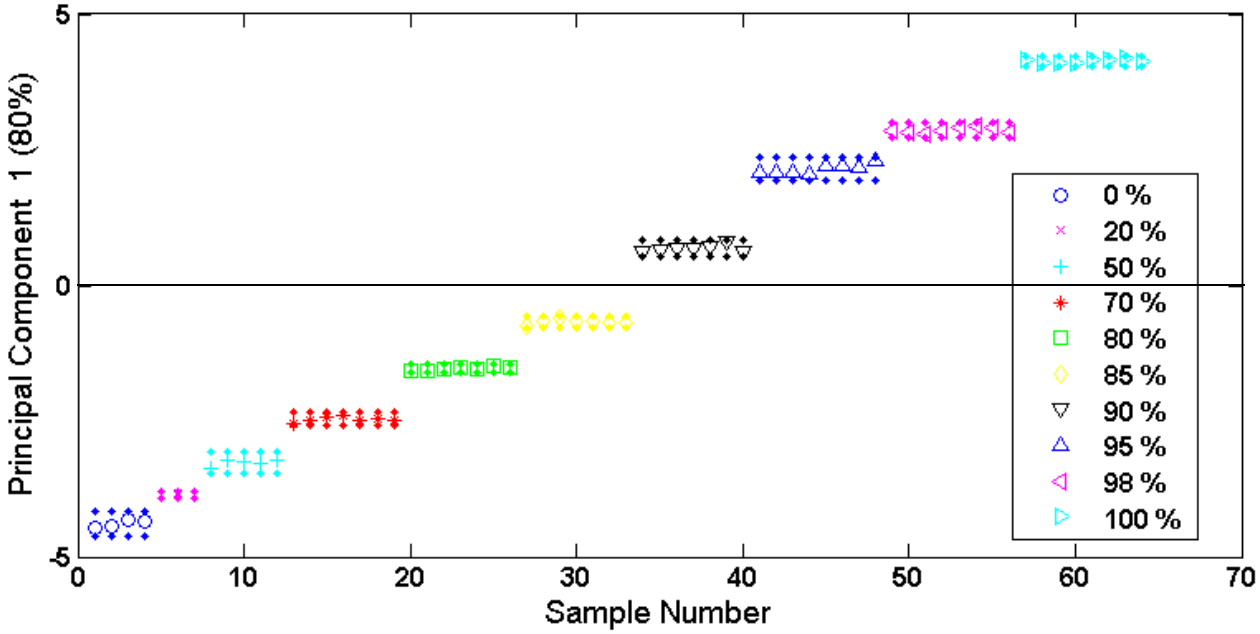
-Normalized Total intensity
-Mean centered

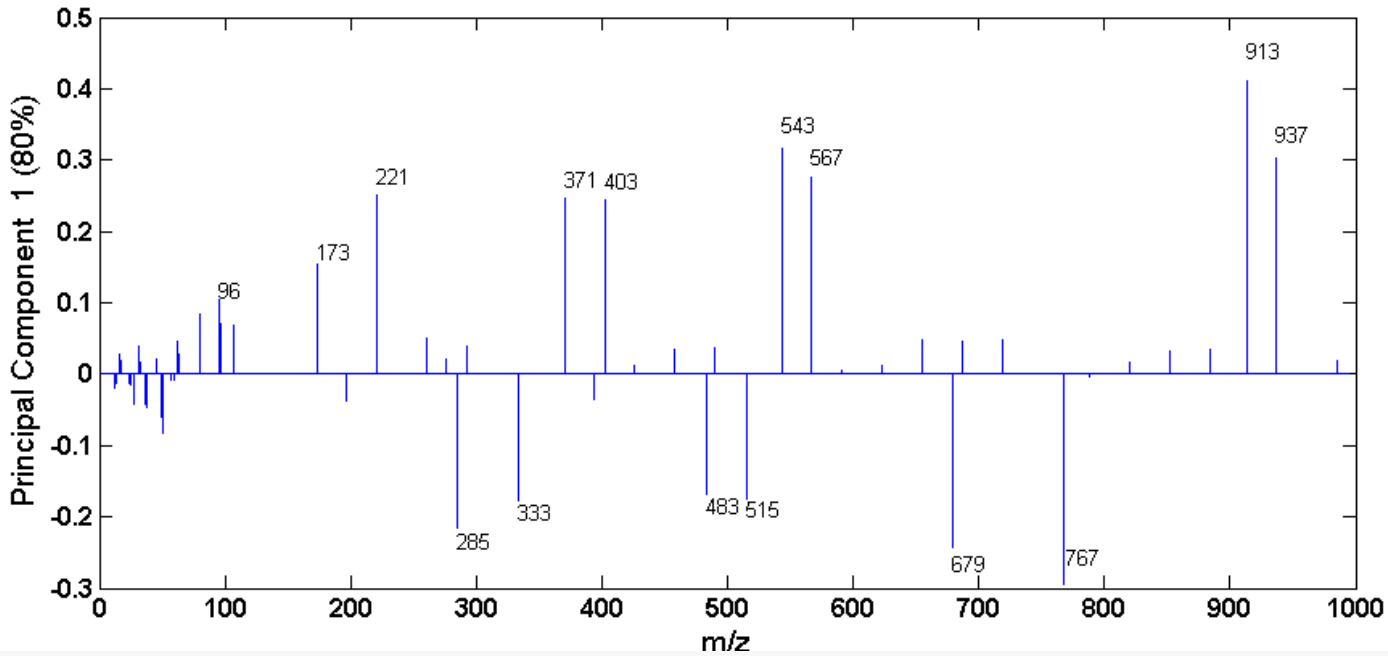




Example: Mixed C10 C18 SAMs

-Normalized Sum of selected peaks
-log10 transformed
-mean centered



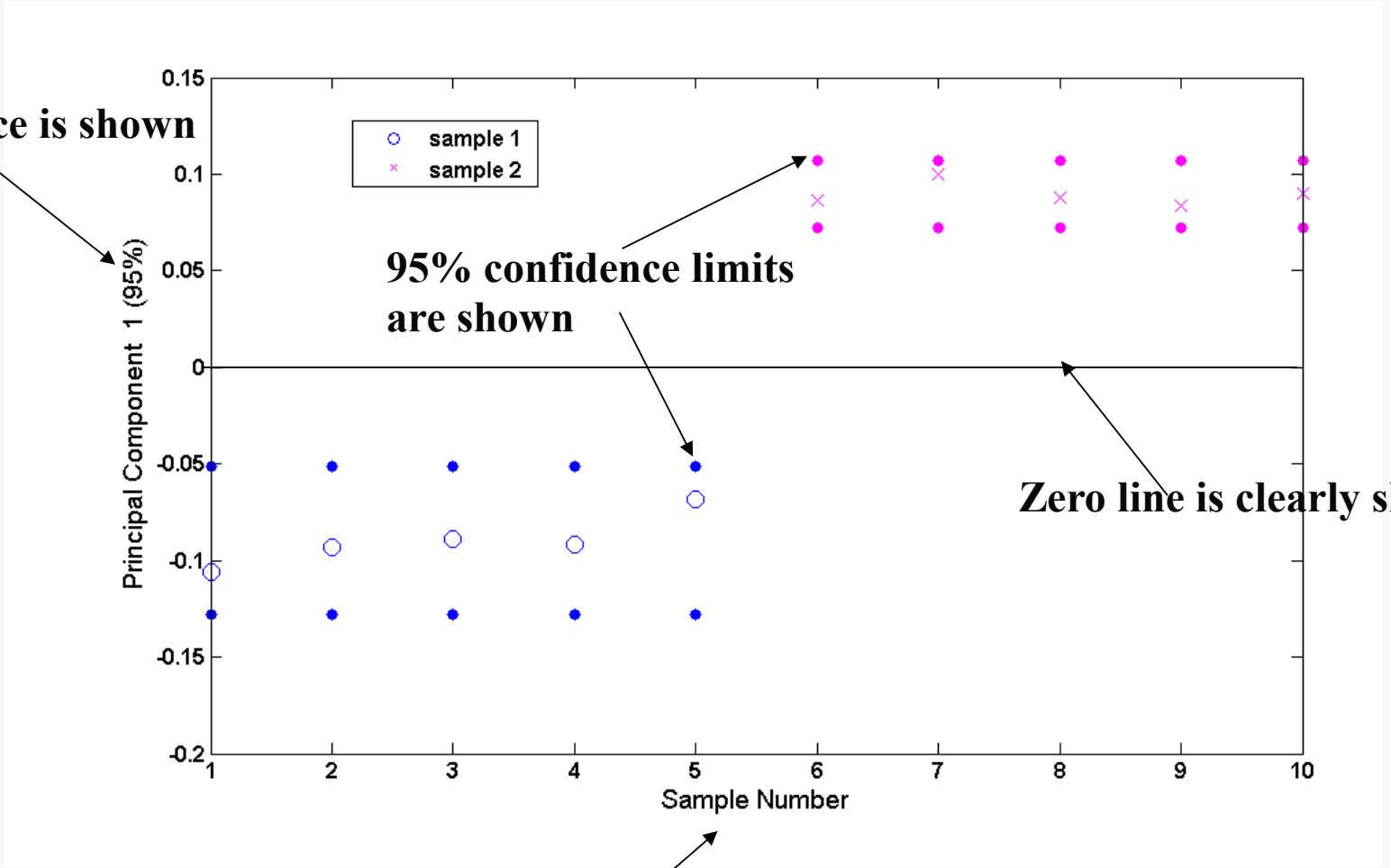


PCA Data Display and Interpretation

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



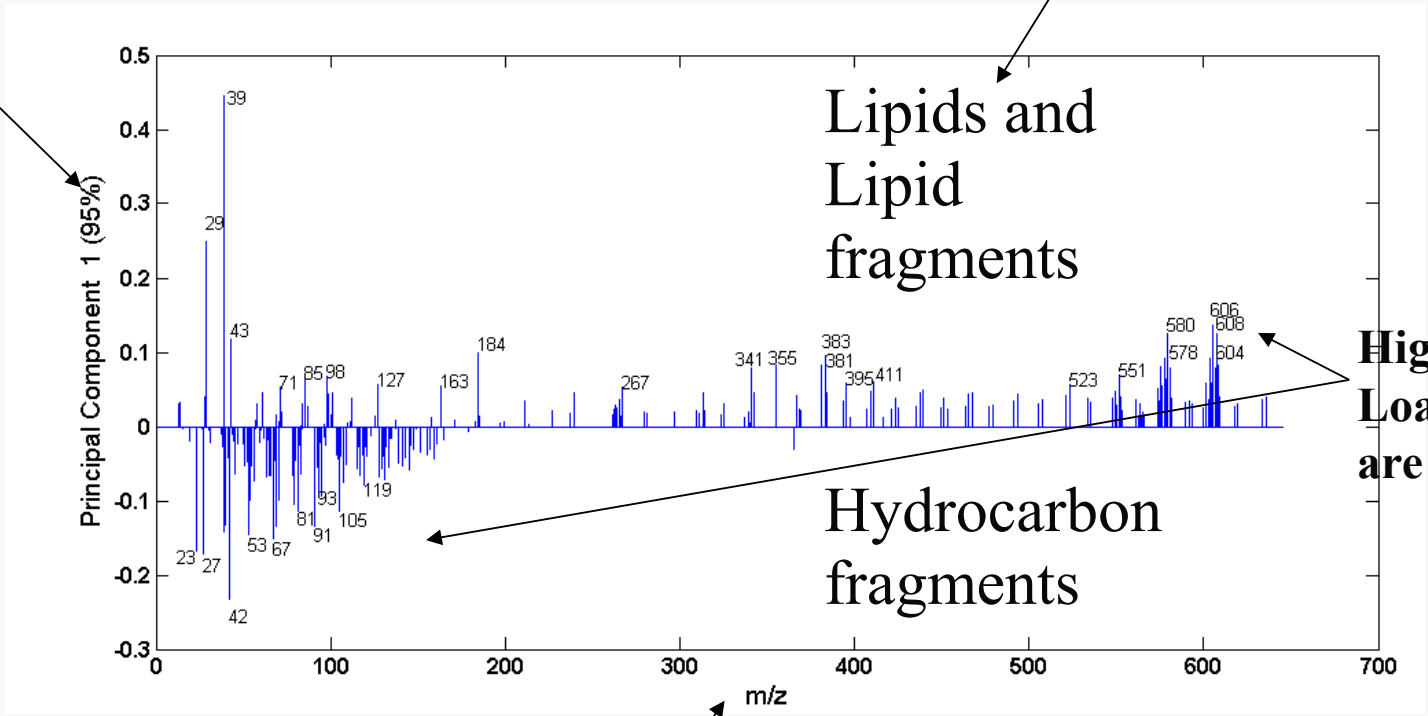
PCA Loadings

- 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



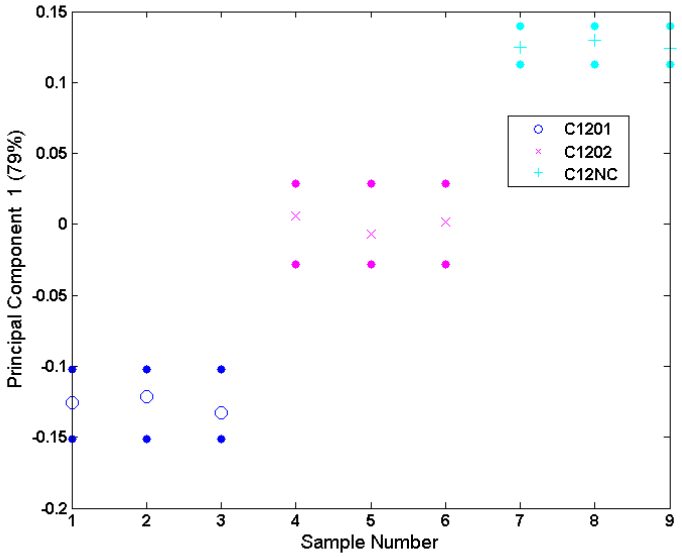
Highest Loadings are labelled

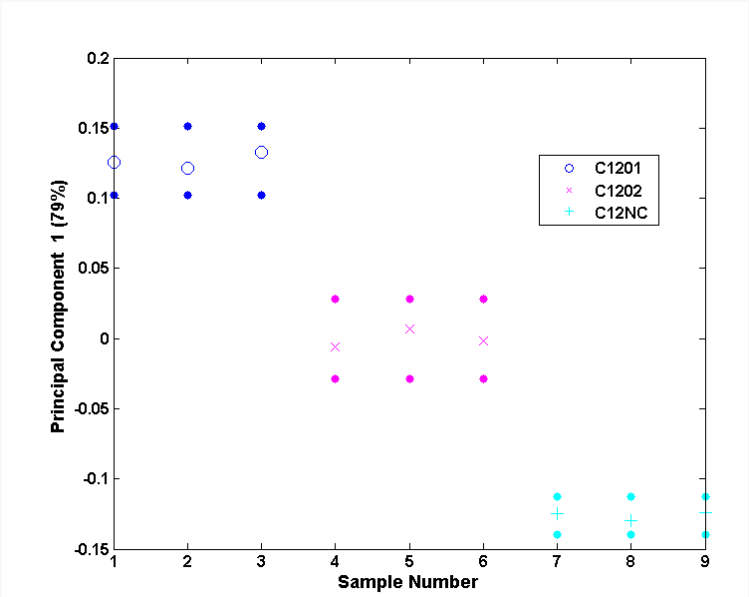
Loadings are plotted versus m/z

PCA: Interpretation

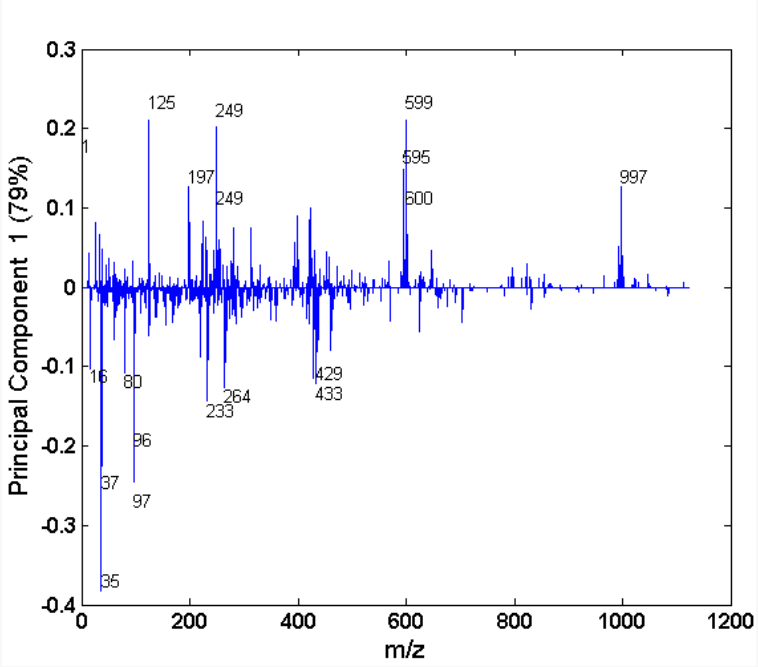
- A note on negative scores and loadings
  - Do not be afraid of negative numbers
  - +/- have no intrinsic meaning
  - Only important to keep + and – scores and loadings together

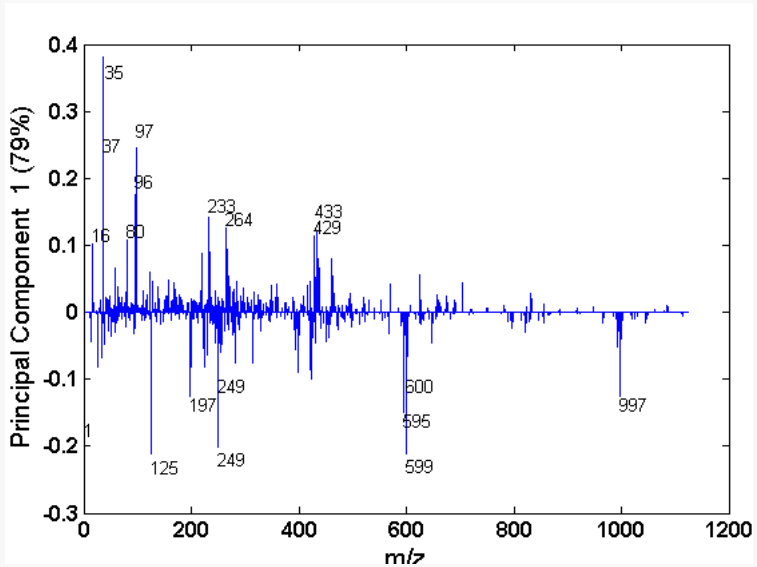
PCA: Interpretation





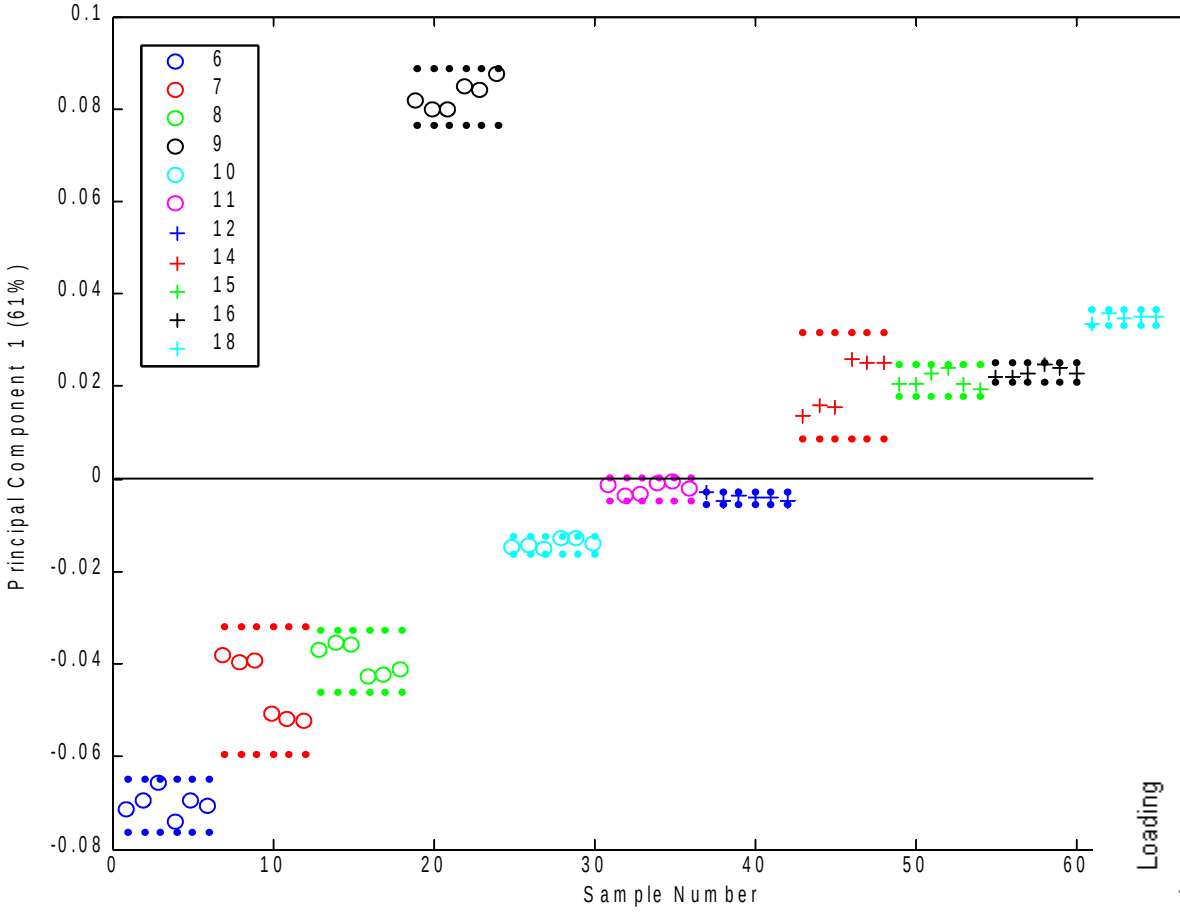
x (-1)

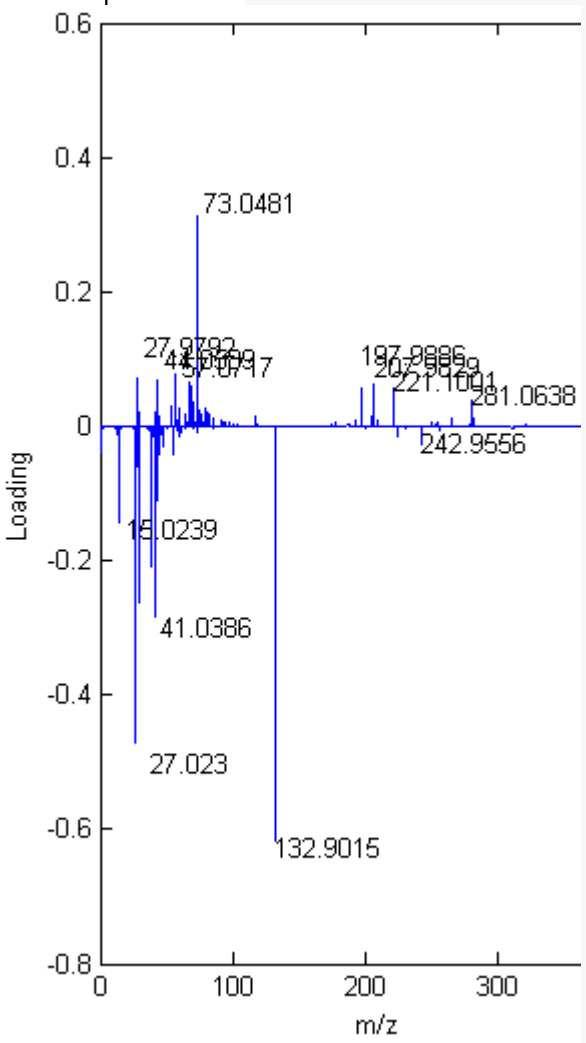




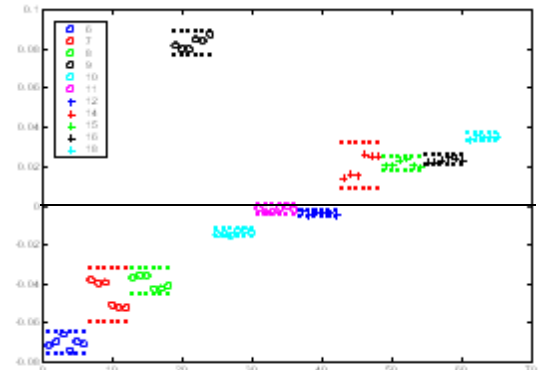
PCA: Interpretation

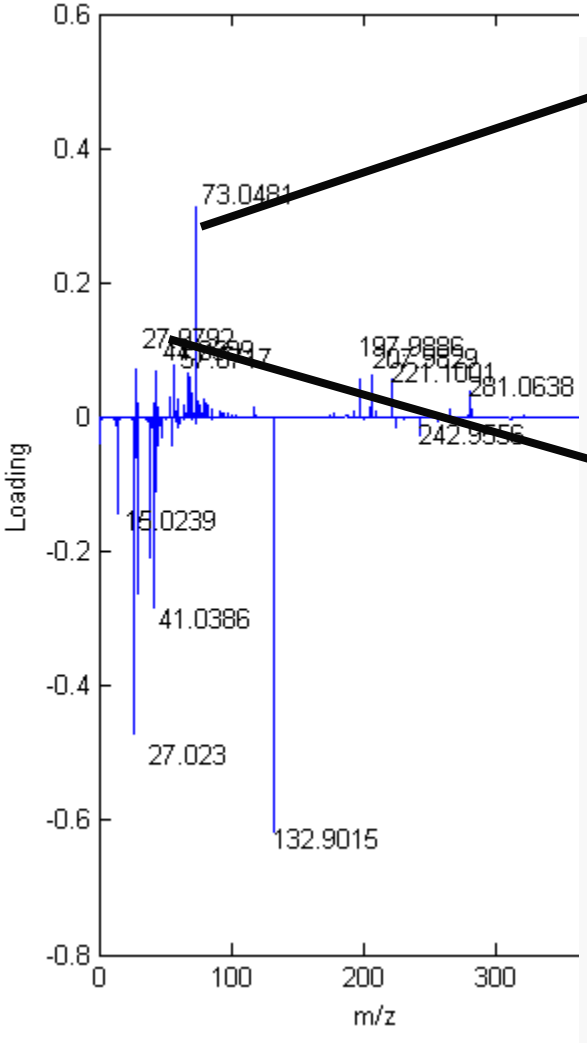
- Scores and Loadings are interpreted in Pairs
  - PC1 scores with PC1 loadings
  - PC2 scores with PC2 loadings
  - Etc…
- Samples with high positive scores on a given PC correspond with variables with high positive loadings
- This means that in general samples with high positive scores on a given PC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

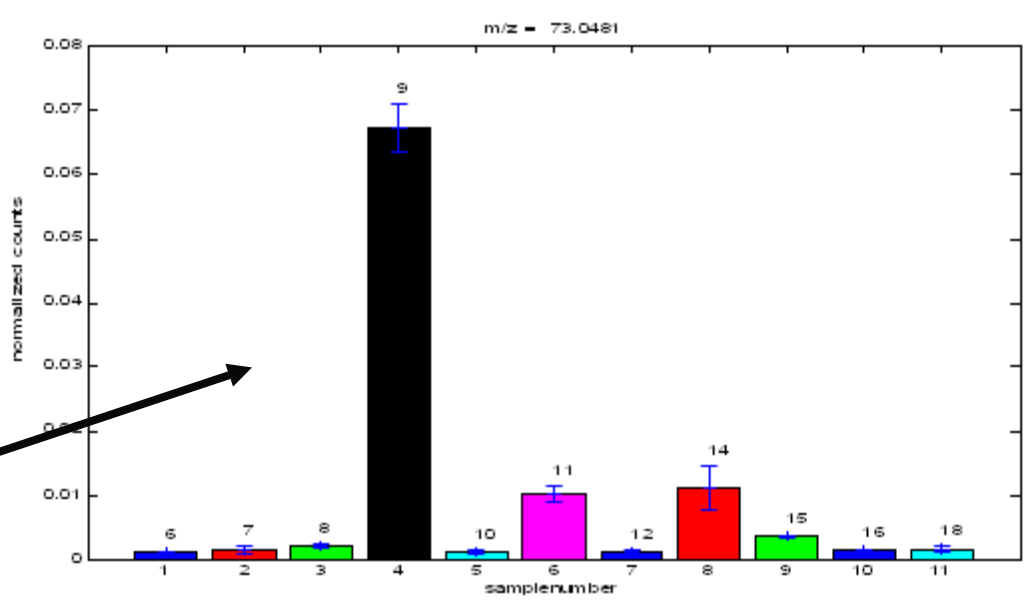


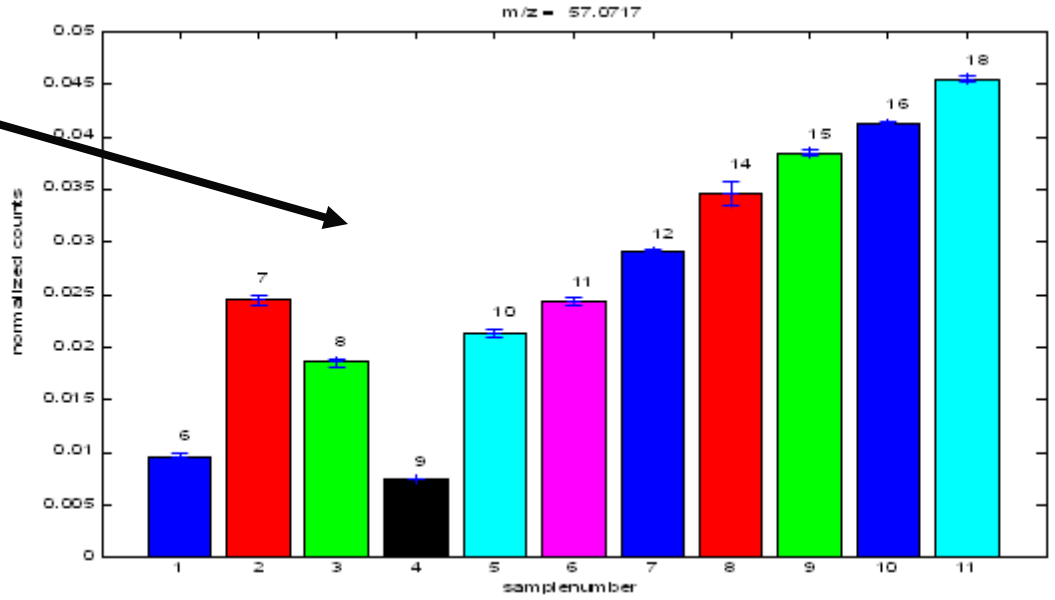


PCA of Methyl SAM series




Loading
m/z
73.0481
27.0232
41.0386
27.023
132.9015
281.0638

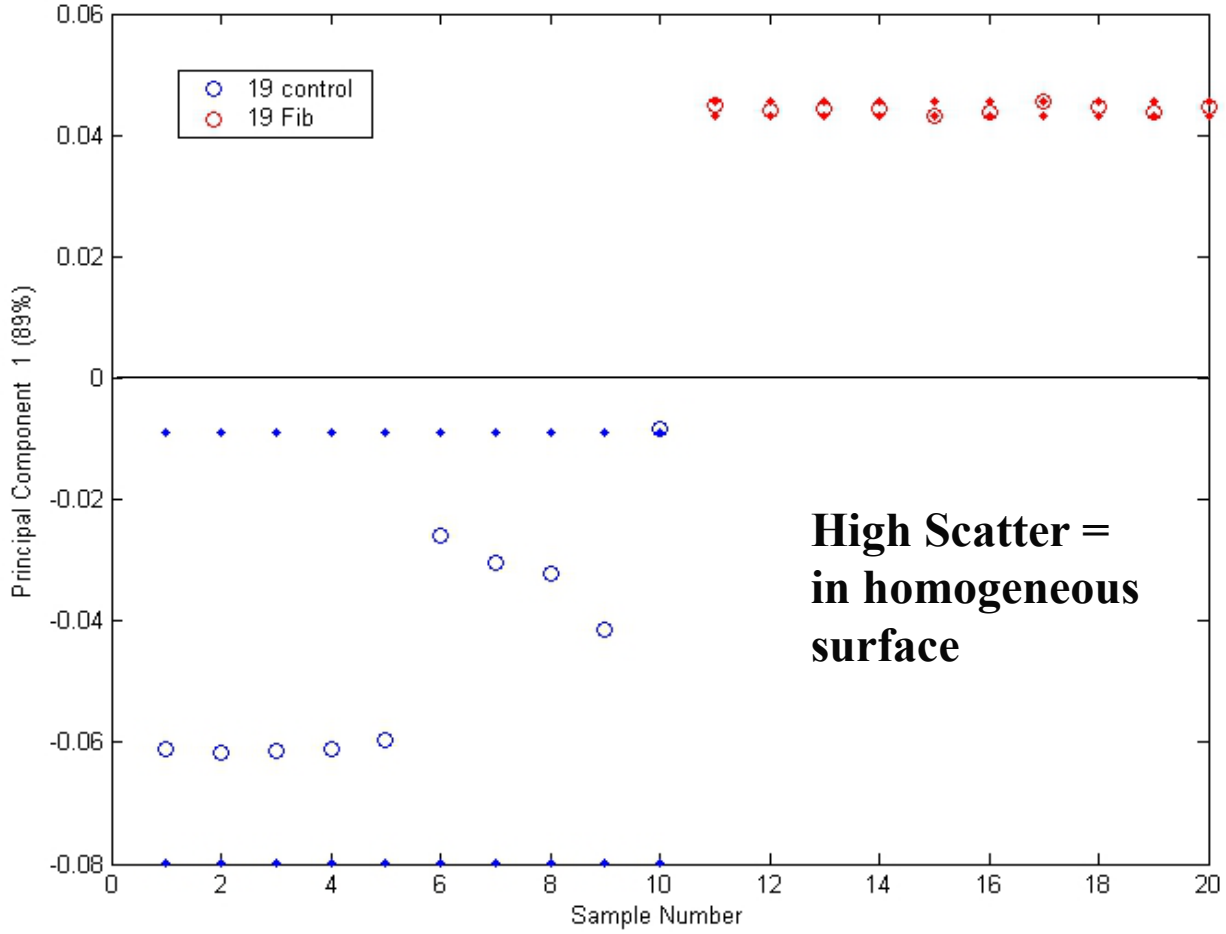

m/z = 73.0481
normalized counts
samplenumber


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

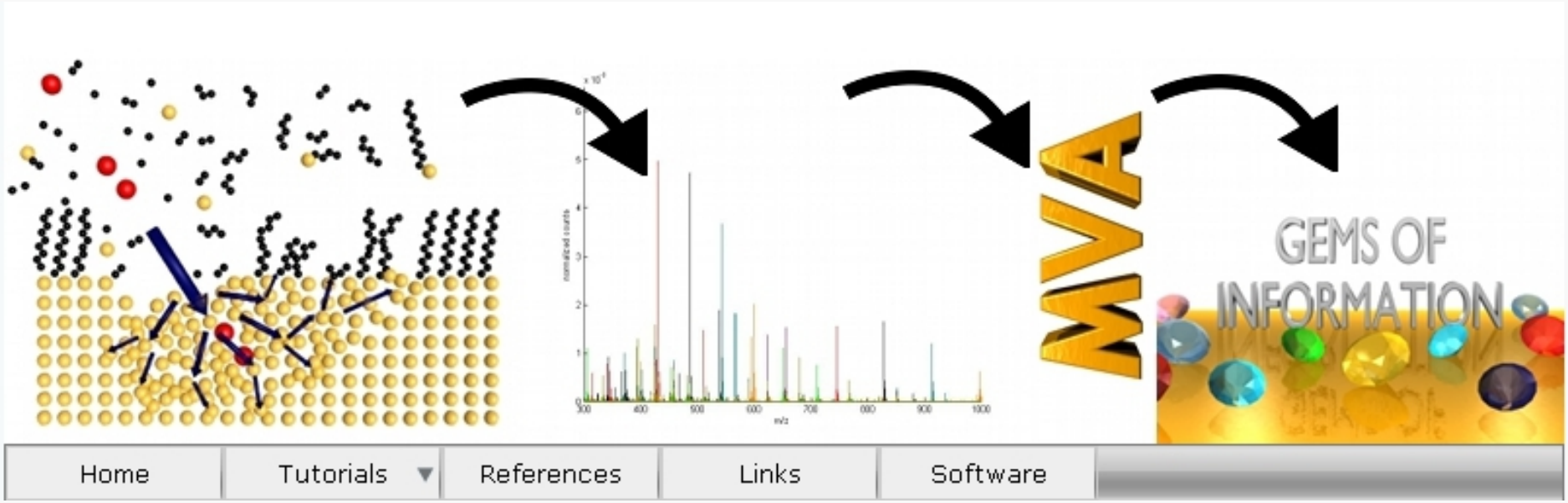
High Scatter = in homogeneous surface

Software Packages

- Commercial
  - SAS
  - SPSS
  - S Plus
  - PLS Toolbox (requires Matlab)
- Open Source or Free
  - NB Toolbox (requires Matlab)
    - http://mvsa.nb.uw.edu/
  - R Project
  - Octave
  - Scilab

Summary

- What are Multivariate statistics?
  - Powerful Tools for
    - Dealing with large data sets
    - Removing user bias
      - Finding patterns and trends
      - Building models
      - Prediction
      - classification
  - Hypothesis Generators
  - NOT a substitute for good scientific methodology
      - Well designed experiments
      - Replicates and controls
      - Validation of hypotheses



- **mvsa.nb.uw.edu**
  - Tutorials
  - References
  - Links
  - Software

djgraham@uw.edu

National Institute of Biomedical Imaging and Bioengineering

NESAC/BIO

The Complexities of The Basics of PCA (Part I)

Dan Graham, Ph.D.
NESAC/BIO
University of Washington

An NIH National Center for Research Resources
(NIH grant EB-002027)
http://www.nb.uw.edu

NIH National Institute of Biomedical Imaging and Bioengineering

NESAC/BIO

This tutorial will cover some of the basic concepts about PCA. Though it is focused on PCA, the guidelines provided herein can be applied to any multivariate analysis method.

Overview

- **Why Use Multivariate Methods?**
- **What are Multivariate Statistics?**
- **How to use MVA with ToF-SIMS data**
  - **Basic complexities of running PCA - step by step**
  - **Basic complexities of displaying and interpreting results**

In this tutorial I will briefly cover why MVA is required and what MVA is. The majority of the tutorial will cover how to use PCA for ToF-SIMS data starting from collecting data through interpreting the PCA results. As mentioned before, though this focuses on PCA, the concepts and methods covered apply to other MVA methods too.

WHY?

Why Use Multivariate Statistics?

- Reduce Size of Huge Data Sets
  - –Keep important information
- Remove User Bias
- Efficiently Use the Data
- Chemical Signatures are Multivariate
- Biological Problems are Multivariate
- Quantitative Analysis
- NOTE: MVA is a tool, not a replacement for knowing what you are doing

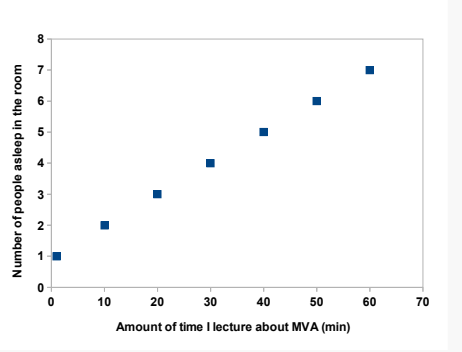
When you collect ToF-SIMS data from a set of samples it is of interest to know how the spectra change and in particular which peaks are different from sample to sample. Due to the large number of peaks in a typical set of ToF-SIMS spectra, it is often difficult to manually find all the changes throughout a data set. MVA provides a way to process large data sets, such as ToF-SIMS spectra, and determine what is changing between samples. Since MVA methods are designed to process large data sets, on can look at all the peaks within a ToF-SIMS data set and therefore remove user bias introduced by selecting only a few peaks. This more efficiently uses the data and enables finding trends in the data that otherwise may have been missed.

MVA is also useful because many problems we try to address with ToF-SIMS are multivariate in nature, meaning that multiple peaks can track changes on the surface and sometimes there may be multiple things changing within a sample set.

Though MVA is a powerful set of tools for data processing, it should be remembered that they are ONLY tools. This means that MVA will not analyze your data for you. It will only provide a way of determining where to focus your analysis. MVA is a set of tools that can help you, not do your work for you.

MVA Sounds Cool, Should I Always Use It?

- No. If you only have a few variables, it probably doesn't make sense to use MVA.
- If another method can answer your question, then use the other method.



Some data sets do not require MVA. For example, if you have previous knowledge about a data set that shows that a small set of peaks can be used to track the changes in a given system, MVA may not add anything to your analysis. Also, if you have only taken 1 or 2 spectra on each sample using MVA may not make sense. MVA methods are statistically based methods and therefore you need to take sufficient data to adequately describe the variance within your sample set.

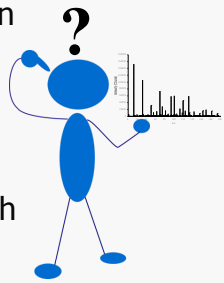
MVA Can Be Very Useful If:

- You have data sets with a large number of samples and variables
- You want to remove potential user bias
- You want to identify samples or predict a response based of experimental measurements

General slide on some of the benefits of MVA.

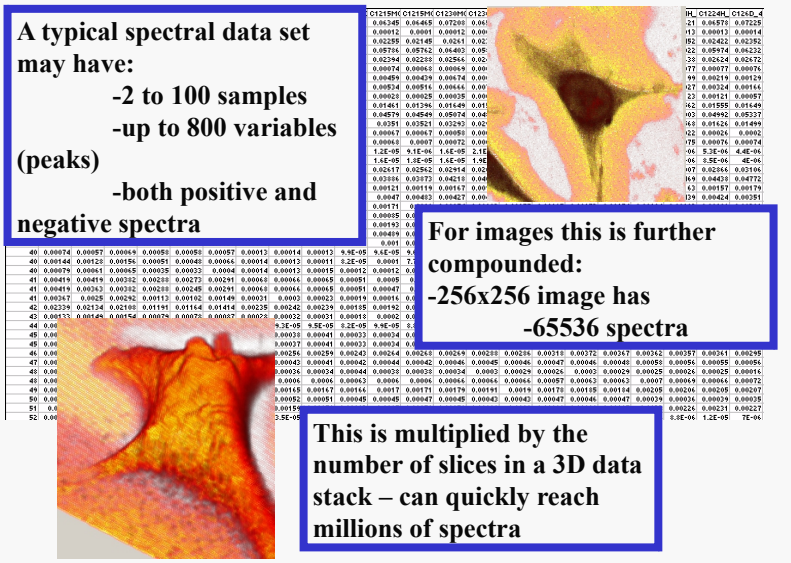
MVA Is Useful for ToF-SIMS Because SIMS Data Is Complicated

- Spectra contain hundreds of peaks
- Images contain thousands of spectra
- Peak intensities can be interrelated
- Matrix effects can cause non-linear changes in peak intensities
  - Due to sample composition
  - Due to the presence of oxides
  - Due to the presence of salts
- Peak intensities may or may not correlate with surface composition
- Peak intensities may vary due to differential sputtering
- Often heavy fragmentation and lack of molecular ion signals (assuming you know what signal to expect)



There is no way around it. ToF-SIMS data is complicated. There can be multiple factors that can change the relative intensities of peaks within a spectrum or image. These changes can be due to instrumentation, sample preparation, sample composition and more.

Data Overload



ToF-SIMS generates a lot of data. It is of interest to use all of this data in order to understand more about the system of interest.

WHAT?

"Multi" "Variate"

- Multivariate = more than 1 variable
- So multivariate analysis pertains to the analysis of multiple variables
  - Response from multiple measurements/instruments
    - XPS, ToF-SIMS, Contact Angle
  - Multiple responses from single measurement/instrument
    - Multiple ToF-SIMS peak intensities
- MVA – looks at the dependence (covariance) between different variables

Multivariate simply implies working with multiple variables. These variables can be anything that is measured. For ToF-SIMS the variables are typically the integrated peak areas or possibly the intensities of individual mass bins.

MVA looks at the covariance between all variables in the system. This means it looks at how each variable varies (or changes) compared to all the other variables.

Multivariate Analysis Methods

- Many different methods available
  - Principal component analysis (PCA)
  - Factor analysis (FA)
  - Discriminant analysis (DA)
  - Multivariate curve resolution (MCR)
  - Maximum Autocorrelation Factors (MAF)
  - Partial Least Squares (PLS)
  - PCA-DA, PLS-DA, CA
- We will focus on PCA
  - Most commonly used method
  - Successful with SIMS data
  - Forms the basis for many other methods

There is a whole alphabet soup of MVA methods. Many of these are variations of factor analysis. Each of the available methods can be useful and were designed to find new ways to process and interpret large data sets. No one method is better than another. Each has strengths and weaknesses. You should choose a method based on which one you think will answer the questions you are working on.

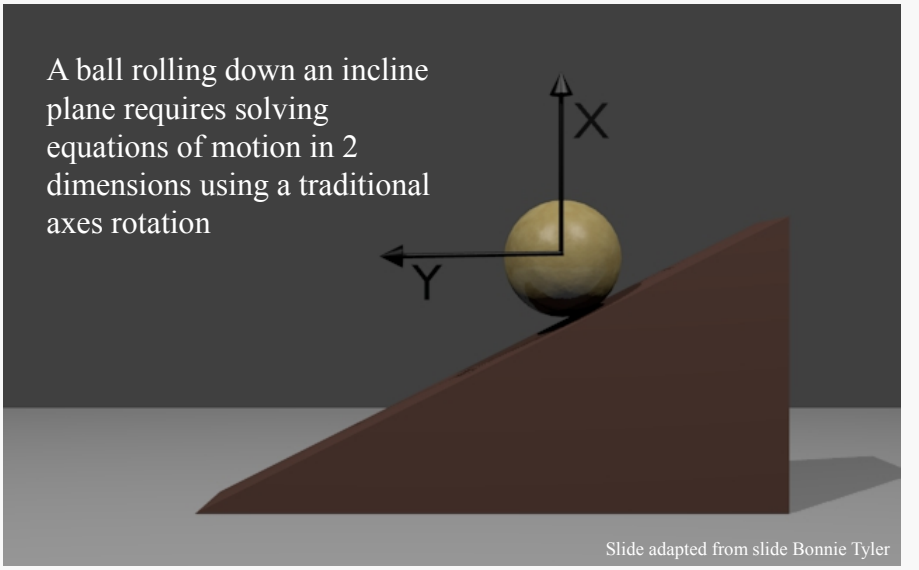
When Should MVA be Used?

- MVA 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

MVA works particularly well when you limit the number of variables that are changing in a given system. Here I refer to external variables such as concentration, temperature, or exposure time. Variables that you can control during your experiment. If you limit the number of variables you change within a given system to 1, then you can interpret your MVA results in reference to that one variable. As always, it is best to have a central question you are trying to answer, and to use a well controlled set of samples. Your question should be part of your experimental design.

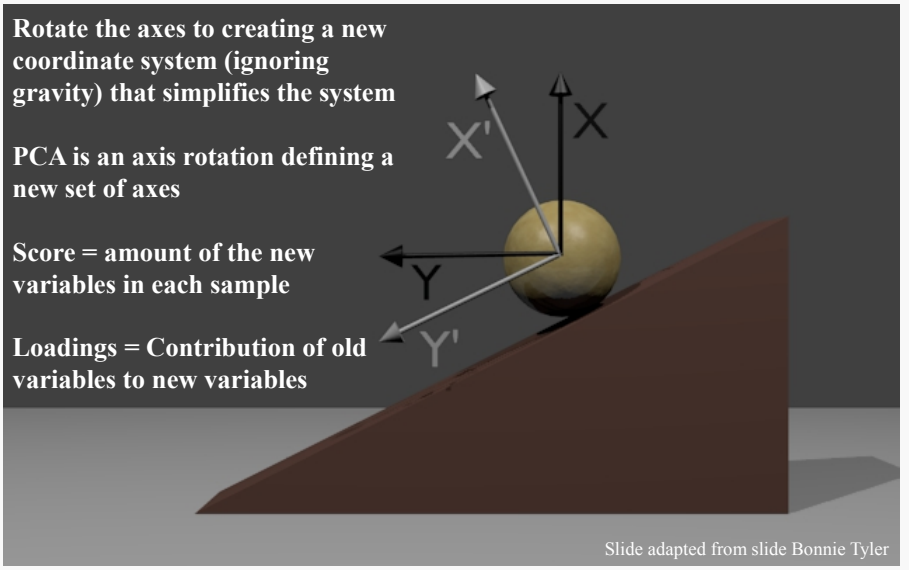
MVA is also good for determining differences between sets of samples. However, I would caution you on using MVA as a last resort to try and figure out something from a confusing data set. Without a good experimental design, MVA may only be able to confuse you further. That doesn't mean that it cannot be helpful with unknown samples, just that you need to think about what you are doing.

Multivariate Analysis
Reducing the Dimensionality of a Problem



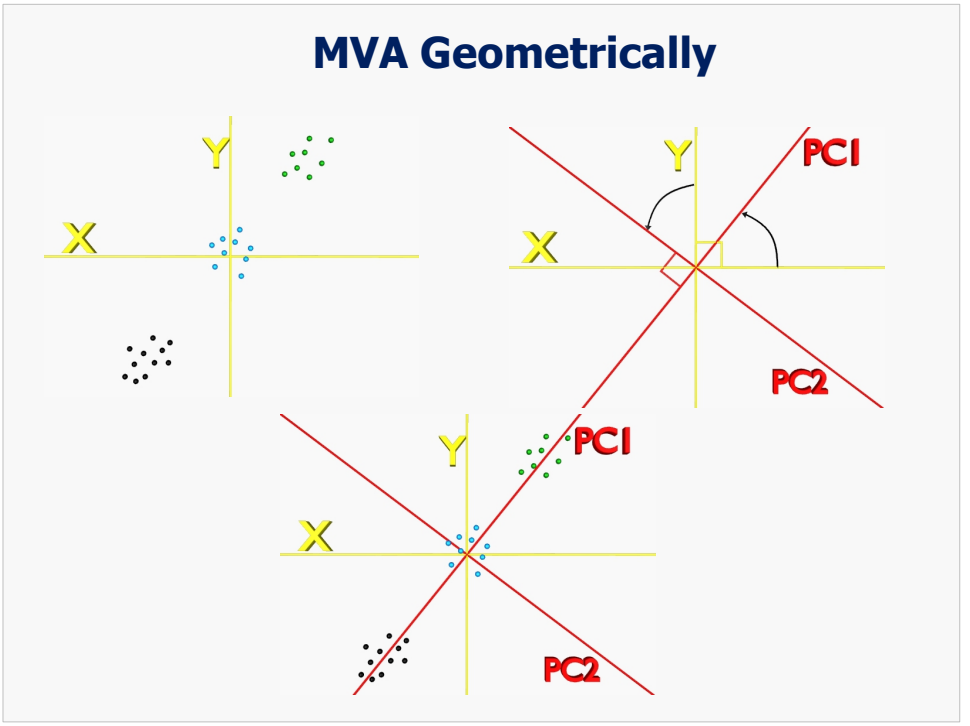
Many MVA methods involve doing an axis rotation. Axis rotations can be useful to simplify the number of variables within a system. For example, for a ball rolling down an inclined plane using a standard set of axes (pictured above), you must solve the equations of motion in 2 dimensions (ignoring gravity).

Multivariate Analysis
Reducing the Dimensionality of a Problem



If we rotate the axes so that the Y' axis is parallel to the surface of the inclined plane, we can reduce the problem to be 1 dimensional.

PCA is a type of axis rotation. The axes of a system are rotated to define a new set of axes that capture the major differences between the samples. This axis rotation creates new sets of data called the scores and loadings. The scores tell the amount of the new variables in each sample. The loadings tell the contributions of the old variables to the new variables. We'll talk more about this later.

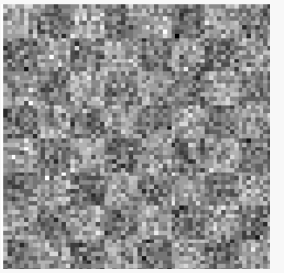


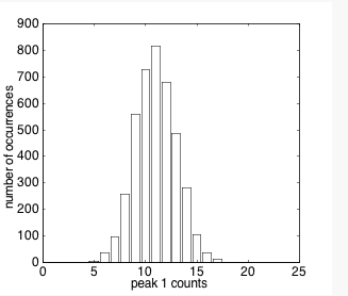
This slide illustrates how the first 2 PC axes are defined. The first PC axis is rotated so that it captures the largest direction of variance within the data. In the case of this simulated data set, the largest variation within the data is the spread of the samples along the diagonal through the 3 clouds of data points. PC1 is placed so that it is aligned with these sample differences. PC2 is placed orthogonal to PC1 in the next greatest direction of variation. In this case PC2 is capturing the spread in the samples on either side of the PC1 axis.

How Axis Rotation Can Help

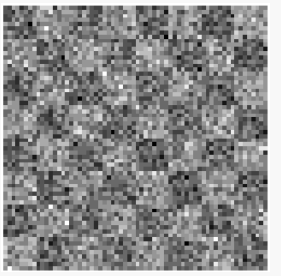
Computer Generated Image

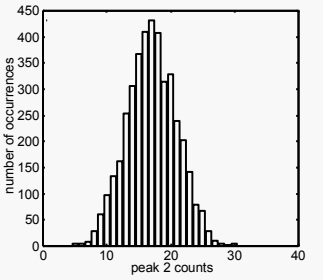
Peak 1 Image





Peak 2 Image



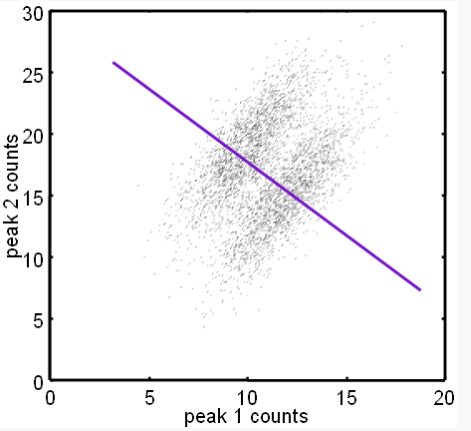


Slide courtesy of Bonnie Tyler

This is a nice example provided by Bonnie Tyler of how axis rotation can help. In this slide two peak images are shown along with the intensity histogram of the peak 1 and peak 2 counts. One can just barely make out the square pattern in the peak area images.

How Axis Rotation Can Help

Computer Generated Image



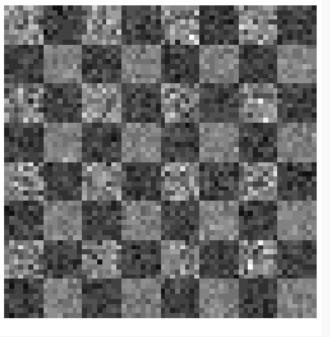
Slide courtesy of Bonnie Tyler

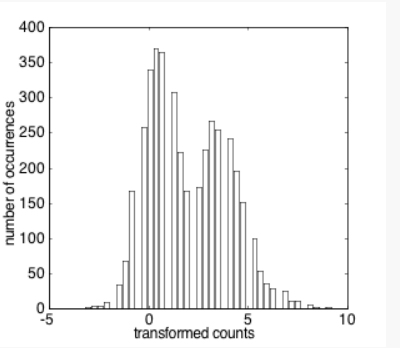
On this slide the peak 1 counts are plotted against the peak 2 counts. On this plot it is apparent that there are two clouds of data. If one created a new axis that cuts through the short axis of these clouds and then projected the data onto this new axis line one could see that the two peaks could be better separated.

How Axis Rotation Can Help

Computer Generated Image

Transformed Variable Image and Histogram





Slide courtesy of Bonnie Tyler

This slide shows the data projected onto the new axis (shown in the last slide). The square pattern can now be easily seen in the peak area image. Also, it can be seen that the histograms for both peaks are preserved on the transformed variable space. So all pertinent information about the data set can be captured by projecting the data onto this new axis system.

The New Coordinate System

How do we decide how to rotate the axes?

Problem	Method
Sources of Variation in a Data Set	PCA, MCR, MAF
Prediction of Properties from Data Set	PLS, PCR
Differences Between Analysis,HCA, Groups or Classes	Discriminant PLS-DA
Find Features in Noisy Images	MAF

- Choose the method based off the goals of the analysis
  - No one method is "Better" than another
  - They are all just tools

How you rotate the axes depends on the goals of the analysis. This slide lists a few examples showing the type of problem to be solved, and the possible MVA methods that could be used.

The MVA method used should be based off the goals of the analysis. Remember, no one method is "better" than another. They are all just tools to help you understand your data. Also, you should understand how the methods work and how to interpret the results you get before you start using them routinely.

Math

- MVA methods are based of linear algebra and matrix math
- Good reviews of PCA math can be found in the literature
  - Chemom. Intell Lab Sys, 1987, 2, 37-52
  - J Qual Technol, 1980, 12, 201-213
  - Jackson JE, 1991, User's guide to principal components, John Wiley & Sons, NY

Very nice overviews of PCA mathematics can be found in the articles listed above. It is highly recommended that you read through these (and other) articles to get a good understanding of how the math works and what it is doing.

Principal Components Analysis

- **Variance**
- **A measure of the spread in the data**

$$S^2 = \frac{\sum x - \bar{x}}{n - 1}$$

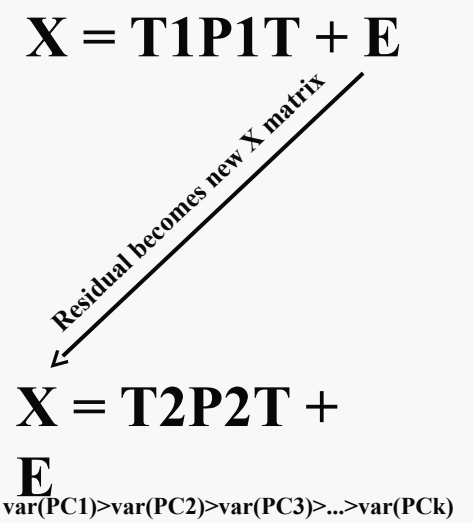
- **Covariance**
- **A measure of the degree that two variables vary together**
- **PCA is calculated from the covariance matrix**

$$\mathrm{cov}(X) = \frac{X^T X}{m - 1}$$

Here I will just cover some basic concepts about PCA math. First we need to understand what variance is. Variance is a measure of the spread in the data. You can calculate the variance for a given variable using the equation at the top of the slide. For MVA we want to know how each variable varies with respect to all of the other variables. For this we need to calculate the covariance matrix. This matrix contains the variance of all variables with respect to all other variables. PCA is calculated from the covariance matrix.

PCA Methodology

- **PCA determines sequential orthogonal axes that capture the greatest direction of variance within the data**



PCA determines sequential orthogonal axes that capture the greatest directions of variance within the data set. What this means is that PCA first determines the greatest direction of variance within the data set and defines the PC1 axis along this direction. PC1 is then subtracted from the data set and the residual matrix (E) becomes the new X axis that is used to find PC2. So each subsequent PC is calculated from the residual found by subtracting all previous PCs from the input data matrix. At some point the residual matrix E contains only noise.

The variance captured in PCA always follows the rule that var(PC1)>var(PC2)>var(PC3)>…>var(PCk)

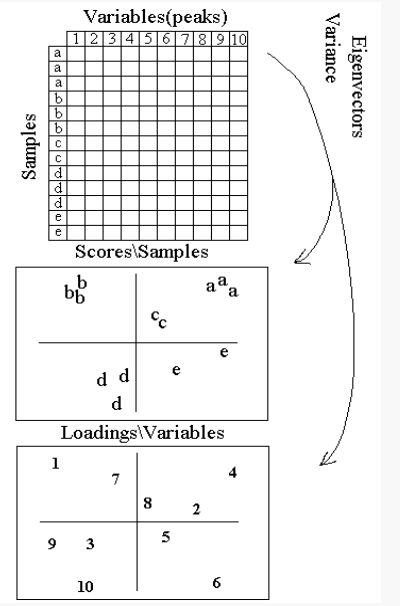
What the Math Means

- Scores ---> "concentration"
- Loadings ---> "spectra"
- However these are not usually "concentrations" or "spectra" of a pure component or even an individual physical factor

Scores can be thought of as an "amount" or "concentration" and loadings can be thought of as "spectra", but not in the sense of pure component or a physical factor.

PCA

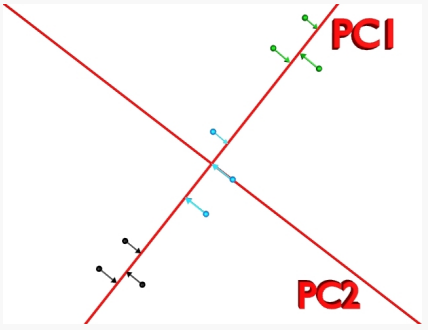
- **Looks at the variance patterns of a data matrix**
- **Reduces data dimensionality**
- **Gives simple graphical presentation of data**
- **Determines relationship of samples and variables based on the variance in the data**
- **No external constraints needed**
- **Original matrix is reconstructed into new matrices that define the major patterns of the data in multivariate space**
  - **SCORES -> Describe relationship between samples (spread) as described by PC's**
  - **LOADINGS -> Describe how the variables relate to the PC's**



This slides gives an overview of PCA. The figure on the right illustrates how the original data matrix is reconstructed into two new matrices (the scores and loadings). The scores show the relationship between the samples and the loadings show which variables are responsible for the differences seen between the samples. For example in this simulated data set samples with different letters are different from each other since they cluster together and are separated from samples with different letters. It can be seen by looking at the scores and loadings that samples b correspond more with peaks 1 and 7 (seen in the same quadrant as samples b) and samples a and c correspond more with peaks 4, 2 and 8 (located in the same quadrant as samples a and c).

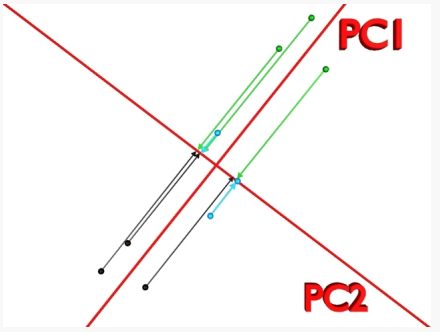
Scores

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



Projection onto PC1

Scores tell the relationship (spread) between the samples

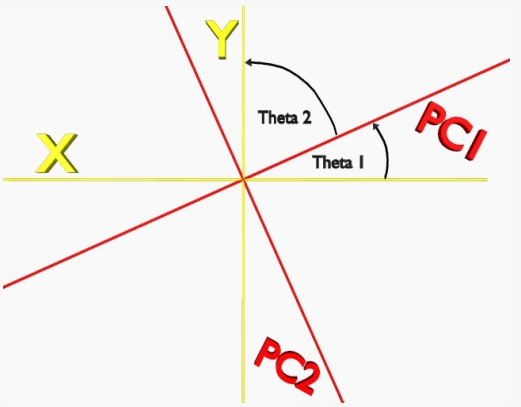


Projection onto PC2

The scores are a projection of the samples onto the new PC axes. So the PC1 scores are found by projecting the samples onto the PC1 axis. The PC2 scores are found by projecting the samples onto the PC2 axis. As can be seen, for this example data set, by projecting the samples onto PC1 the samples will be separated from each other (they have different scores). On PC2 the samples will overlap with each other.

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**

The loadings are the direction cosines between the new axes and the original variables. Cosine(0)=1 and cosine(90)=0. This means that variables that are more closely aligned with the new PC axes will have a higher loading (smaller angle = larger cosine = larger loading). Variables with larger loadings have a larger influence on the separation of the samples on the given PC axis. Conversely, variables (peaks) with small loadings have a smaller influence on the sample separation. In the case of a small loading, the angle between these variables and the new PC axis would be large (large angle = small cosine = small loading).

HOW?

The next section will cover how to collect data, pre-process it and run PCA.

Background Information

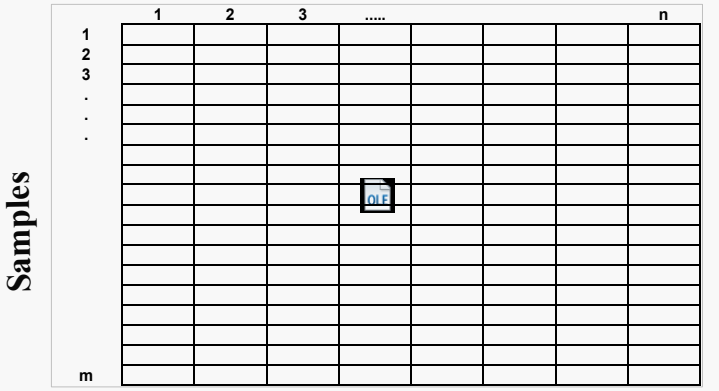
- Data is arranged in matrices
  - samples in rows
  - variables in columns
- m = number of samples
- n = numbers of variables
- k = number of PCs
- T = scores matrix
- P = loadings matrix

For PCA the data must be arranged in a matrix. Samples (spectra) should be in the rows, and the variables (peak areas) should be in columns.

This slide provides definitions that are used later.

Data Matrix

Variables



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**

This slide shows a blank matrix showing that samples should be in rows and variables in columns. For SIMS variables are peak areas (or counts from data bins), and samples are spectra.

PCA: Things to know

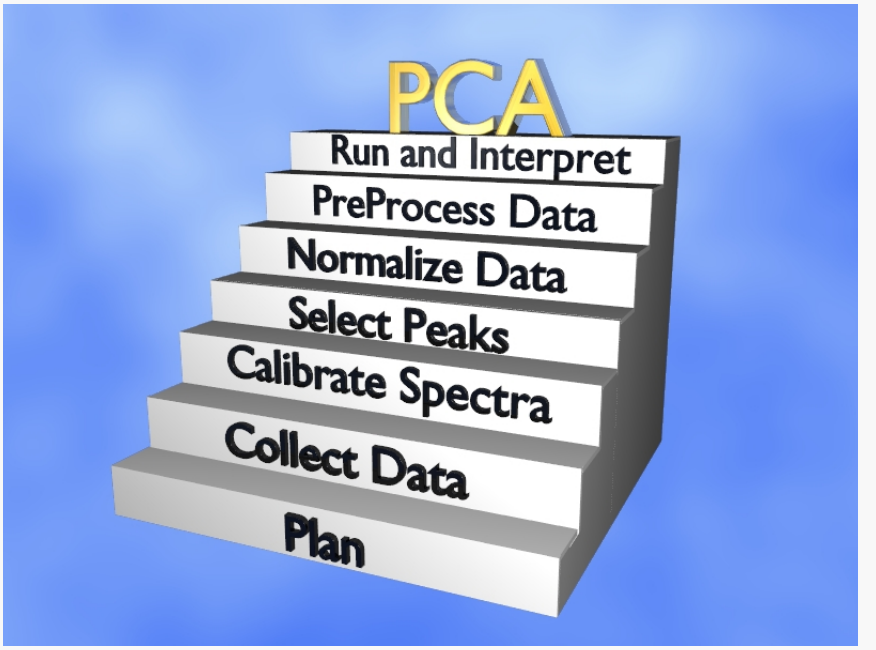
- PCA assumes linear relationships between variables
- PCA is scale dependent
  - variables with larger values look more important
- PCA looks at variance in the data
  - It will highlight whatever the largest difference are
  - To make sure you are comparing things properly it is common to preprocess the data
    - Remove any instrument variation, or other non-related variance (normalization)
    - Make sure data is compared across a common mean (centering)
    - Make sure data is compared across common variance scale (autoscaling, variance scaling, etc)

These are important assumption made when running PCA.
PCA assumes a linear relationship between variables (peaks). This may not always be true with ToF-SIMS due to matrix effects.

PCA is scale dependent. This means the answer you get will be dependent on any pre-processing you do with the data. Variables with larger values will look more important in the loadings plots simply due to the magnitude of the variable. Data pre-processing can be done to weight peaks in a way to give more influence to different peaks. This also means it is important to understand how and why you are pre-processing your data.

PCA looks at variance patterns. This means it will find differences no matter what they are. What I mean by this is that if you have contamination on some of your samples, it is likely that the largest PCs will find differences due to the contamination before it finds any differences due to your chemistry of interest.

Steps to PCA



This is a slide to illustrate the “steps” required to use PCA or any other MVA method.

Proper ToF-SIMS Analysis Requires:

- Good experimental plans (controls)
- Proper sample preparation
- Careful data collection
- Consistent data calibration
- Sound understanding of the fundamentals of mass spectral analysis
- Knowledge of how to properly use the available tools to help with the analysis

One should always plan carefully when doing any experiment. This is particularly important when using ToF-SIMS. Controlling extraneous variables and sources of potential variation within a data can be critical to the success of an experiment.

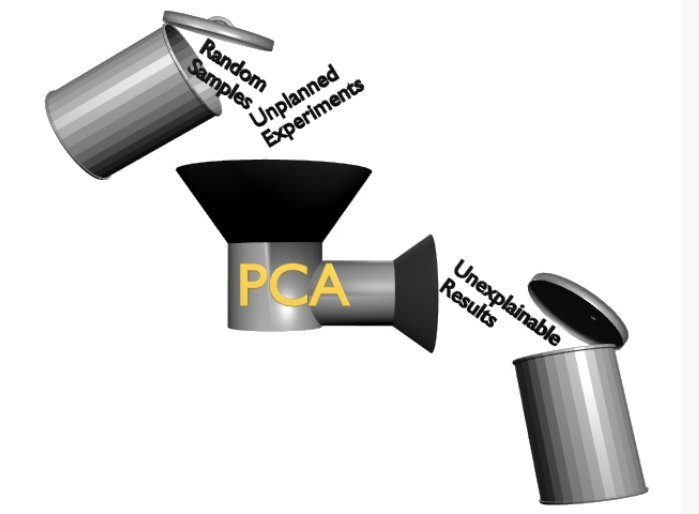
Remember MVA methods are just tools, you still need to understand the basics of mass spectral analysis.

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

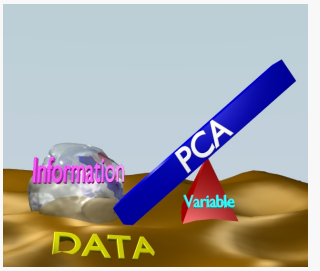


You will get garbage out!!!

If you put garbage in…you get garbage out. This slide is to remind you that you need to plan your experiments carefully.

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
    - Non-homogeneous => 5 to 7 spots on 3 to 5 samples



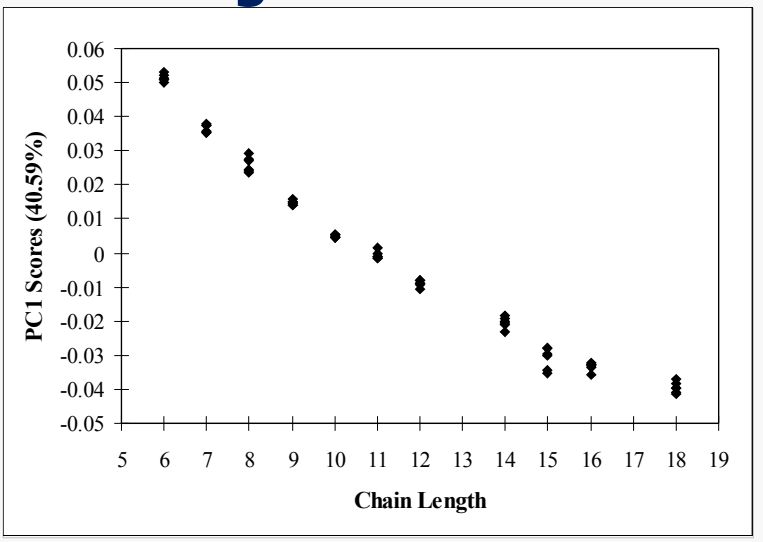
Planning your experiments properly and limiting the number of variables that are changing can allow you to pull out information from your PCA results that can help you interpret and understand your data set. For this you will need to run appropriate controls and collect enough data to adequately model the variance in your system.

This is a very important point. MVA methods are statistically based methods. You must acquire enough data to make sure your results are statistically significant. For homogenous samples (very reproducible spectra) 3 to 5 spectra across 2 samples (6 to 10 spectra total per sample type) is usually sufficient. For non-homogenous samples (where the spectra vary a lot), you need to take more data. Usually 5 to 7 spots across 3 to 5 samples is sufficient (15 to 35 spectra total per sample type). It is a good idea to repeat your experiments at least twice on separate days.

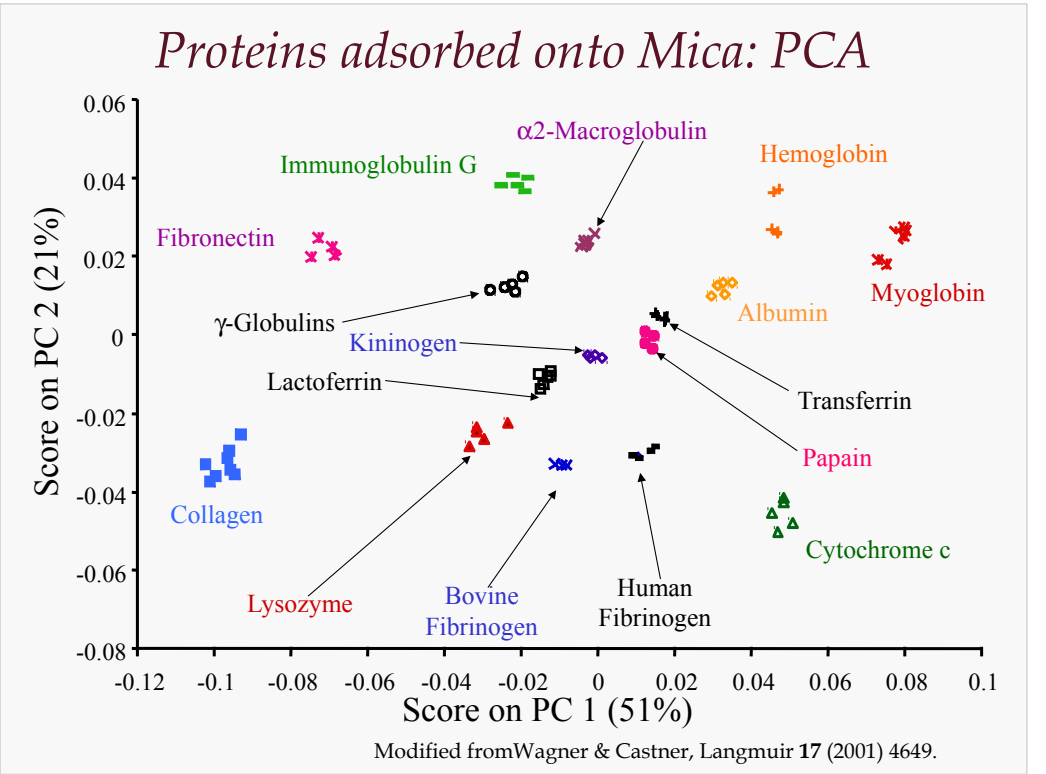
It is also important to always use 95% confidence limits when plotting PCA scores. This can help determine the significance of the separation seen between samples.

SAMs – typically very homogeneous

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

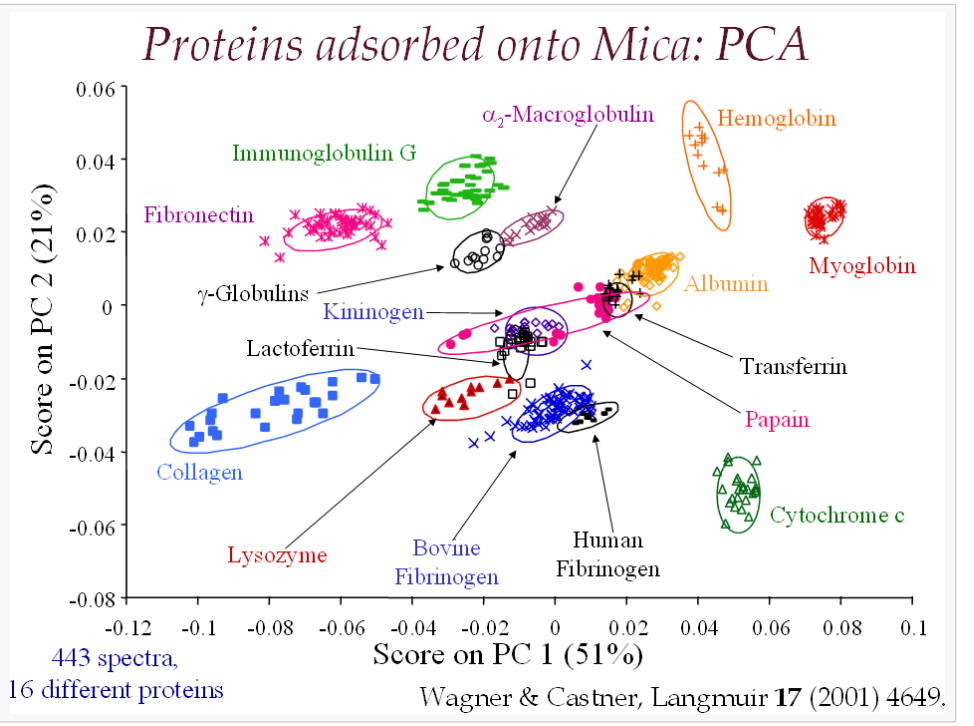


This is an example of very reproducible spectra. The data shown is from a series of alkane thiol monolayers on gold with chain lengths varying from 6 to 18. 6 spectra are shown for each sample. As can be seen, for many of the samples the spectra almost completely overlap each other. So 6 spectra per sample is sufficient for this type of sample.



This slide was adapted from the work of Matt Wagner and Dave Castner. In this slide I have deleted most of the data points from the protein data set generated by Matt. Using this set of data points one could look at the figure and conclude that all of the proteins are clearly separated and that the scatter in the data is minimal.

However if you at all the data points, the story changes....



This is the full data set from the previous slide. It can be seen that most of the proteins are separated from each other. However, there is significant scatter in the data from many of the proteins. For most biological samples (proteins, cells, tissues) you need to collect a lot of data in order to truly characterize the variance in the systems and extract useful, valid information. Just a few data points is not sufficient.

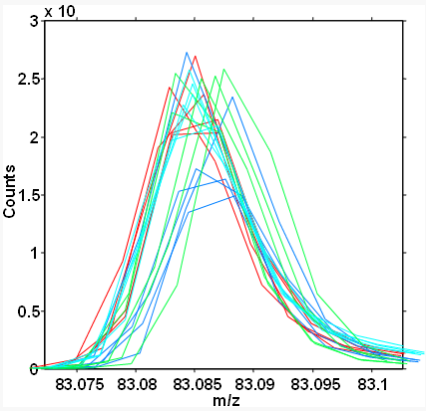
Data calibration

- 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
- Always double check autocalibration functions
  - They can make mistakes

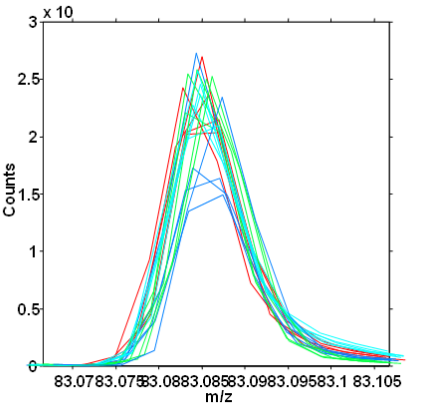
Once the data has been collected, it must first be calibrated before applying PCA. Calibration is included as a step to successful PCA because it is important that all the spectra within a sample set are calibrated properly and in the same way. To aide in the accuracy of high-mass peak identification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

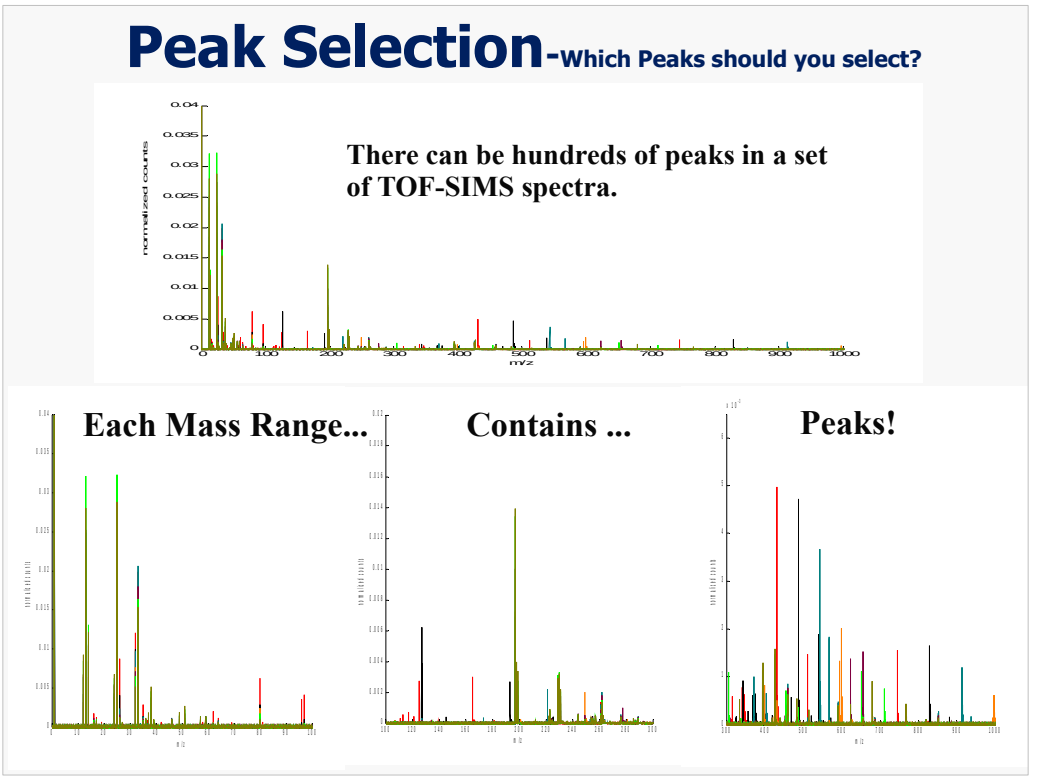


After Checking Calibration to assure consistency



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.

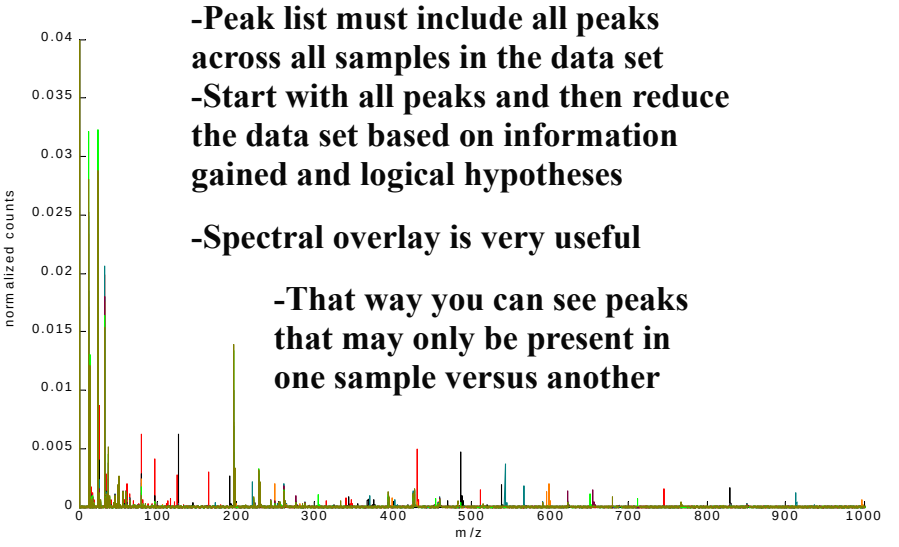
If this were not corrected, errors could be made in placing the integration limits for the peaks in the data set, and variance could be introduced into the peak areas that is not due to real sample differences.



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?



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.

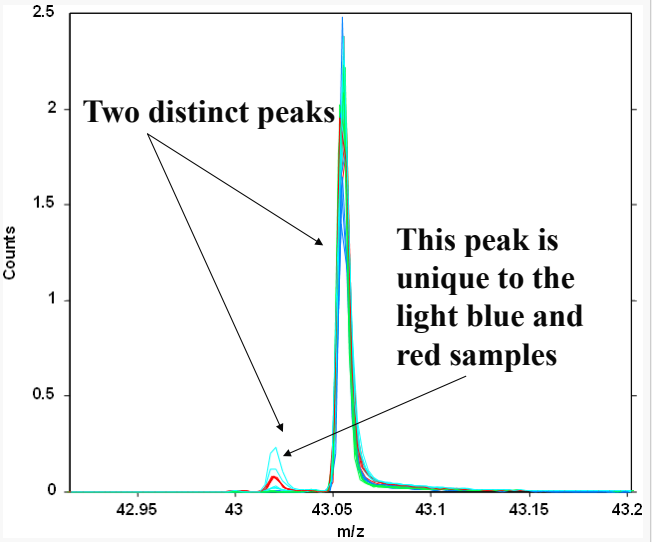
Since the same peak set must be used for all spectra that are to be used in PCA, it is useful to do peak selection from overlaid spectra. This allows the user to see peaks from all spectra on the same axis and helps avoid missing peaks that only show up in the spectra from 1 sample type within the set.

Peak Selection Continued

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

-Manual peak selection is recommended

-It is time consuming, but not prohibitive

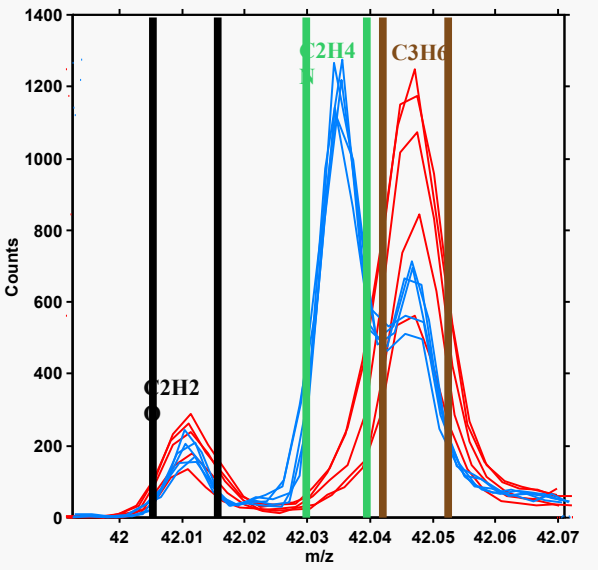


Though some programs contain routines to automatically select peaks from a spectrum, it is recommended to do peak selected manually. This will make sure that all the necessary peaks are properly chosen. Also most automatic selection routines are not able to set proper integration limits for the peaks. 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.

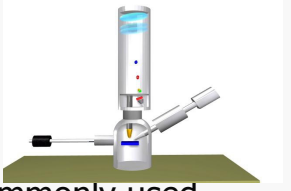
PCA data Pretreatment

- No standards have been set for data pretreatment
- Some common trends include
  - normalizing the data (many different ways)
  - Square root transform or divide by the square root of the mean and then mean center for TOF-SIMS spectra
  - Poisson scaling or square root transformation for TOF-SIMS images

This slide summarizes some general trends for data pre-processing, however no real standards have been developed.

Normalization

- Data normalization helps account for differences in the data due
  - *topography*
  - sample charging
  - instrumental conditions

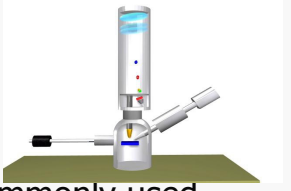


- Many different methods are commonly used
  - Total intensity
  - Sum of selected peaks
  - Highest peak in spectrum
  - User selected peak
  - Total intensity minus H and contaminants
- Know assumptions being made
- Understand that normalization removes information from the data set

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.

Normalization

- Data normalization helps account for differences in the data due
  - *topography*
  - sample charging
  - instrumental conditions

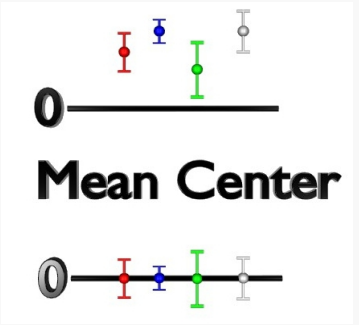


- Many different methods are commonly used
  - Total intensity
  - Sum of selected peaks
  - Highest peak in spectrum
  - User selected peak
  - Total intensity minus H and contaminants
- Know assumptions being made
- Understand that normalization removes information from the data set

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 TOFSIMS 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.

Mean centering



- Mean centering
  - Subtracts the mean of each column (variable) from each column element
  - Centers data so that all variables vary across a common mean of zero

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.

Scaling

- Scaling attempts to account for differences in variance scales between variables
  - Poisson scaling (takes into account Poisson noise structure that is often seen for SIMS data)
  - Dividing by the square root of the mean
  - Square root transform
  - Many others
- Need to know why you are using a given method
- Need to understand assumptions
  - Are the assumptions valid?

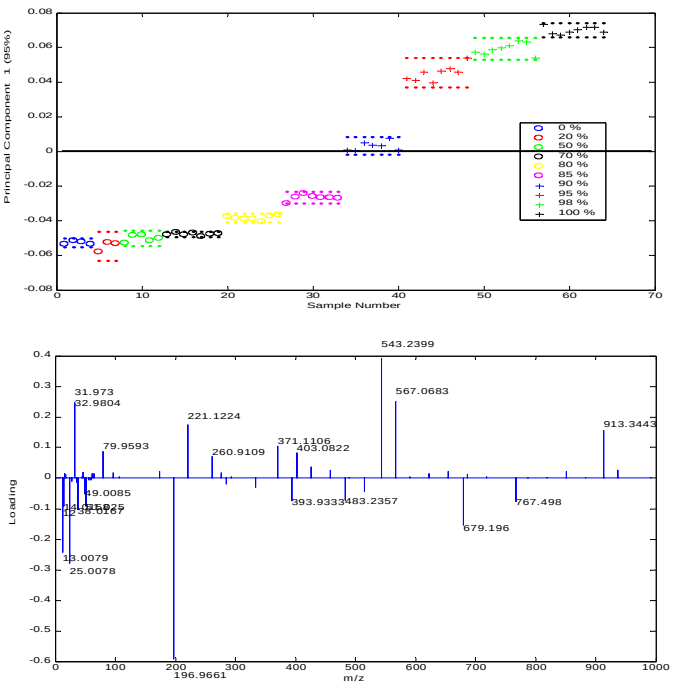
Data scaling is done to account for differences in the variance scales between variables Normalization can be considered a scaling operation since with normalization we are dividing or multiplying by some value to adjust for unwanted variances in the data. One common scaling method used with PCA is autoscaling. 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

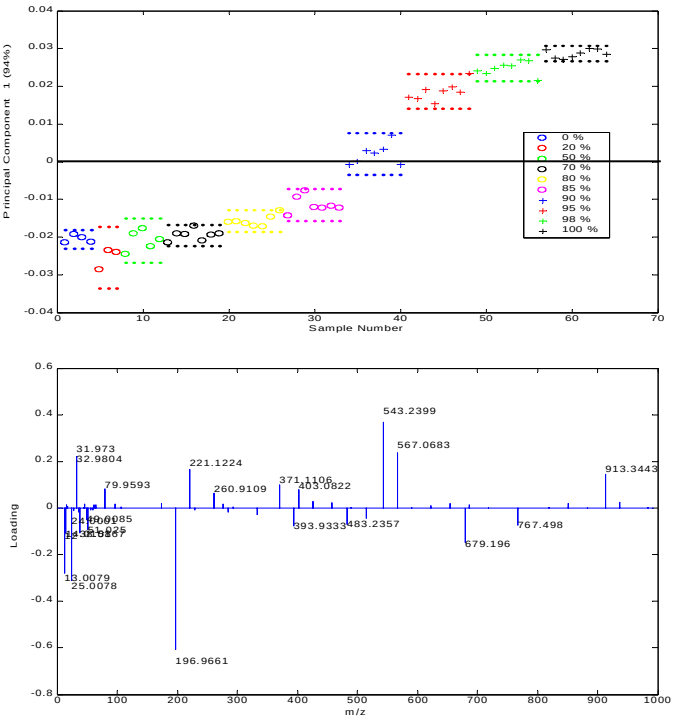


The following few slides illustrate the affects that data preprocessing can have on PCA processing of TOF-SIMS data.
The data presented here comes from a set of spectra taken from mixed monolayers of decanethiol (C10) and octadecanethiol (C18). The monolayers were assembled from different solution percentages 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

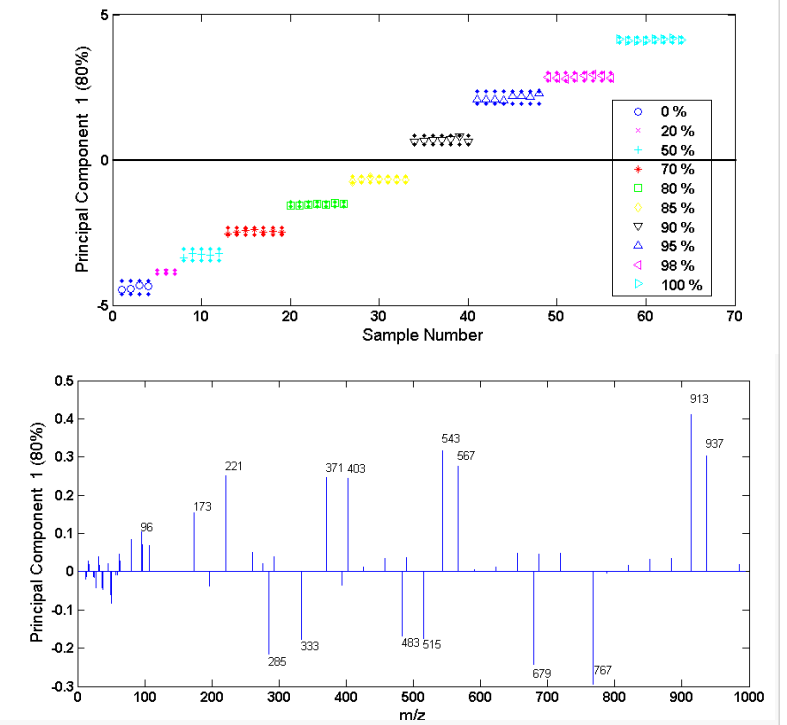
-Normalized Total intensity
-Mean centered



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 Sum of selected peaks
-log10 transformed
-mean centered



This slide shows PCA of the C10/C18 data set after normalization to the Sum of selected peaks, 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.

PCA Data Display and Interpretation

The following slides will cover the basics of PCA data display and interpretation.

Plotting Scores

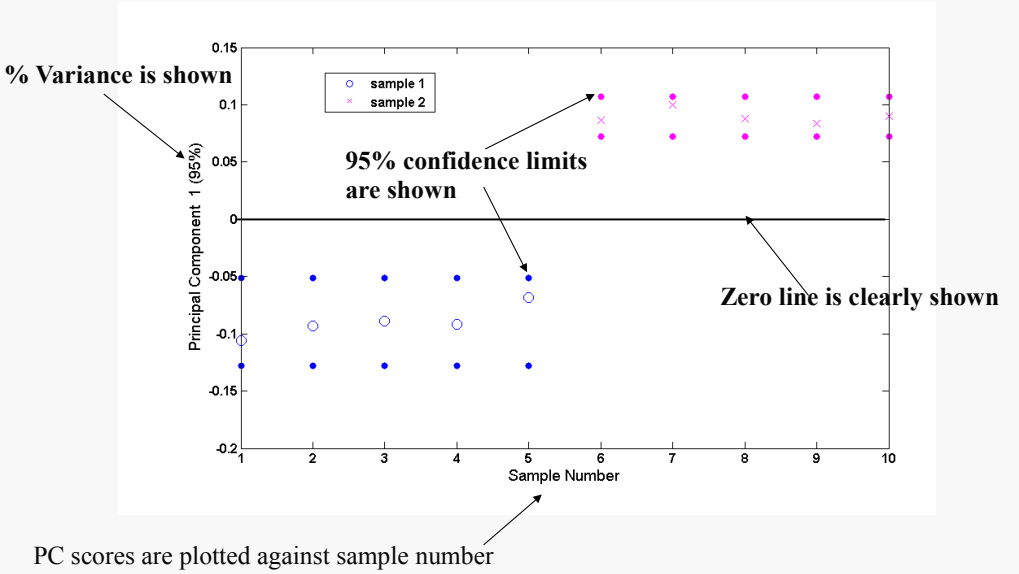
- Plotting software may vary
- It is easiest to interpret data in 1 dimension at a time
  - Plot PC vs Sample
  - If samples vary in systematic way you can plot PC vs variable of interest
- Sometimes it is necessary to plot 2 PCs against each other to see sample separation
- Always show % variance captured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



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 Loadings

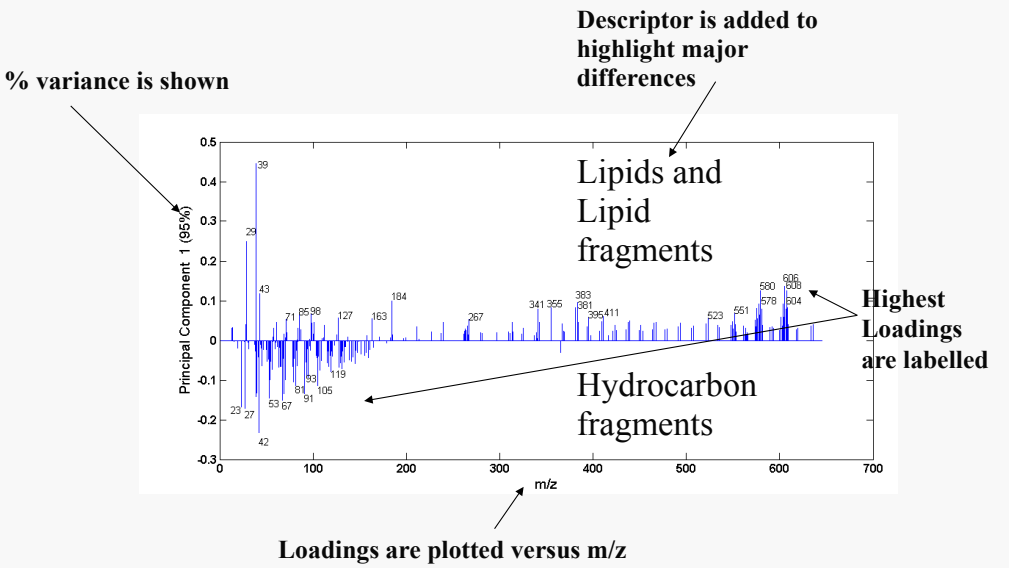
- Plotting software may vary
- It is easiest to interpret data in 1 dimension at a time
  - Plot PC vs m/z
    - This makes it so the loadings look more like a mass spectrum
- Always show % variance captured for each PC
- Only label highest loads to maintain clarity
  - You can explain other peak loadings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



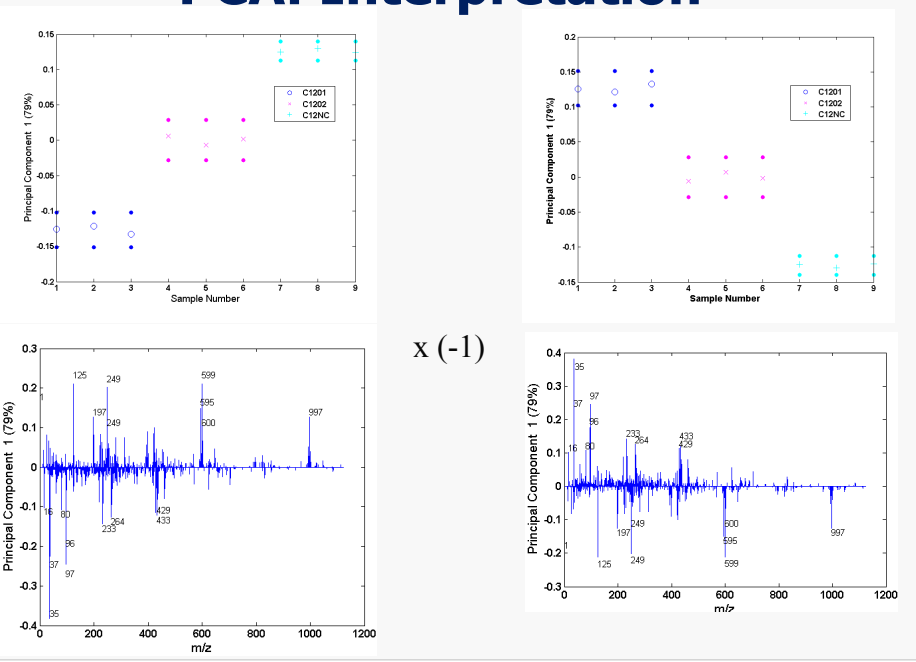
This slide shows an example PCA loadings plot. As seen in the figure, the loadings are plotted against m/z. The PC axis is clearly labeled. Peak labels are provided for the peaks with higher loadings values. It can also be seen that a general descriptor has been placed on either side of the PC1 axis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

PCA: Interpretation

- A note on negative scores and loadings
  - Do not be afraid of negative numbers
  - +/- have no intrinsic meaning
  - Only important to keep + and – scores and loadings together

This slide is just to enforce the concept that the + and – on PCA score and loadings axis do not have any specific meaning. They are only important to keep positive scores with positive loadings and negative scores with negative loadings.

PCA: Interpretation



In this example I show that the data on the left, is exactly the same as the data on the right, with the exception that the data on the right has been multiplied by -1. This does not change the relationship between the scorse and loadings. It is however important to remember that if you multiply the scores by -1 you must also multiply the loadings by -1.

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 correspond 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

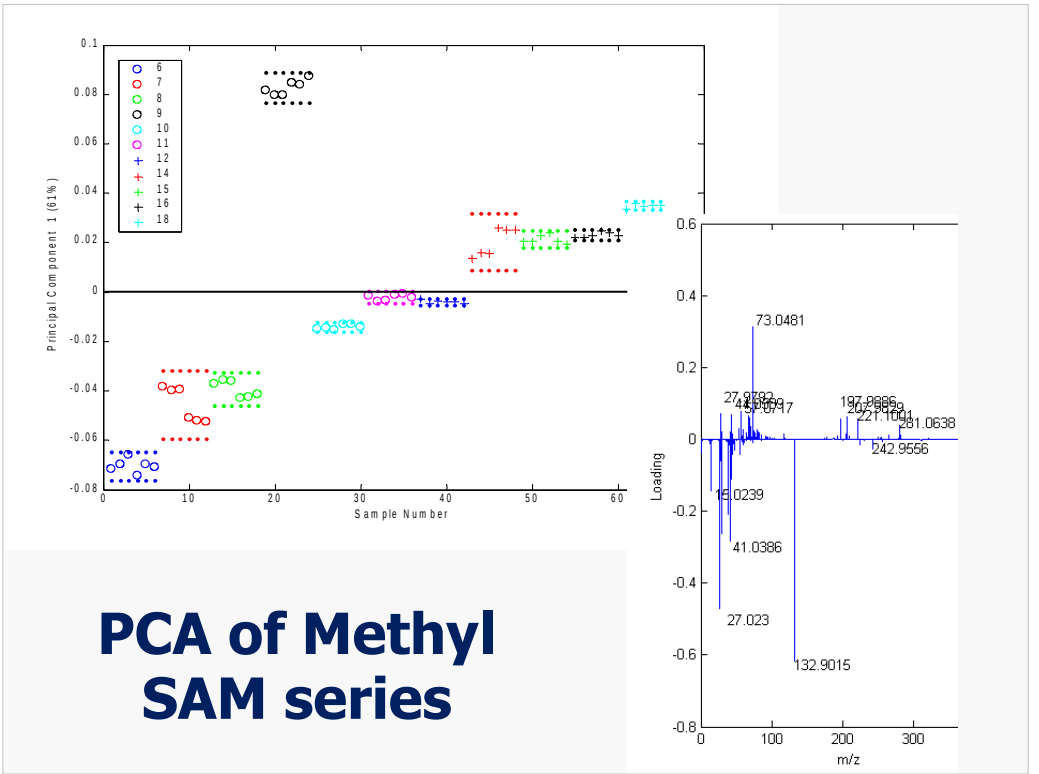
Interpretation of PCA results is done using the scores and loadings plots together. You cannot interpret PCA results by looking at either one alone. This is because the two plots contain complimentary information and either one without the other is incomplete. For example you can see clear separation 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.

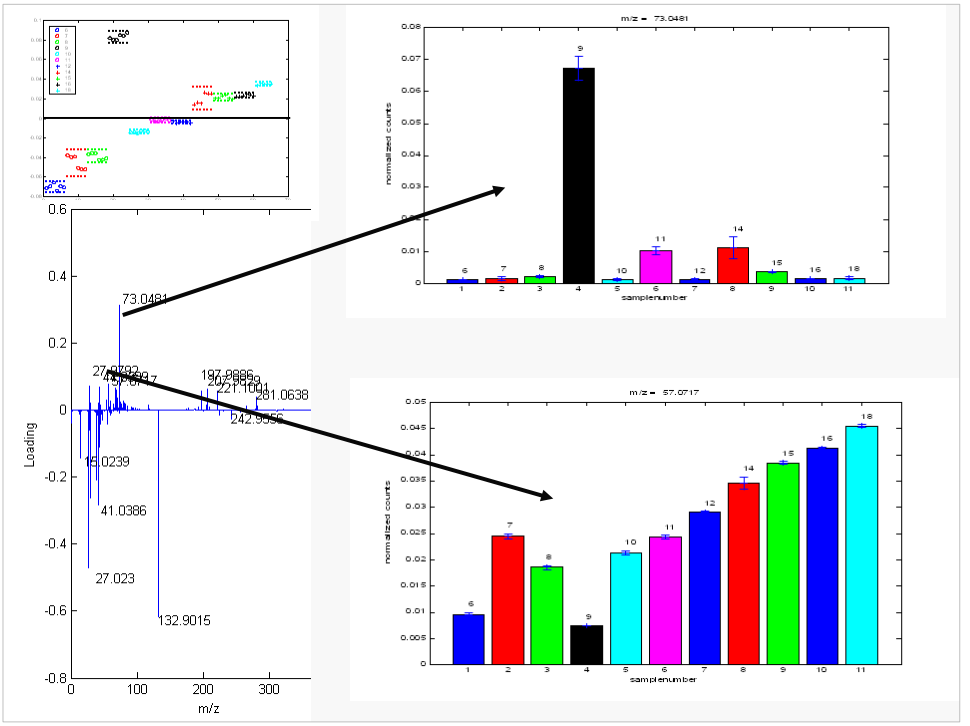
It is also true that samples with positive scores are negatively correlated with variables with negative loadings and that samples with negative scores are negatively correlated with variables with positive loadings.

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 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).



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

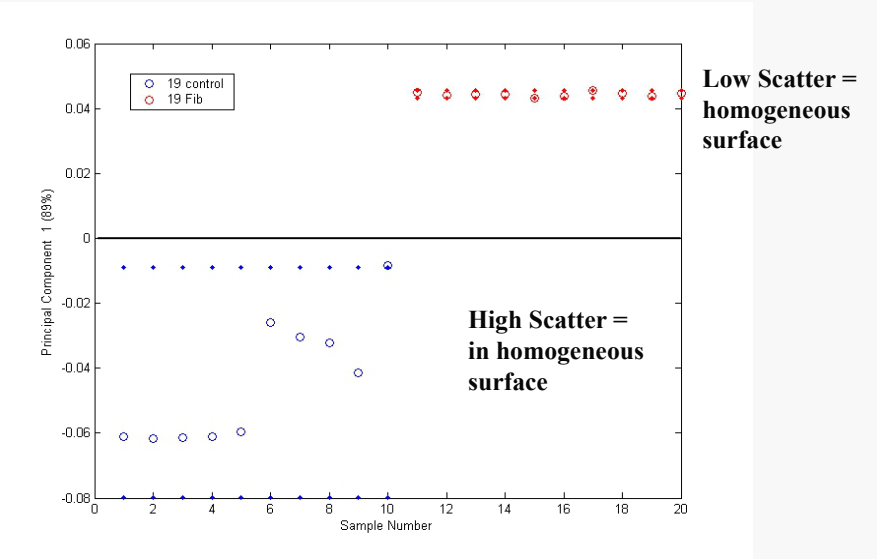
- Scores Plots
  - –Samples with similar scores are similar (clustered together)
  - –Samples with very different scores are different (separated from each other)
- Scatter in the scores for a given sample type suggests inhomogeneities in the sample
- Tight grouping of scores for a given sample suggests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



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.

Software Packages

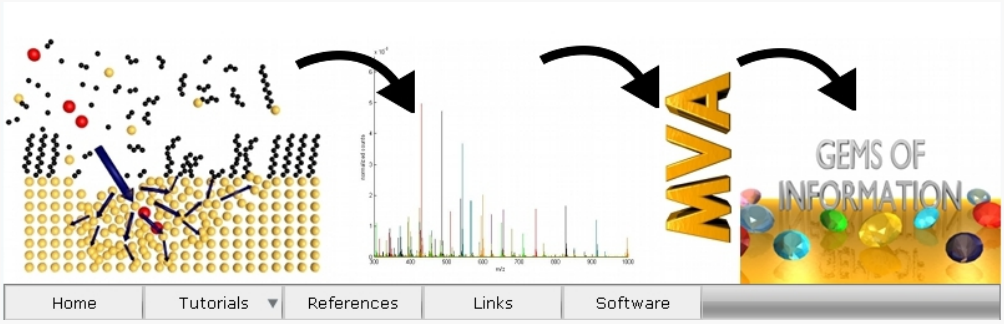
- Commercial
  - SAS
  - SPSS
  - S Plus
  - PLS Toolbox (requires Matlab)
- Open Source or Free
  - NB Toolbox (requires Matlab)
    - http://mvsa.nb.uw.edu/
  - R Project
  - Octave
  - Scilab

There are many programs that can be used to run PCA or other MVA methods. A few are listed on this slide.

Summary

- What are Multivariate statistics?
  - Powerful Tools for
    - Dealing with large data sets
    - Removing user bias
      - Finding patterns and trends
      - Building models
        - Prediction
        - classification
  - Hypothesis Generators
  - NOT a substitute for good scientific methodology
      - Well designed experiments
      - Replicates and controls
      - Validation of hypotheses

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.



- **mvsa.nb.uw.edu**
  - Tutorials
  - References
  - Links
  - Software

djgraham@uw.edu

NIH National Institute of Biomedical Imaging and Bioengineering

NESAC/BIO