Introduction to Principal Component Analysis (PCA)

NESAC/BIO

NESAC/BIO

Daniel J. Graham PhD

University of Washington
NESAC/BIO

MVSA Website
2010

Multivariate Analysis

- Multivariate analysis (MVA) methods have been applied to complex data systems for years
- Examples of MVA can be found for
  - IR
    - Anal. Chem. 1991, 63, 936-944; Anal. Chem. 1988, 60, 1202-1208; Anal. Chem. 1988, 60, 1193-1202; Appl. Spectrosc. 1985, 39, 73-84; Appl. Spectrosc. 1997, 51, 340-345.
  - ESCA
    - Surf. Interface Anal. 1997, 25, 942-947; Colloid Polym. Sci.1999, 277, 627-636; Surf. Interface Anal. 1997, 25, 105-110; Air and Waste 1993, 43, 729-735.
  - STM
    - Surf. Sci. 1994, 321, 276-286.
  - AFM
    - Thin Solid Films 1995, 264, 282-290.
  - Auger
    - J. Appl. Surf. Sci. 1993, 64, 41-57.
  - Other Mass Spec
    - Anal. Appl. Pyrolysis 1985, 9, 1-17; Anal. Chim. Acta 1997, 348, 389-407; Anal. Chem. 1997, 69, 4381-4389; Anal. Chem. 1989, 61, 715-719; Anal. Chem. 1983, 55, 81-88.; Int. J. Mass Spectrom. Ion Processes 1989, 89, 111-124. J. Chromatogr., A 1999, 840, 81-91; Int. J. Mass Spectrom. Ion Processes 1989, 89, 157-169; Anal. Chim. Acta 1983, 150, 45-52.

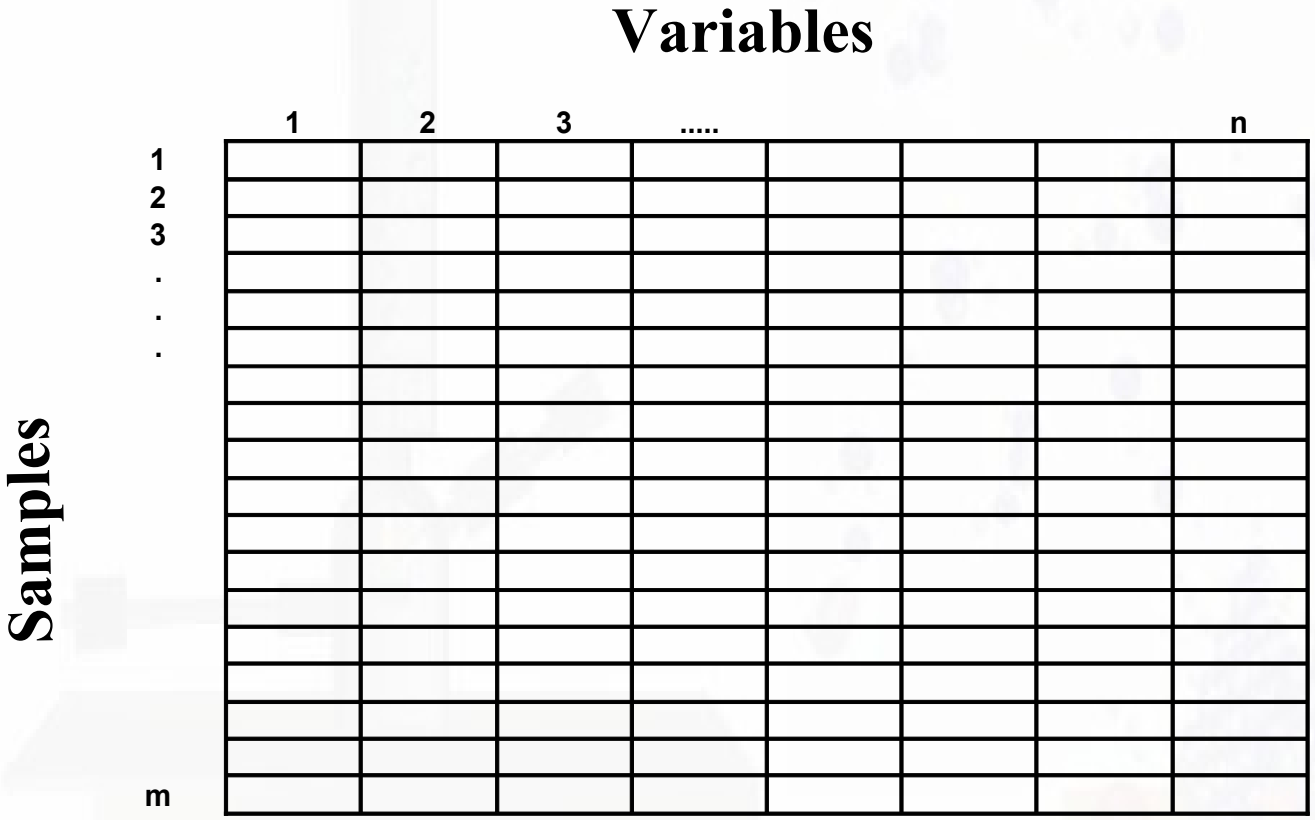
Multivariate Analysis Methods

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

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



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)

PCA data Pretreatment

- No standards have been set for data pretreatment
- Some common trends include
  - normalizing the data (many different ways)
  - mean centering for TOF-SIMS spectra
  - Autoscaling for TOF-SIMS images

Review of the literature using MVA for the interpretation of SIMS spectra

Problem	Data Pretreatment	MVA Technique	References
Polymer analysis			
Quantification of the composition of plasma polymerized thin films	P1+N2+S1	PLSR	[51,52]
Quantification of polymer molecular weight	P2+N3 or N5+S1	PCA	[53–56]
Quantification of polymer blend composition	P2+N3+S1	PCA	[57]
Discrimination between different polymers	P2+N2+S2+S1P3	PCA, NN	[58,59]
Characterization of multilayer polymer films	P1+N3+S1	PCA	[60]
Correlation of polymer surface chemistry with protein adsorption and cell growth	P1+N2+S1	PLSR	[61–63]
Detection and quantification of polymer additives on polymer surfaces	P2+N3+S1	PCA	[64,65]
Quantification of cross-linker density in polymer films	P1+N5+S1	PLSR	[66]
Protein adsorption			
Identification of adsorbed proteins	P2+N3+S1 or S4	PCA, LDA	[32,67–70]
Classification of adsorbed proteins	P3	NN	[71]
Detection of adsorbed protein conformation	P2+N3+S1	PCA	[72–74]
Quantification of protein adsorption on polymer surface	P2+S2+S1	PLSR	[75]
Detection of low amounts of adsorbed protein	P2+N3+S1	PCA	[76]
Quantification of multicomponent protein mixtures	P2+N3+S1	PLSR, SIMCA	[77–80]
Others			
Discrimination between solid-phase extraction stationary phases	P1+S2+S1	PCA, PLSR, NN	[81,82]
Analysis of SIMS spectra from alkanethiol SAMs	P1+N1+S1	PCA	[83,84]
Discrimination between bacterial samples	P2+N2+S1P3+N1+S3	PCAPLSR, LDA	[85,86]
Tracking assembly of a supported lipid monolayer	P1+N3+S1	PCA	[87]
Discrimination between different paint surfaces	P3+N1+S1	PCA, PLSR	[88]
Discrimination between different atmospheric aerosol particles	Not described	PCA	[14,16]
Interpretation of SIMS depth profiles	P2+N3	FA	[6,89]
Discrimination between different calcium phosphate phases	P1+N1+S1	PCA	[90]

Peak selection abbreviations P1: All peaks in the mass spectrum above background were selected; P2: Only peaks characteristic to the system of interest were selected; P3: The spectra were binned to 1 amu bins and all bins were used.

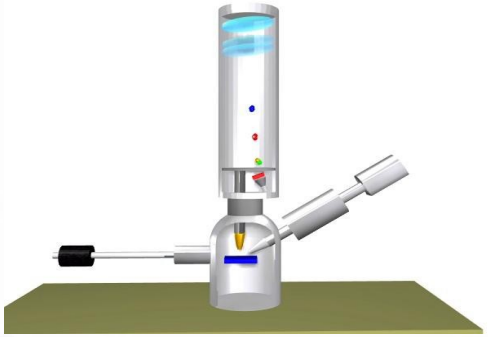
Normalization abbreviations N1: The intensity of each peak in each spectrum was normalized to the total secondary ion counts for that spectrum; N2: The intensity of each peak in each spectrum was normalized to the intensity of the most intense peak in that spectrum; N3: The intensities of the selected peaks were normalized to the sum of the intensities of the selected peaks for each spectrum; N4: The intensity of each peak in each spectrum was normalized to the total secondary ion counts for that spectrum less contributions from hydrogen (H^+/H^-) and contaminants (e.g. PDMS); N5: The intensity of each peak in each spectrum was normalized to a selected peak in the spectrum.

Centering/scaling/transformation abbreviations S1: Mean-centered; S2: The logarithm (either natural logarithm or $\log_{10}$) of the data set was taken before analysis; S3: Each peak was scaled to the mean of the highest peak within a 25 *m/z* window of the peak; S4: Autoscaled.

MVA abbreviations FA: Factor analysis; LDA: Linear discriminant analysis; MVA: Multivariate analysis; NN: Neural networks; PCA: Principal component analysis; PLSR: Partial least squares regression; SIMCA: Soft independent modeling of class analogy.

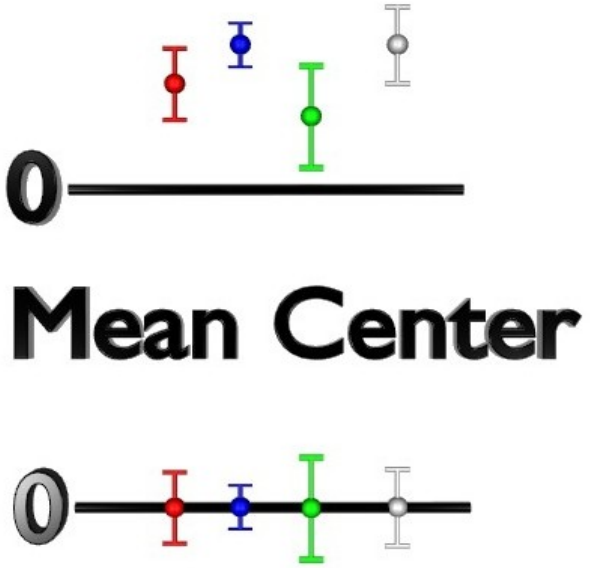
Normalization

- 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



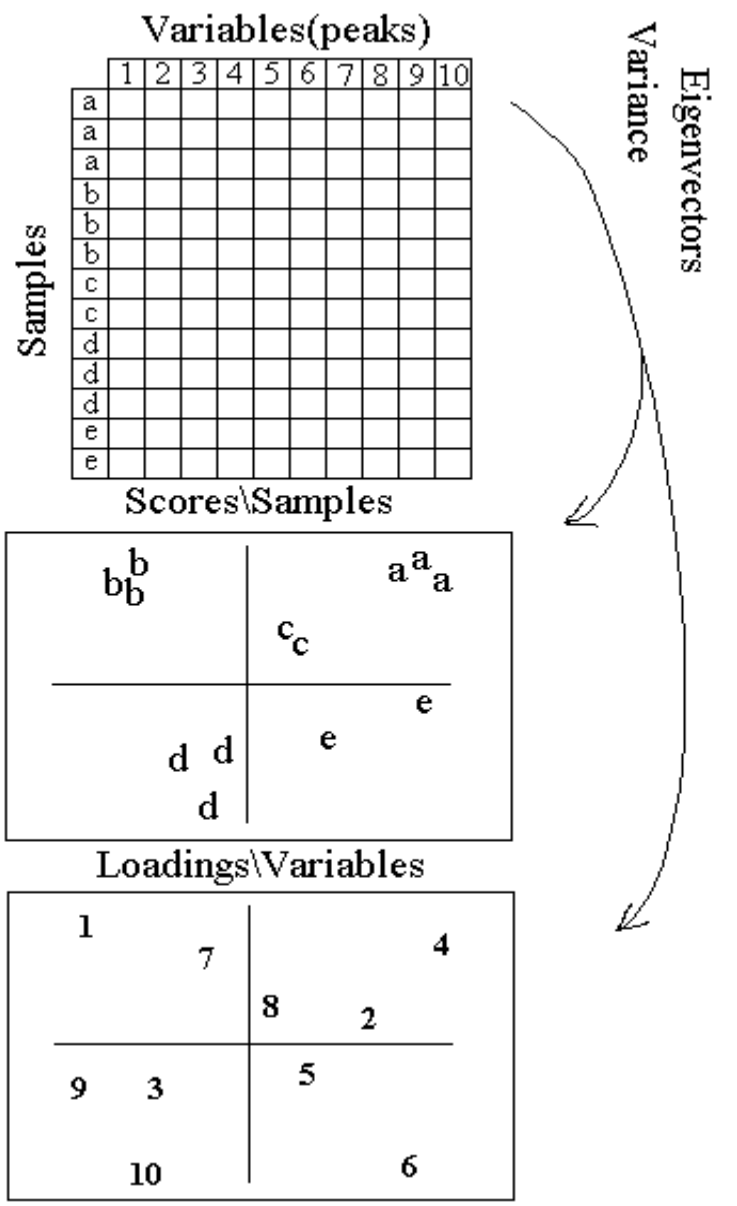
- **Mean centering**
  - **Subtracts the mean of each column (variable) from each column element**
  - **Centers data so that all variables vary across a common mean of zero**

Scaling

- Scaling attempts to account for differences in variance scales between variables
- There is some debate about whether TOF-SIMS data should be scaled or not
- Autoscaling for SIMS images is common
  - Divides mean centered variables by their standard deviation
    - Results in variables with unit variance
- Other scaling methods have been proposed
  - No consensus on what is the “best” way

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



PCA Allows "quick" data summary

- Scores
  - Tell relationship between samples
    - Are they similar or different?
  - Give an idea of the reproducibility of samples
- Loads
  - Show which variables are responsible for sample differences
  - Can help determine sample differences across entire peak set

PCA Mathematics

- **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**
- **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**

$$X = T_1 P_1^T + E$$

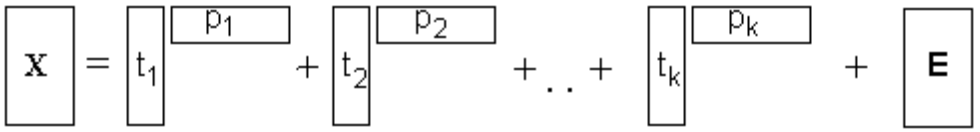
Residual becomes new X matrix

$$X = T_2 P_2^T + E$$

var(PC1)>var(PC2)>var(PC3)>...>var(PCk)

PCA Mathematically

- **PCA decompositon**
  - $X = t_1 p_1^T + t_2 p_2^T + \ldots + t_k p_k^T + E$



P_i (loadings) are the eigenvectors of the covariance matrix

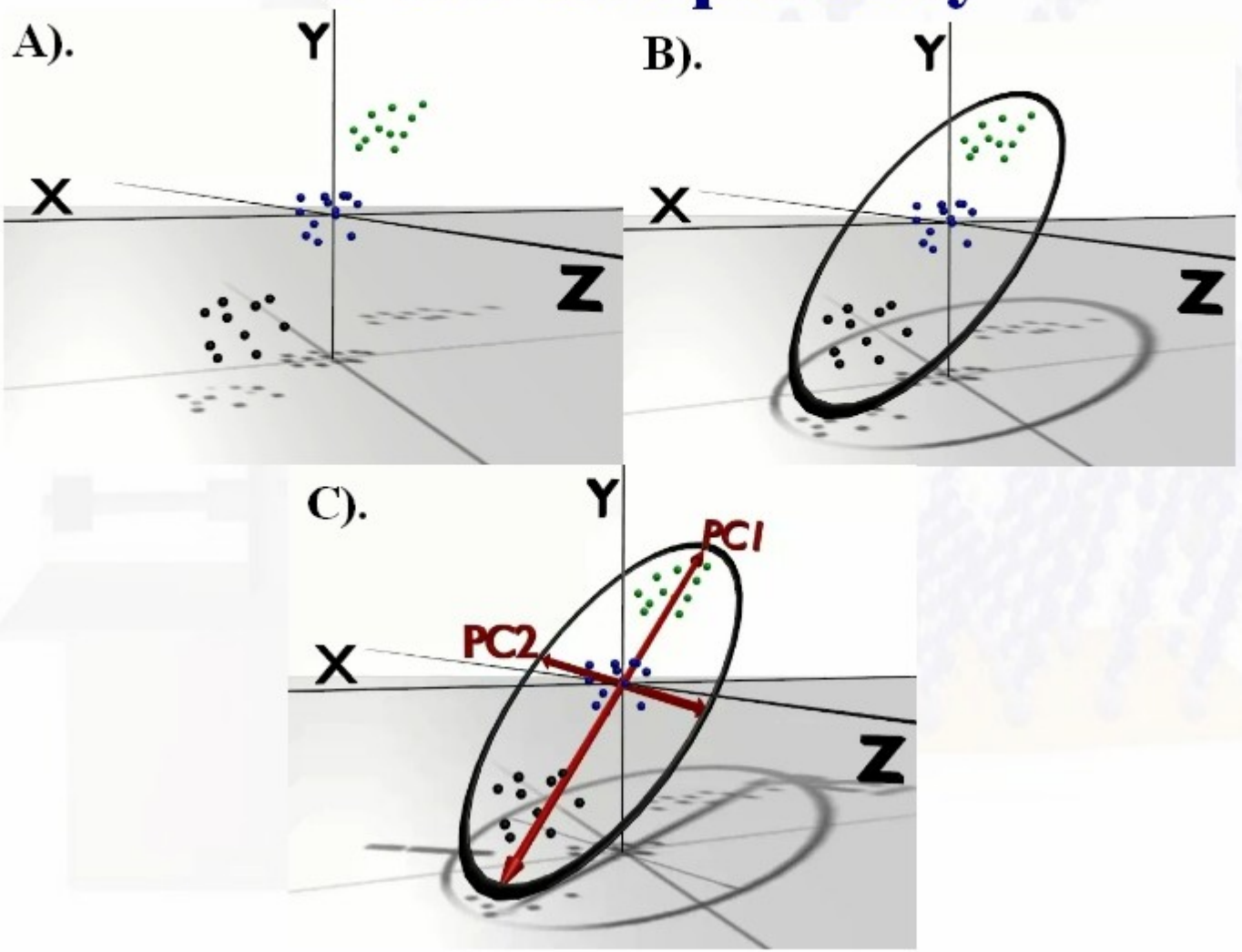
$$\mathrm{cov}(X)\,p_i = \lambda_i p_i$$

λ_i are the eigenvalues. They describe the amount of variance captured by each $t_i p_i$ pair (PC)

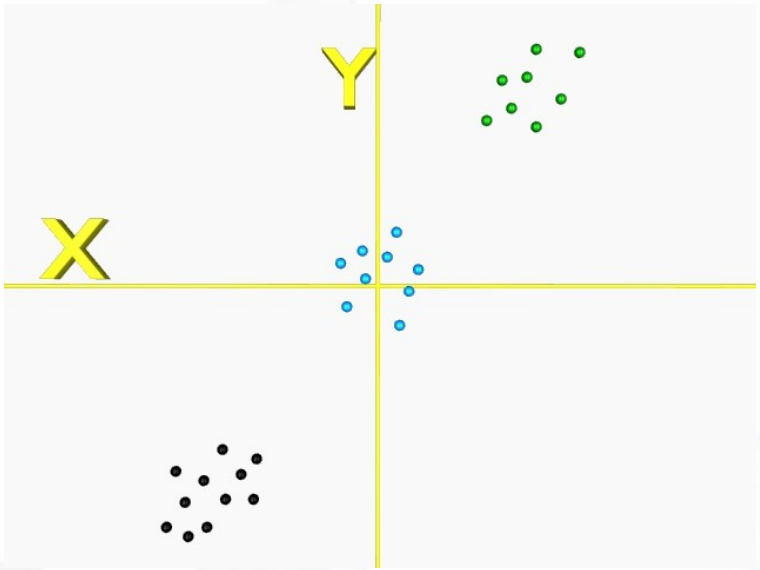
PCA Mathematically Continued

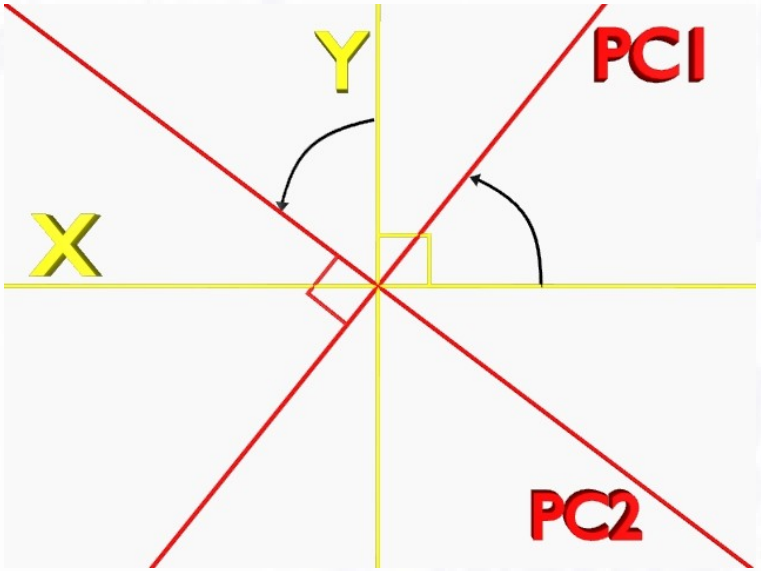
- Original matrix is reconstructed into new set of matrices
  - $X = TP^T + E$
    - T = scores
    - P = loadings
    - E = residual (random noise)
  - One common way of doing PCA is by a singular value decomposition
    - $cov(X) = USV^T$
      - P = V Eigenvectors (loadings)
      - T = US (scores)
      - $S_{ii} = sqrt(\lambda_i)$ (λ_i = eigenvalues)

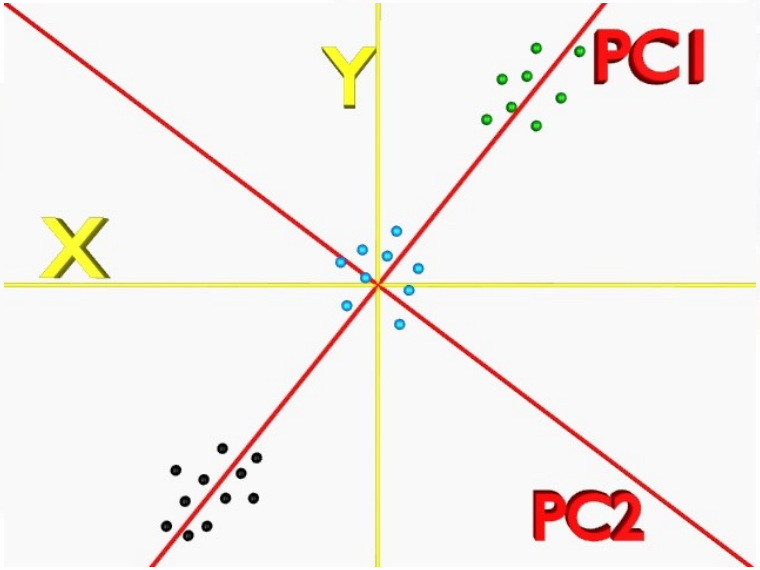
PCA Graphically



PCA Graphically



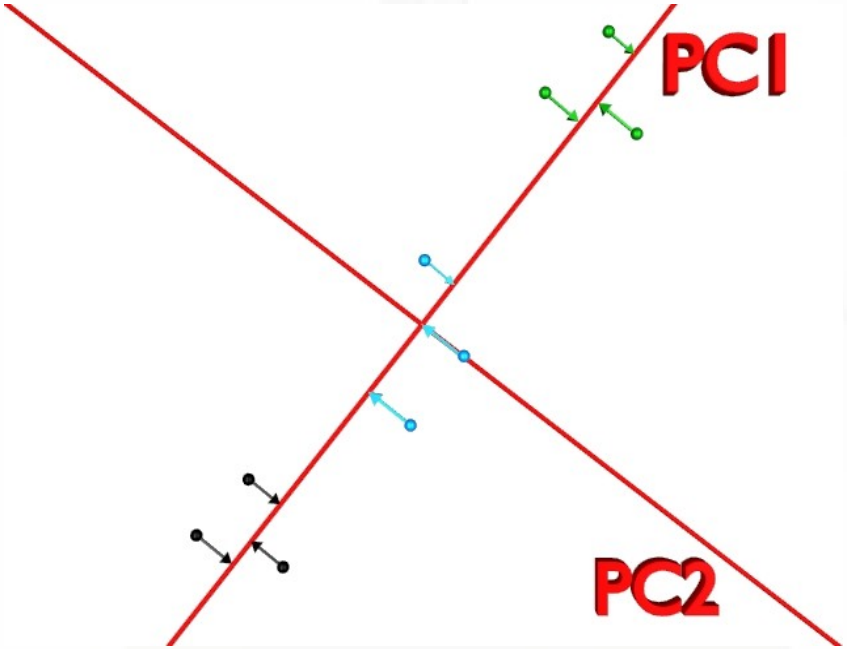




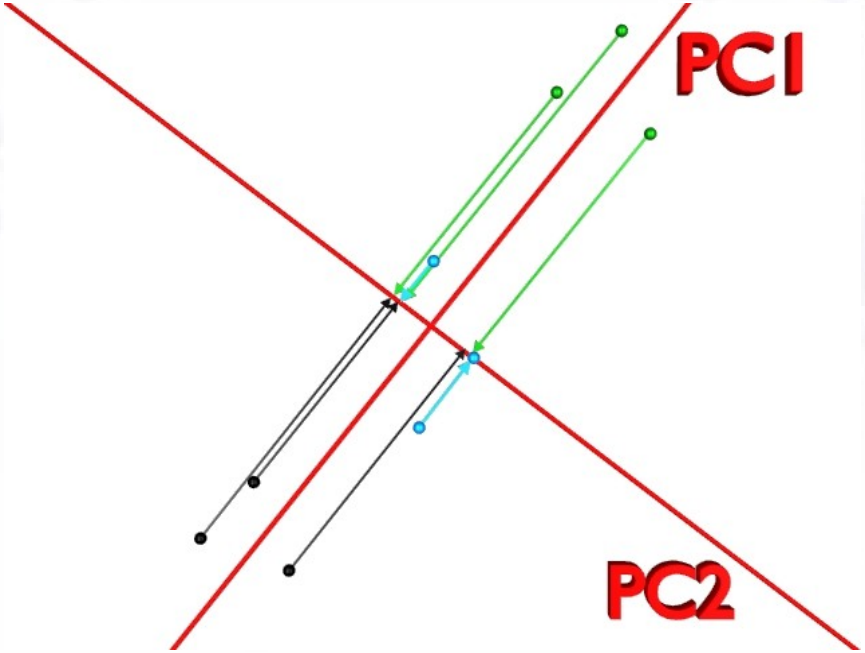
PCA Scores

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

Scores tell the relationship (spread) between the samples



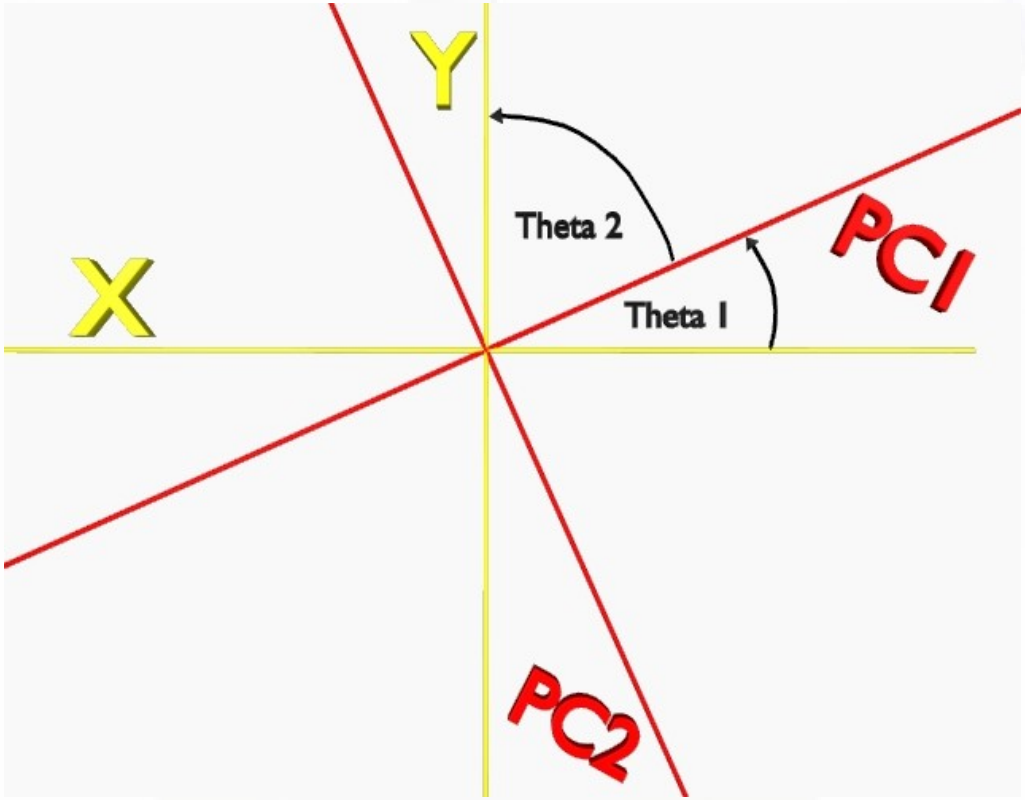
Projection onto PC1



Projection onto PC2

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

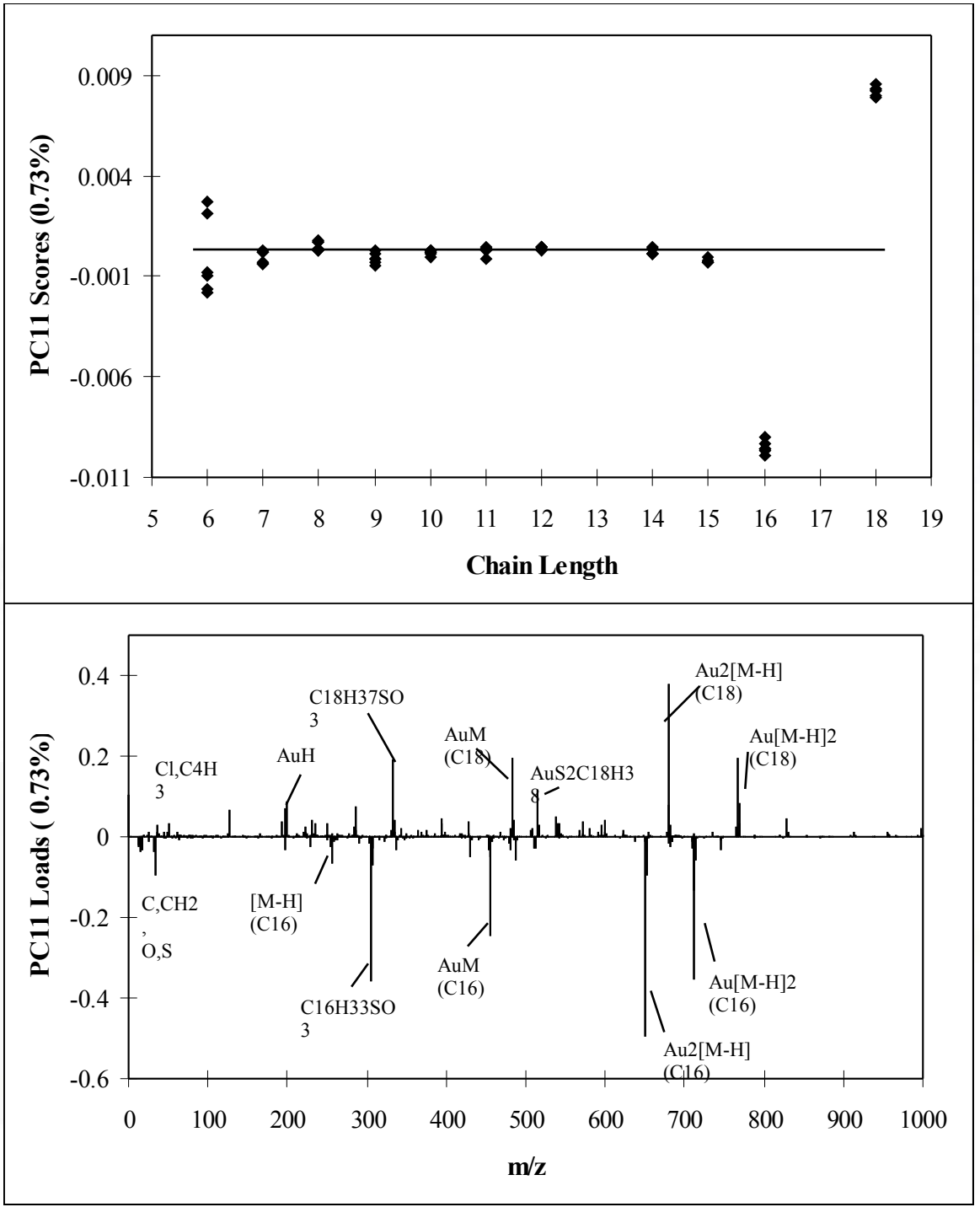
The loadings tell which variables are responsible for the separation seen between samples

I ran PCA now what do I have?

- A set of PC scores and loadings
  - Each PC captures the greatest amount of variation in the given direction
  - %var PC1 > PC2 > PC3 > PC4 ... > PCn
- % variance tells relative amount of information captured by a given PC, but you need to check the scores and loadings to determine if the PC contains useful information

Quick Example %variance

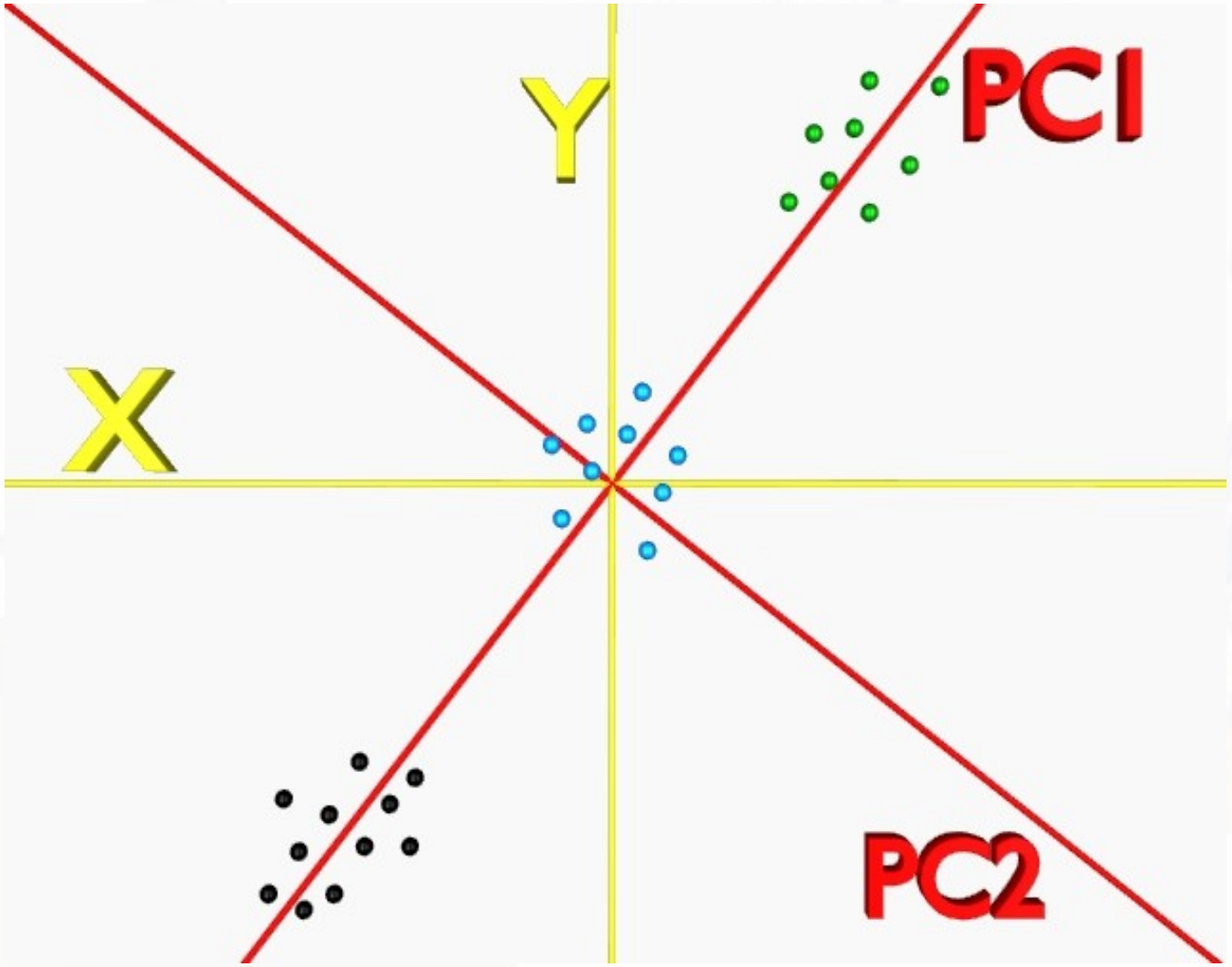
- Even PC11 contains useful information about the samples even though it only captures 0.73% of the variance in the data
- Separates C16 and C18 samples



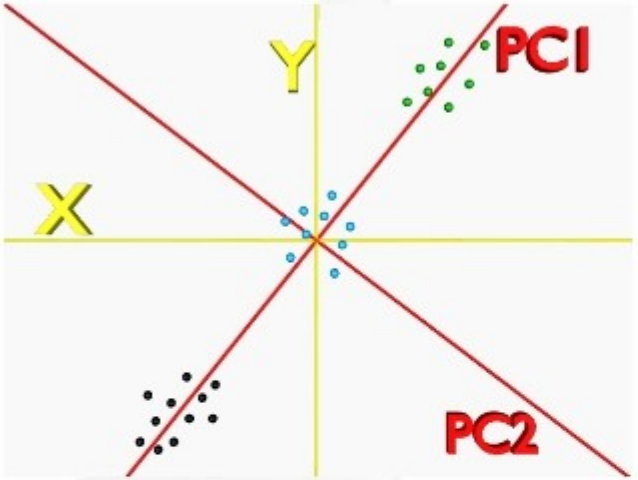
PCA Interpretation

- Most easily visualized by a set of plots of the scores and loadings
- Scores and loadings are compared together
- Trends seen in the PCA plots should be verified by the raw data

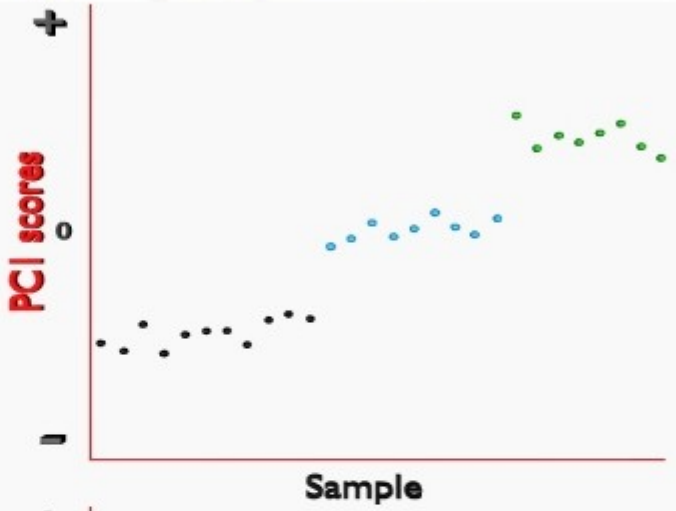
PCA Example: Synthetic data

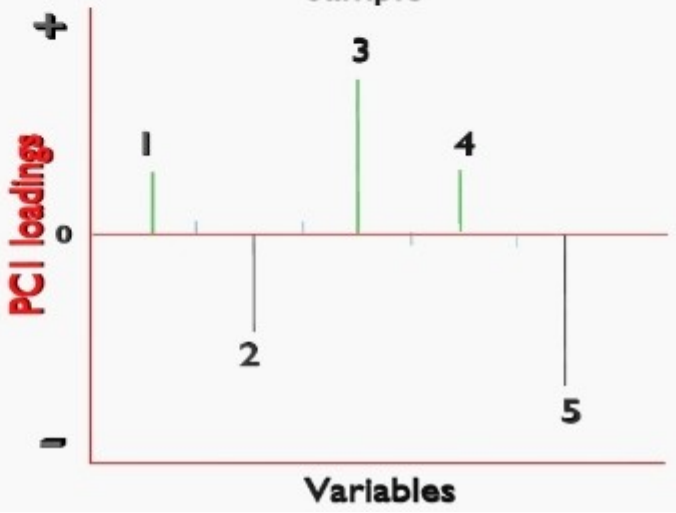


PCA Interpretation

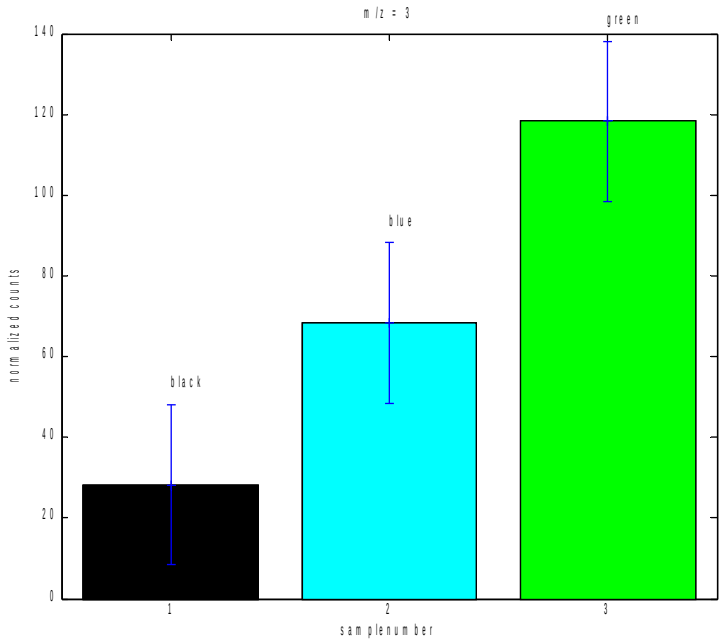


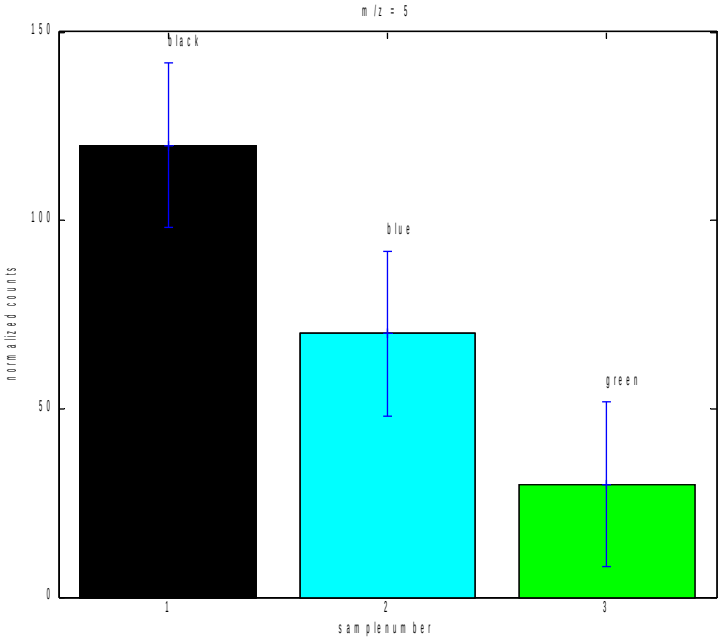
- Samples separated on PC1
- Loadings show variables responsible for separation

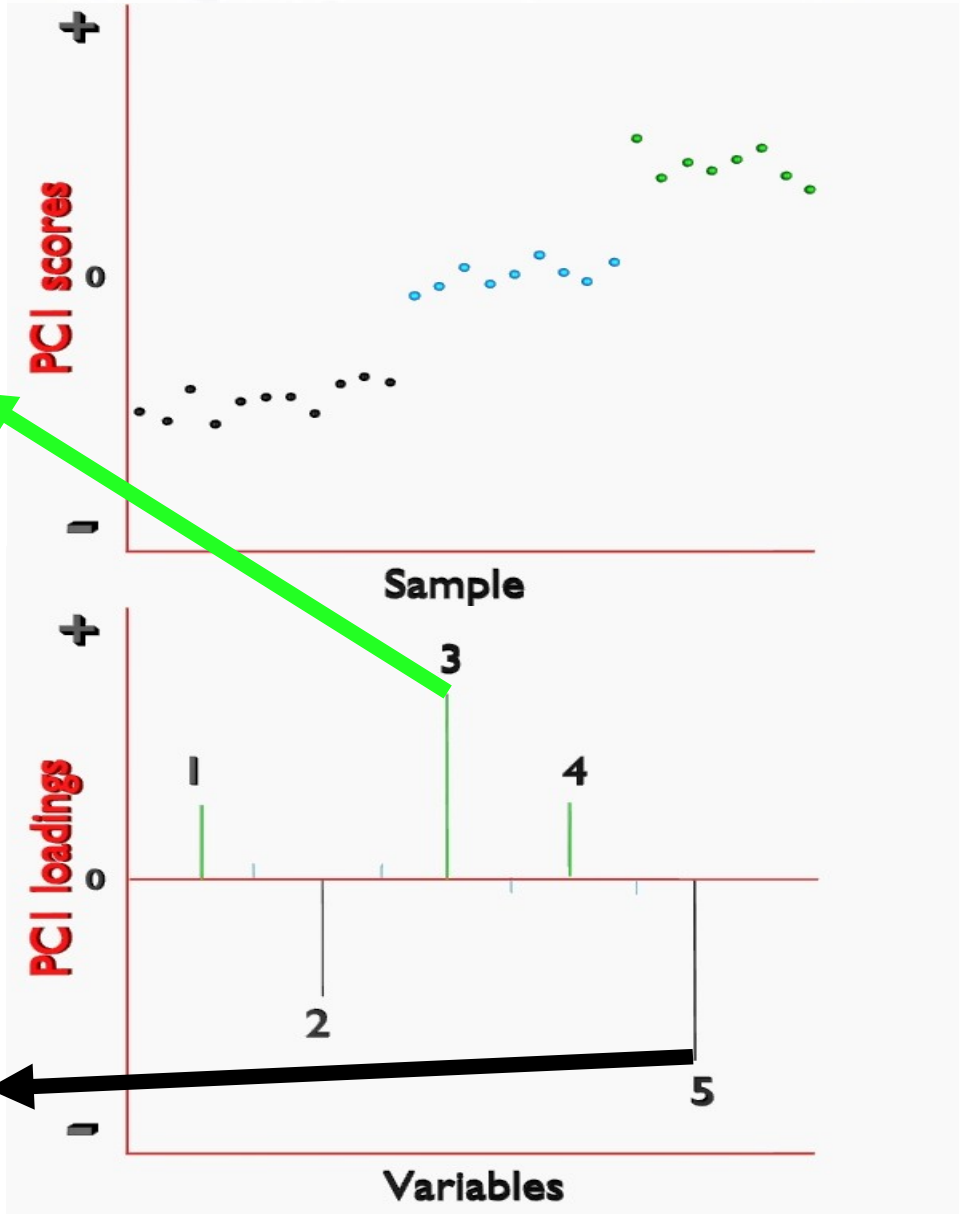




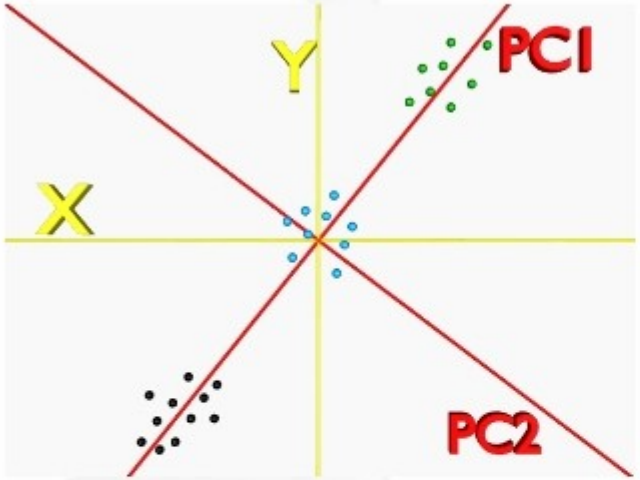
rpretation


m/z = 3
green
blue
black
normalized counts
sample number

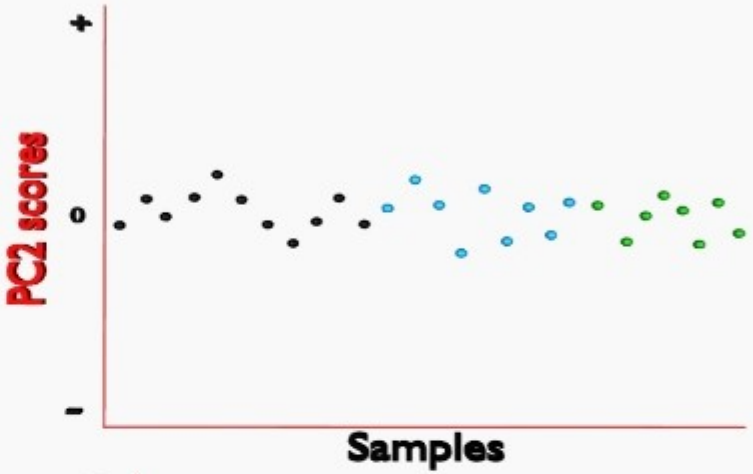

m/z = 5
black
blue
green
normalized counts
sample number

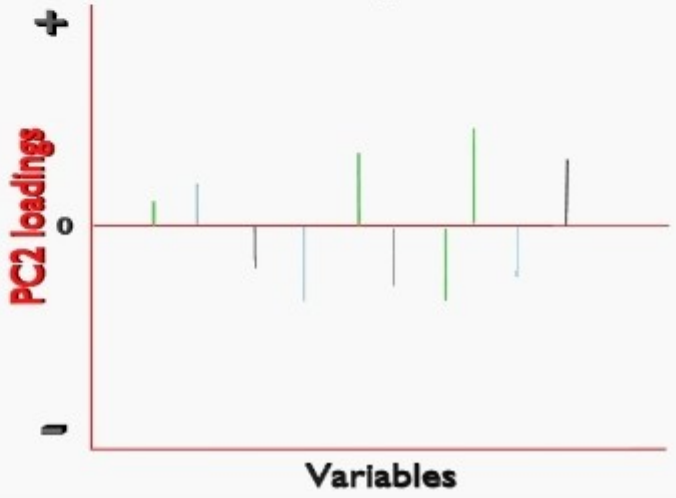

PCI scores
Sample
PCI loadings
Variables
1
2
3
4
5

PCA Interpretation Cont.



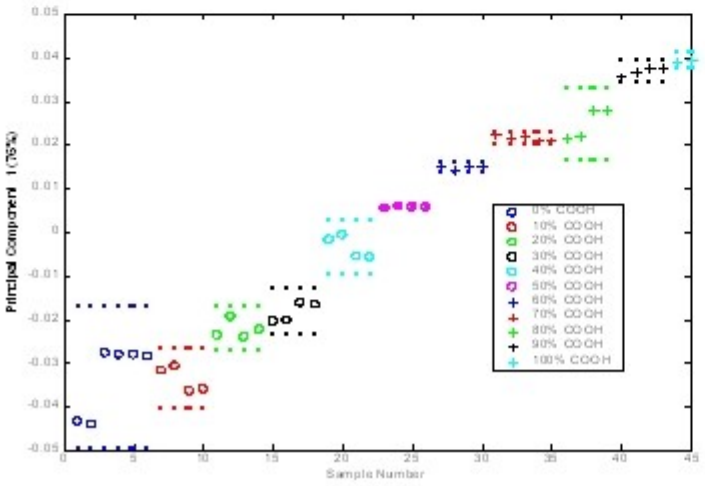
- Samples overlap on PC2
- Loadings are random

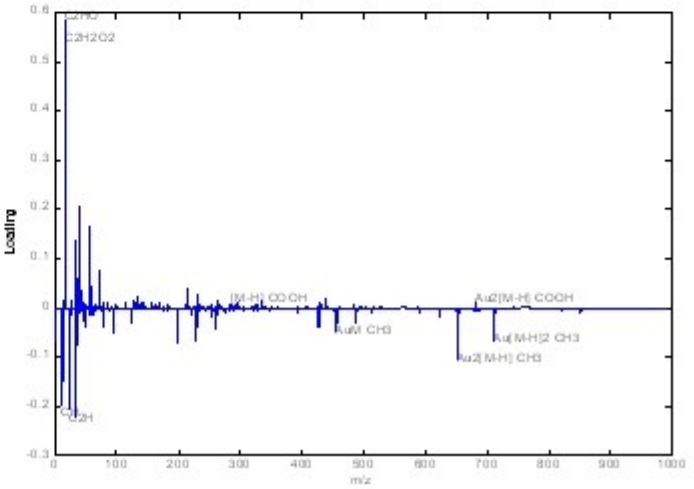


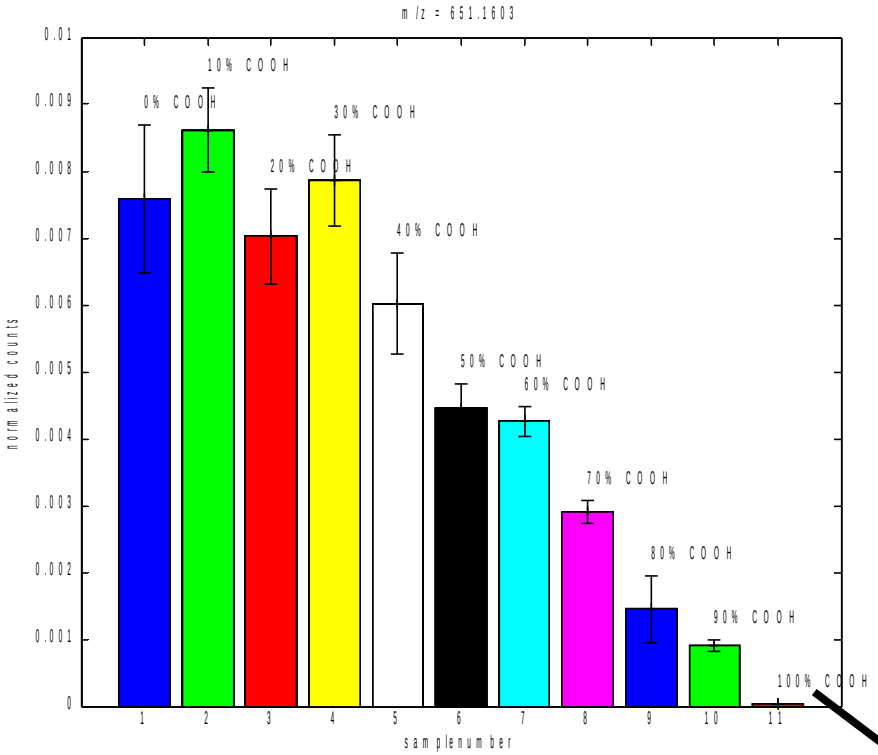


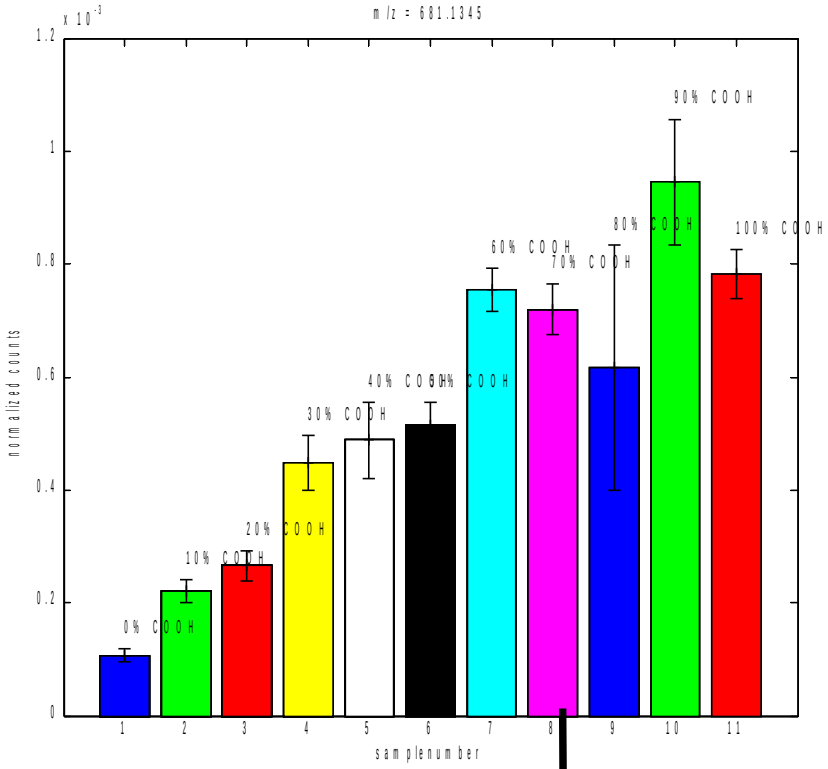
Interpretation

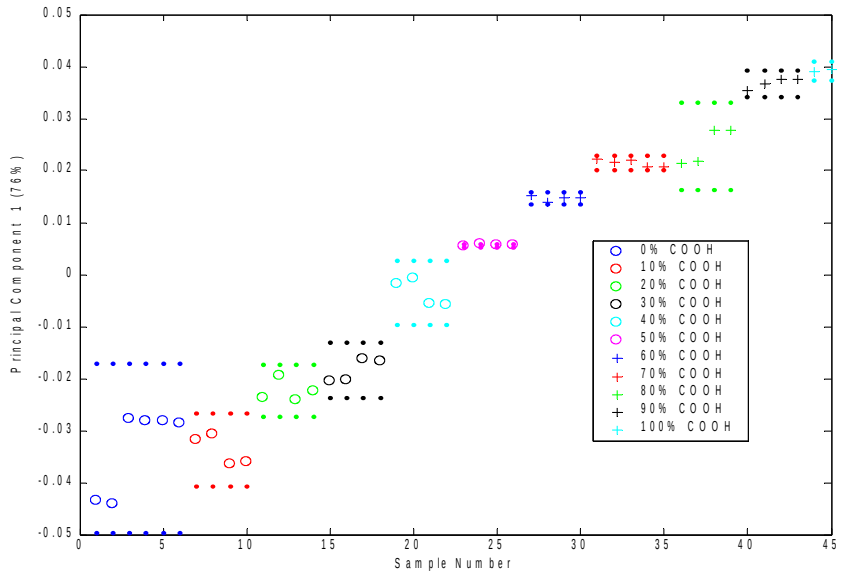
- How to interpret Scores and Loadings plots
- How separated are the samples?
  - Use 95% confidence limits to check
- Check the raw data!

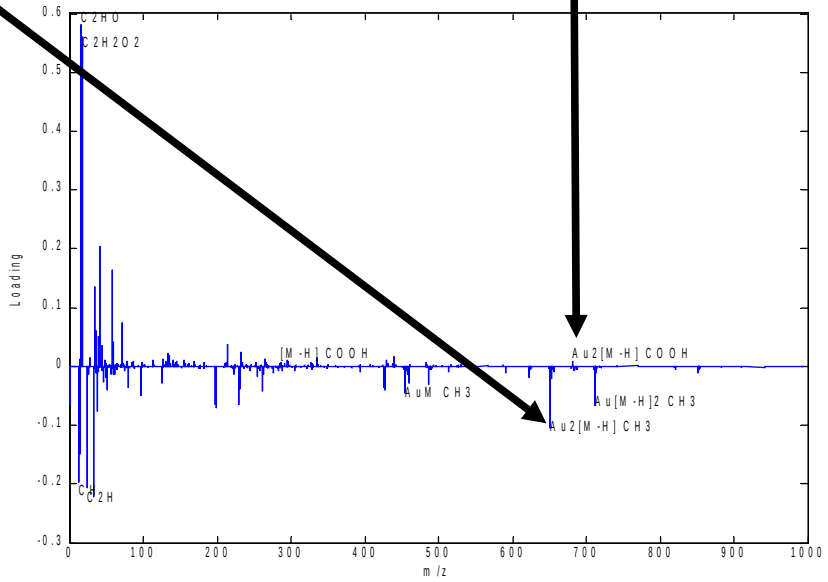


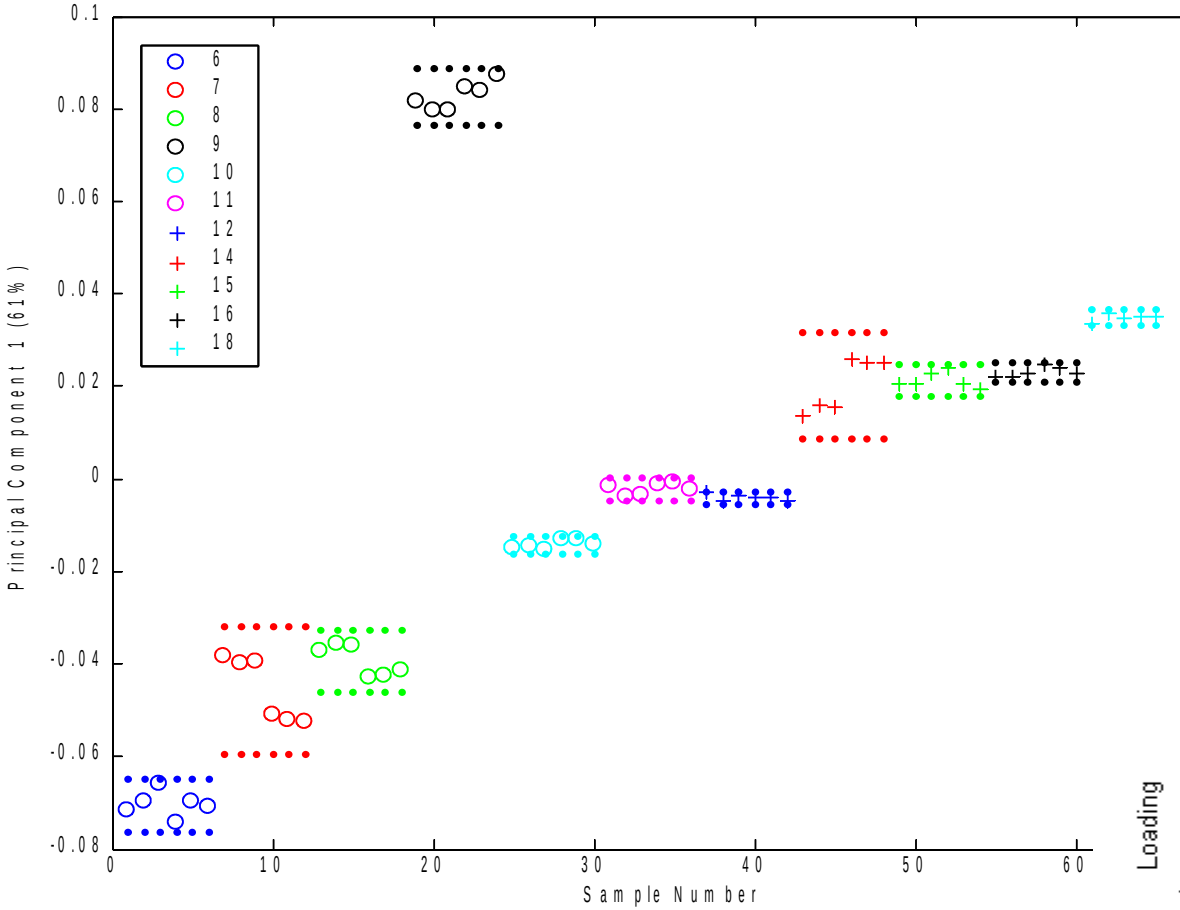


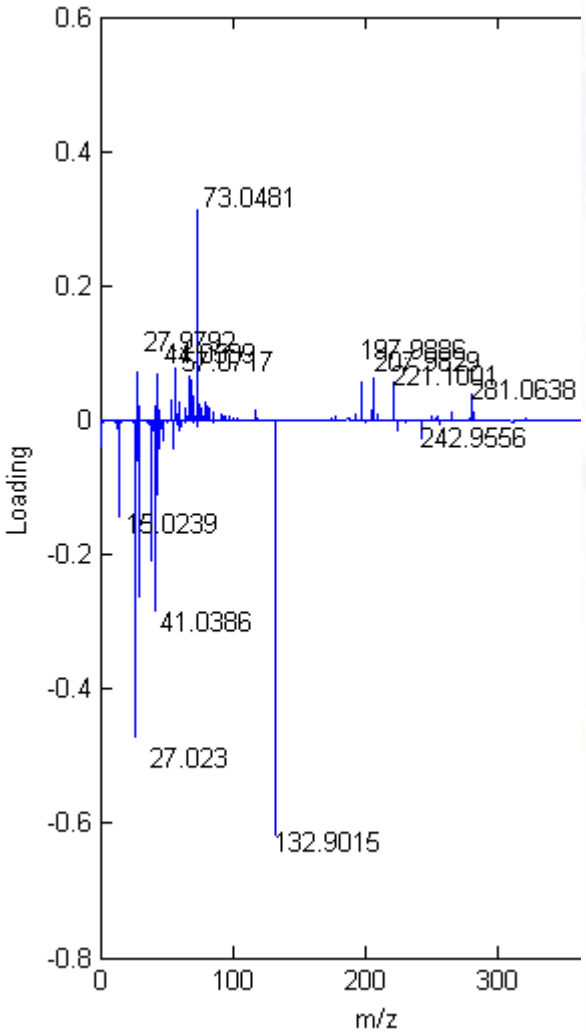

m/z = 651.1603
normalized counts
samplenumber
0% COOH
10% COOH
20% COOH
30% COOH
40% COOH
50% COOH
60% COOH
70% COOH
80% COOH
90% COOH
100% COOH


m/z = 681.1345
x 10^-3
normalized counts
samplenumber
0% COOH
10% COOH
20% COOH
30% COOH
40% COOH
50% COOH
60% COOH
70% COOH
80% COOH
90% COOH
100% COOH

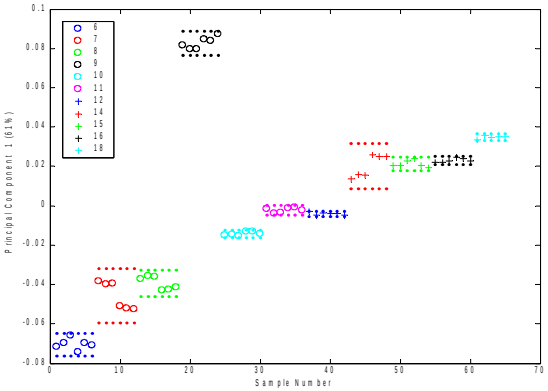

Principal Component 1 (76%)
Sample Number
0% COOH
10% COOH
20% COOH
30% COOH
40% COOH
50% COOH
60% COOH
70% COOH
80% COOH
90% COOH
100% COOH

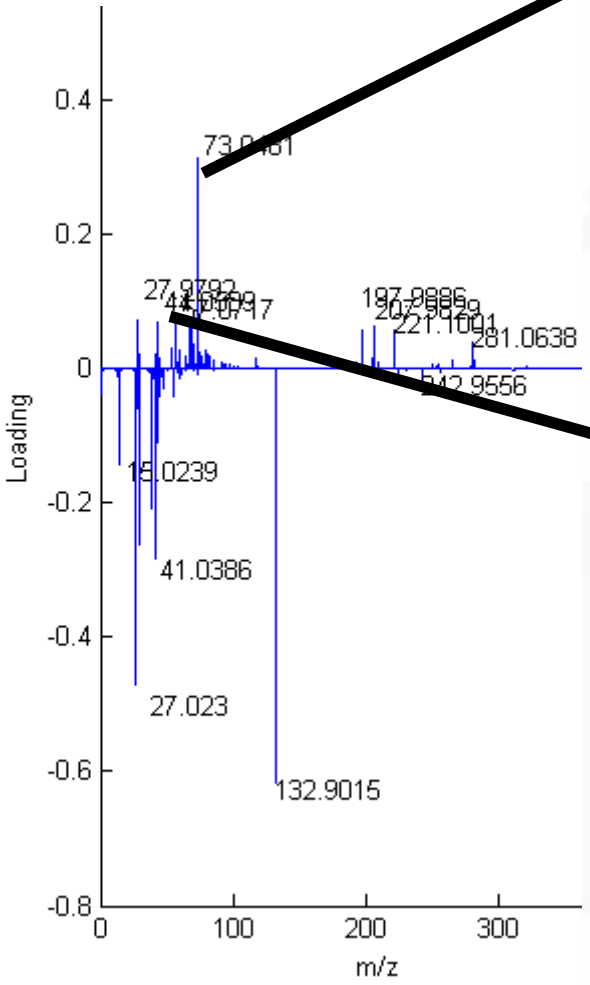

Loading
m/z
C2HO
C2H2O2
[M-H]COOH
AuM CH3
Au2[M-H]COOH
Au[M-H]2 CH3
Au2[M-H]CH3
C2H

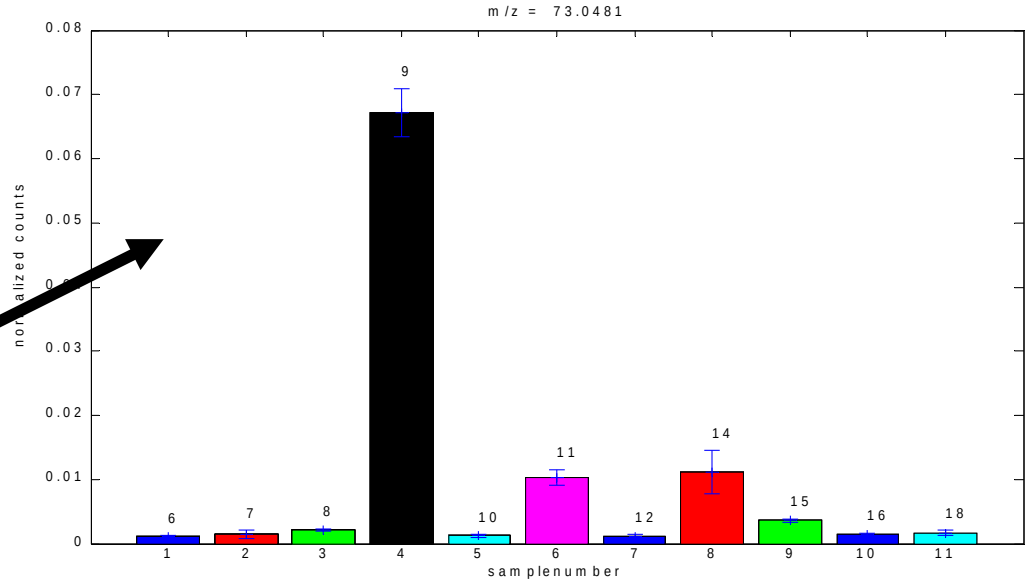


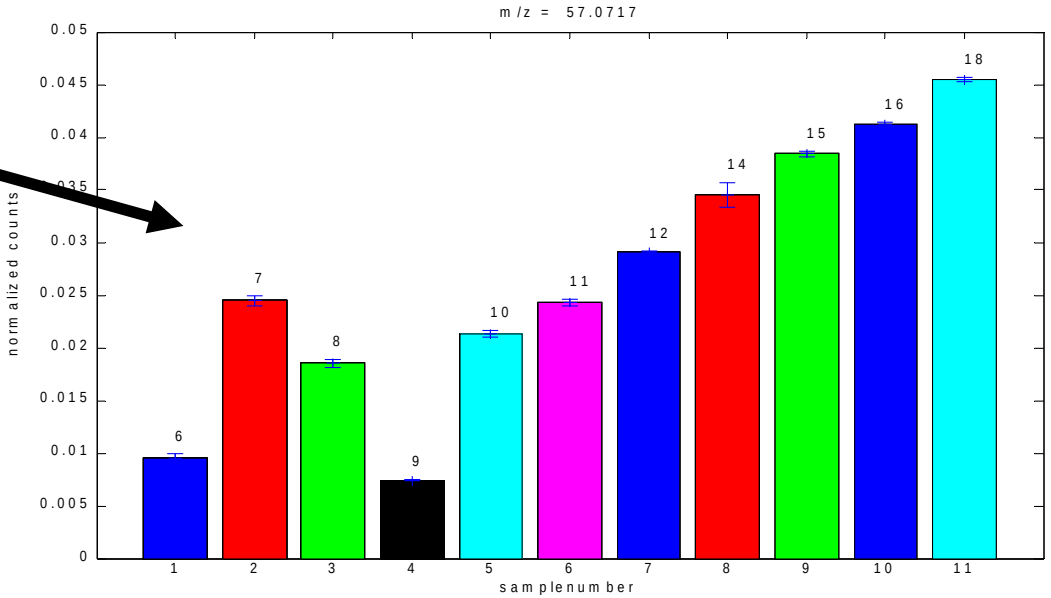


PCA of Methyl SAM series

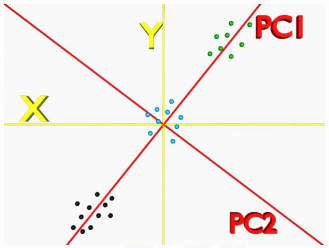

Principal Component 1 (61%)
Sample Number


Loading
m/z
73.0481
27.9782
44.0717
197.9886
221.1091
281.0638
342.9556
15.0239
41.0386
27.023
132.9015


m/z = 73.0481
normalized counts
samplenumber


m/z = 57.0717
normalized counts
samplenumber

PCA things to consider



- **Determines the largest directions of variance within the data regardless of what that variance is from**
  - **PCA is an unsupervised method (no input to what the samples are)**
  - **If 1 sample is contaminated, PCA will likely separate that sample from the rest in one of the first PC's**
  - **If sample differences are not the greatest source of variance in the data set, PCA may not separate out the samples**

PCA things to consider

- The decision to use PCA should be part of your experimental plan, not just an afterthought after collecting your data
- You should understand what you are doing and how it works
- You should understand the assumptions made for running PCA
- You should check your results with the raw data

For More Information about PCA See ...

- JE Jackson (1980) Principal Components and Factor Analysis: Part I - Principal Components. Journal of Quality Technology 12:201-213
- JE Jackson (1991) A Users's Guide to Principal Components. John Wiley & Sons, Inc.,New York
- S Wold, K Esbensen and P Geladi (1987) Principal Components Analysis. Chemometrics and Intelligent Laboratory Systems 2:37-52

Introduction to Principal Component Analysis (PCA)

Daniel J. Graham PhD

University of Washington
NESAC/BIO

MVSA Website
2010

This presentation is aimed at providing a general overview of principal components analysis (PCA) as applied to TOF-SIMS data. This introduction is not meant to provide a rigorous explanation of PCA mathematics. For this the reader is directed to the LINKS section of this website where adequate references are provided for the interested reader.

This site has been put together by the voluntary effort of many people with the best intentions of providing accurate, useful information regarding MVA of TOF-SIMS data. Any mistakes are purely unintentional.

Multivariate Analysis

- **Multivariate analysis (MVA) methods have been applied to complex data systems for years**
- **Examples of MVA can be found for**
  - **IR**
    - **Anal. Chem. 1991, 63, 936-944; Anal. Chem. 1988, 60, 1202-1208; Anal. Chem. 1988, 60, 1193-1202; Appl. Spectrosc. 1985, 39, 73-84; Appl. Spectrosc. 1997, 51, 340-345.**
  - **ESCA**
    - **Surf. Interface Anal. 1997, 25, 942-947; Colloid Polym. Sci.1999, 277, 627-636; Surf. Interface Anal. 1997, 25, 105-110; Air and Waste 1993, 43, 729-735.**
  - **STM**
    - **Surf. Sci. 1994, 321, 276-286.**
  - **AFM**
    - **Thin Solid Films 1995, 264, 282-290.**
  - **Auger**
    - **J. Appl. Surf. Sci. 1993, 64, 41-57.**
  - **Other Mass Spec**
    - **Anal. Appl. Pyrolysis 1985, 9, 1-17; Anal. Chim. Acta 1997, 348, 389-407; Anal. Chem. 1997, 69, 4381-4389; Anal. Chem. 1989, 61, 715-719; Anal. Chem. 1983, 55, 81-88.; Int. J. Mass Spectrom. Ion Processes 1989, 89, 111-124. J. Chromatogr., A 1999, 840, 81-91; Int. J. Mass Spectrom. Ion Processes 1989, 89, 157-169; Anal. Chim. Acta 1983, 150, 45-52.**

Multivariate methods have been applied to many different analytical techniques over the years. These include, but are not limited to IR, ESCA, STM, AFM, Auger and various mass spectrometric techniques. The application of multivariate methods to these techniques has been largely driven by the need to process large sets of complex data in a reasonable amount of time, while maintaining the ability to extract all the pertinent chemical information within the data.

As with many of the methods listed above, TOF-SIMS is ideally suited for multivariate analysis. A typical TOF-SIMS spectra contains hundreds of peaks. The intensities of many of these peaks are correlated due to the fact that they come from the same surface species. Determining how these peaks are related and how their relative intensities correspond to the differences in the given surface chemistries is a challenge well suited for multivariate analysis.

Multivariate Analysis Methods

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

There is an alphabet soup of multivariate analysis methods available for data processing. These include factor based methods such as PCA, FA, and DA and other methods such as MCR, and neural networks.

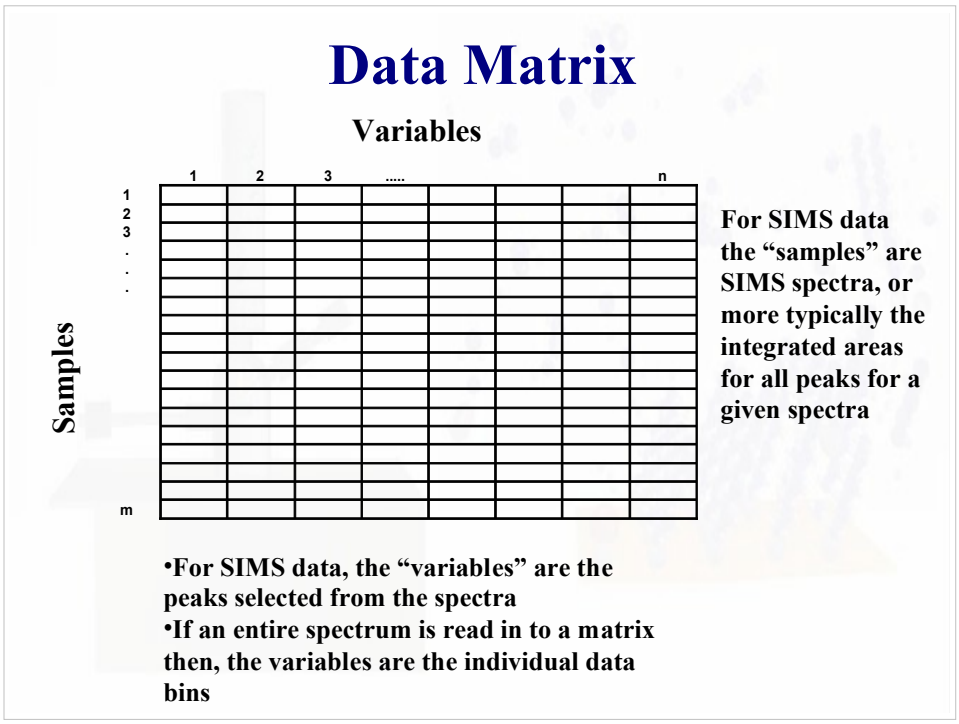
This presentation will focus on PCA since it is the most commonly used method for TOF-SIMS analysis and because it forms a basis for many other methods. PCA is a multivariate analysis method that determines the largest directions of variance within a data set. PCA mathematics are founded in linear algebra. The PCA solution is determined by finding the eigenvectors and eigenvalues of the variance covariance matrix of a data set. Graphically PCA is an axis rotation that creates a new set of axes that define the major directions of variance within a data set.

The results of PCA are the scores and loadings. The scores give the relationship between the samples and tell the amount of each sample on the PC axes. The loadings show which variables are responsible for the differences seen in the scores plot. Mathematically the loadings are the direction cosines between the original variables and the new PC axes. The loadings are the weightings given to the original variables to produce the new PC axes. The PCA axis rotation enables reducing data sets with potentially hundreds of variables down to a few, relatively easy to interpret, variables.

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**

Before we begin looking into PCA, we must first establish some definitions. Within this presentation it will be assumed that all data is arranged in an m x n matrix where samples (m) are in rows and variables (n) are in columns. 'k' will represent the number of PCs in a given PCA model. 'T' will represent the scores matrix and 'P' will represent the loadings matrix.



This slide shows the arrangement for a data matrix as discussed in this presentation. The 'samples' for TOF-SIMS data will represent spectra, or the measured areas of peaks from a given spectrum. The 'variables' will be the selected peaks from the given spectra, or the individual data bins if the entire spectrum is read into the matrix.

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)**

A few general things to consider before beginning with PCA. First PCA assumes a linear relationship between variables. Second, PCA is scale dependent. This is an important point to note. This means that variables with high relative intensities and higher relative variances will be highlighted more in PCA than lower intensity variables. With TOF-SIMS data this often means that low mass, less chemically specific peaks, have higher loadings than high mass, chemically specific peaks, solely due to the differences in relative intensity of these peaks. Finally PCA looks at variance patterns in the data. It will determine the largest sources of variation within a data set, but it doesn't know anything about your data. So if the largest source of variation in the data is from noise, or a contaminant, PCA will highlight these differences first.

PCA data Pretreatment

- **No standards have been set for data pretreatment**
- **Some common trends include**
  - **normalizing the data (many different ways)**
  - **mean centering for TOF-SIMS spectra**
  - **Autoscaling for TOF-SIMS images**

As with many multivariate methods, it is common to pre-process the original data matrix in order to highlight the true chemical differences in the samples and not just differences in absolute intensity or differences from the mean. There are many different ways to pre-process a data matrix. Each of these data pretreatments can affect the results obtained from PCA and each method carries with it a set of assumptions. It is important to understand what these assumptions are and determine whether these assumptions are valid. Currently there are no guidelines as to how and when to use one data pretreatment versus another. It is hoped that time and effort will be spent on research that will shed light in this area.

Review of the literature using MVA for the interpretation of SIMS spectra

Problem	Data Pretreatment	MVA Technique	References
Polymer analysis			
Quantification of the composition of plasma polymerized thin films	P1+N2+S1	PLSR	[51,52]
Quantification of polymer molecular weight	P2+N3 or N5+S1	PCA	[53–56]
Quantification of polymer blend composition	P2+N3+S1	PCA	[57]
Discrimination between different polymers	P2+N2+S2+S1P3	PCA, NN	[58,59]
Characterization of multilayer polymer films	P1+N3+S1	PCA	[60]
Correlation of polymer surface chemistry with protein adsorption and cell growth	P1+N2+S1	PLSR	[61–63]
Detection and quantification of polymer additives on polymer surfaces	P2+N3+S1	PCA	[64,65]
Quantification of cross-linker density in polymer films	P1+N5+S1	PLSR	[66]
Protein adsorption			
Identification of adsorbed proteins	P2+N3+S1 or S4	PCA, LDA	[32,67–70]
Classification of adsorbed proteins	P3	NN	[71]
Detection of adsorbed protein conformation	P2+N3+S1	PCA	[72–74]
Quantification of protein adsorption on polymer surface	P2+S2+S1	PLSR	[75]
Detection of low amounts of adsorbed protein	P2+N3+S1	PCA	[76]
Quantification of multicomponent protein mixtures	P2+N3+S1	PLSR, SIMCA	[77–80]
Others			
Discrimination between solid-phase extraction stationary phases	P1+S2+S1	PCA, PLSR, NN	[81,82]
Analysis of SIMS spectra from alkanethiol SAMs	P1+N1+S1	PCA	[83,84]
Discrimination between bacterial samples	P2+N2+S1P3+N1+S3	PCAPLSR, LDA	[85,86]
Tracking assembly of a supported lipid monolayer	P1+N3+S1	PCA	[87]
Discrimination between different paint surfaces	P3+N1+S1	PCA, PLSR	[88]
Discrimination between different atmospheric aerosol particles	Not described	PCA	[14,16]
Interpretation of SIMS depth profiles	P2+N3	FA	[6,89]
Discrimination between different calcium phosphate phases	P1+N1+S1	PCA	[90]

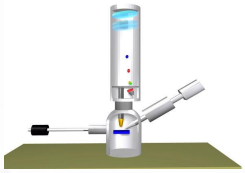
Peak selection abbreviations P1: All peaks in the mass spectrum above background were selected; P2: Only peaks characteristic to the system of interest were selected; P3: The spectra were binned to 1 amu bins and all bins were used.
Normalization abbreviations N1: The intensity of each peak in each spectrum was normalized to the total secondary ion counts for that spectrum; N2: The intensity of each peak in each spectrum was normalized to the intensity of the most intense peak in that spectrum; N3: The intensities of the selected peaks were normalized to the sum of the intensities of the selected peaks for each spectrum; N4: The intensity of each peak in each spectrum was normalized to the total secondary ion counts for that spectrum less contributions from hydrogen (H^+/H^-) and contaminants (e.g. PDMS); N5: The intensity of each peak in each spectrum was normalized to a selected peak in the spectrum.
Centering/scaling/transformation abbreviations S1: Mean-centered; S2: The logarithm (either natural logarithm or $\log_{10}$) of the data set was taken before analysis; S3: Each peak was scaled to the mean of the highest peak within a 25 *m/z* window of the peak; S4: Autoscaled.
MVA abbreviations FA: Factor analysis; LDA: Linear discriminant analysis; MVA: Multivariate analysis; NN: Neural networks; PCA: Principal component analysis; PLSR: Partial least squares regression; SIMCA: Soft independent modeling of class analogy.

This table illustrates the wide variety of data pretreatments that are being used throughout the literature of PCA of TOF-SIMS data. Though each of these methods may be valid, it is important to understand when and why they should be used.

The next few slides will give a brief overview of several data pretreatments and their associated assumptions.

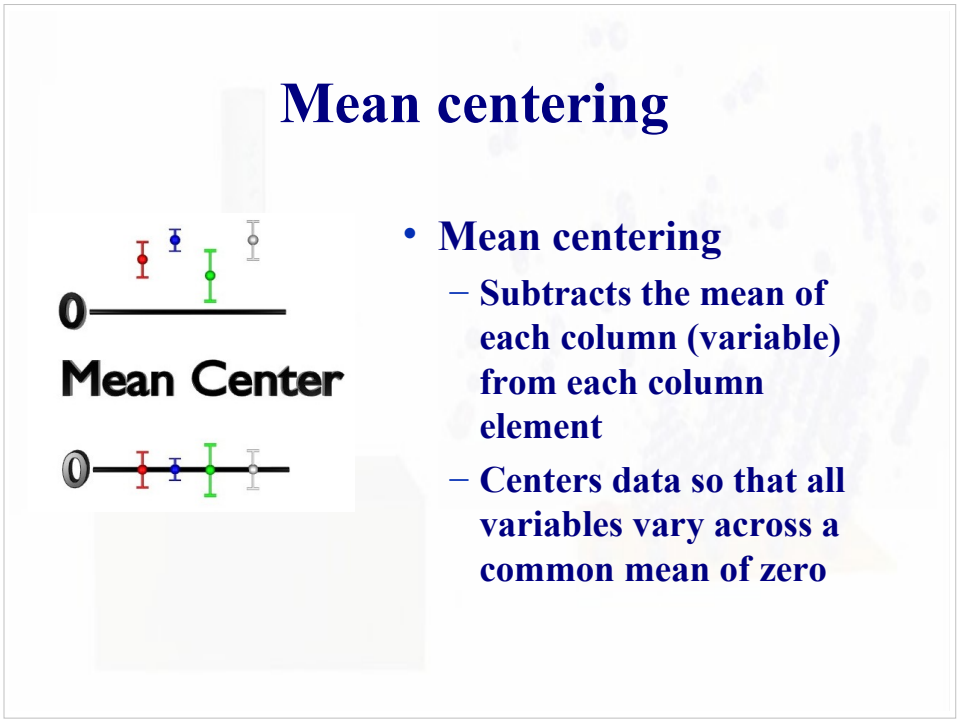
Normalization

- **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.



Another common data pretreatment is mean centering. Mean centering is done by subtracting the mean of each column (variable) from each column element. This centers the data so that all variables vary across a common mean of zero (as illustrated in the figure above). Mean centering makes it so PCA will more likely capture differences in the relative intensities of a given set of variables and not differences in the means of the variables.

Scaling

- **Scaling attempts to account for differences in variance scales between variables**
- **There is some debate about whether TOF-SIMS data should be scaled or not**
- **Autoscaling for SIMS images is common**
  - **Divides mean centered variables by their standard deviation**
    - **Results in variables with unit variance**
- **Other scaling methods have been proposed**
  - **No consensus on what is the "best" way**

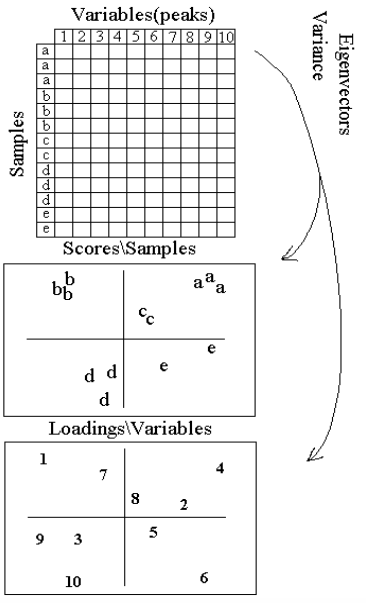
Data scaling is done to account for differences in the variance scales between variables Normalization can be considered a scaling operation since with normalization we are dividing or multiplying by some value to adjust for unwanted variances in the data. One common scaling method used with PCA is autoscaling.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

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**



Regardless of how the data is pretreated, the PCA algorithm is the same. PCA looks at the variance patterns within a data matrix. As mentioned previously PCA is an axis rotation that captures the directions of greatest variance within a data set. The rotation reduces the data dimensionality, allowing a large number of potentially correlated variables to be described by a few uncorrelated variables.

One nice feature of PCA is that the algorithm does not require external constraints or input from the user so it aides in removing user bias from that analysis.

PCA reconstructs the original matrix into a set of new matrices, the scores, the loadings and residuals. The scores describe the relationship between samples as described by the PCs. The loadings describe how the original variable related to the new PC axes. The residuals contain the left over variance from the data set and are assumed to describe noise in the data.

PCA Allows "quick" data summary

- **Scores**
  - **Tell relationship between samples**
    - **Are they similar or different?**
  - **Give an idea of the reproducibility of samples**
- **Loads**
  - **Show which variables are responsible for sample differences**
  - **Can help determine sample differences across entire peak set**

The scores and loadings plots generated from PCA allow simple, quick determination of the differences within a data set. By looking at a scores plot one can determine whether the samples in the data set are similar or different. The spread of samples in the scores plot can also give an idea of the reproducibility of the samples in the data set. If all the samples are tightly grouped together, this suggests the samples are spectrally similar. If the samples are spread across the scores plot it suggest there is variability across the sample set.

The loadings plot shows which variables are responsible for the differences seen between the samples. Loadings plots can typically highlight specific sets of peaks that correspond with a given set of samples. This often provides chemical insight into the sample differences.

PCA Mathematics

- **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 is based on looking at variance differences within a data set. Variance is a measure of the spread within the data. Covariance is a measure of the degree that two variables vary together. PCA is calculated from the covariance matrix of a data set. The covariance matrix is defined as shown above. This covariance matrix can be thought of as containing the covariance of each peak in the data set with all the other peaks in the data set.

PCA Methodology

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

$$X = T_1 P_1^T + E$$

Residual becomes new X matrix

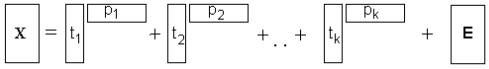
$$X = T_2 P_2^T + E$$

var(PC1)>var(PC2)>var(PC3)>...>var(PCk)

The PCA algorithm calculates each PC axis sequentially. In other words, PCA determines the first direction of greatest variance within the data. This first PC is then subtracted from the data set and the second PC is calculated from the remaining variance in the data. Each sequential PC captures the next greatest direction of variance. This is illustrated in the figure above. The original data matrix X is reconstructed into the product of the scores and the transpose of the loadings from the first PC and a residual matrix E that contains all the variance left after subtracting PC1 from the data matrix. This residual matrix E becomes the new matrix X from which the second PC is calculated. This process is continued until all the variance in the data is captured. For the resulting PCs the variance of PC1 > variance PC2 > variance PC3 >...> variance PCk.

PCA Mathematically

- **PCA decompositon**
  - $\mathbf{X = t_1 p_1^T + t_2 p_2^T + ... + t_k p_k^T + E}$



P_i (loadings) are the eigenvectors of the covariance matrix

$$\mathrm{cov}(X)\,p_i = \lambda_i p_i$$

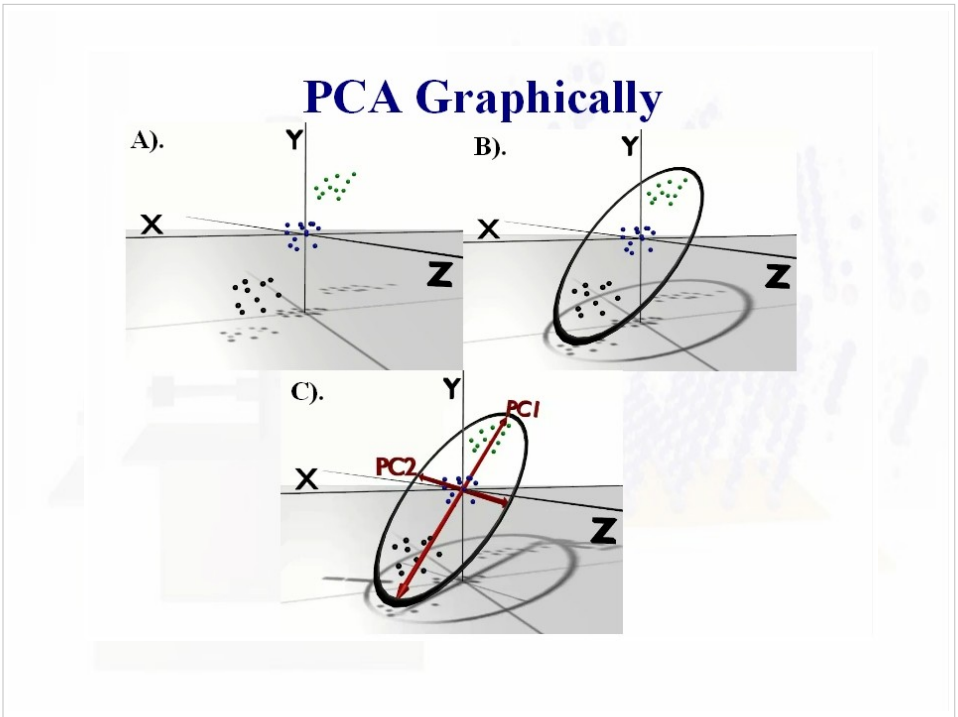
λ_i are the eigenvalues. They describe the amount of variance captured by each $t_i p_i$ pair (PC)

This figures illustrates the PCA decomposition with an equation and a simple graphical representation of the matrices and vectors involved. The loading values from PCA are the eigenvectors of the covariance matrix. The eigenvalues of the covariance matrix describe the amount of variance captured by each PC.

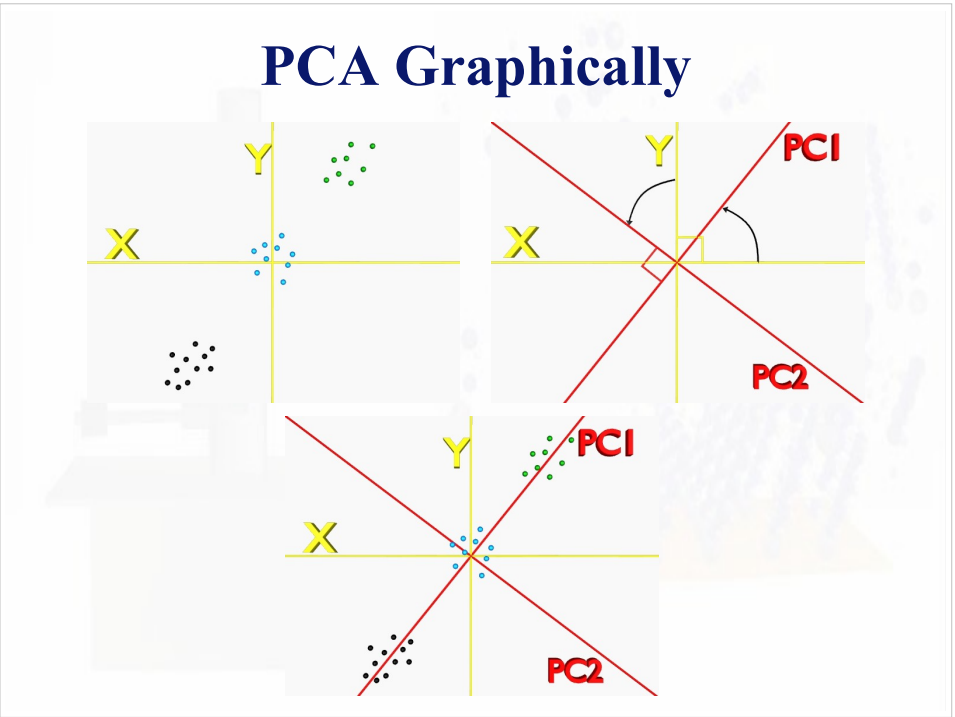
PCA Mathematically Continued

- **Original matrix is reconstructed into new set of matrices**
  - **$X = TP^T + E$**
    - **T = scores**
    - **P = loadings**
    - **E = residual (random noise)**
  - **One common way of doing PCA is by a singular value decomposition**
    - **$cov(X) = USV^T$**
      - **P = V Eigenvectors (loadings)**
      - **T = US (scores)**
      - **$S_{ii} = sqrt(\lambda_i)$ (λ_i = eigenvalues)**

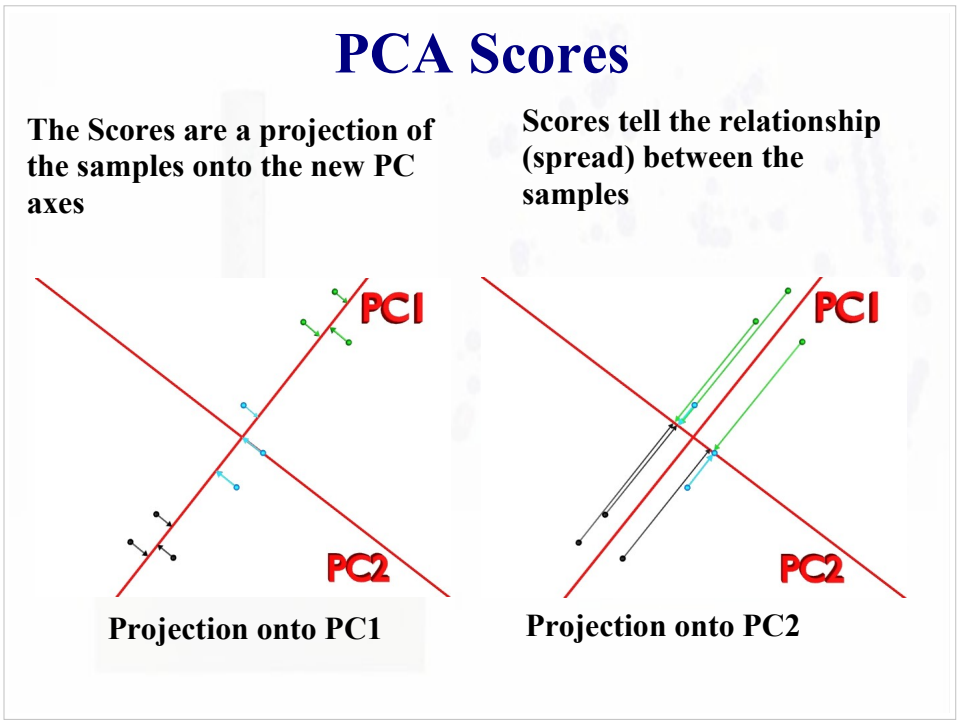
PCA is often calculated from the singular value decomposition of the variance covariance matrix. The general formulas for this are shown above along with the resulting definitions for the scores and loadings.



The figures above present one way of visualizing PCA graphically. Figure A) shows a set of data from 3 different samples plotted in 3 dimensional space. It can be seen that the data from these samples are clearly separated from each other within this data space. It can also be seen that there is some spread within the data points for each group. The ellipse in B) is drawn along the major plane of the data so that it encompasses the space containing all the data points. The major and minor axes of this ellipse are the principal axes of this data space and represent the first and second principal component axes..



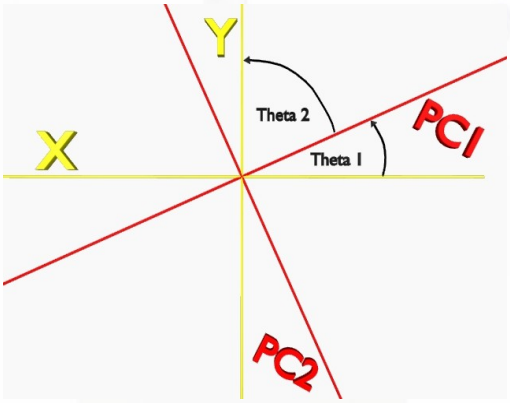
This figure is shown to further illustrate the concept that PCA is an axis rotation. A data set containing 3 sets of samples is plotted in 2 dimensional space (A). PCA would result in an axis rotation (B). PC1 would be defined in the greatest direction of variance within the data set. For this data set PC1 would follow along the spread between the 3 sample groups. PC2 would then be placed orthogonal with PC1 and would capture the spread within the sample groups. (C).



The score value for a given sample can be visualized graphically as the projection of the sample onto the given PC axis. This is illustrated above for a set of 3 different colored samples. The projection of the sample is defined by drawing a perpendicular line from the sample to the PC axes. For the data shown in the figures above it can be seen that the different colored samples will be separated from each other on PC1, while they will all overlap on PC2. This would suggest that PC1 is able to capture differences between the samples while PC2 is capturing some likely random variation in the data.

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 loading values are the direction cosines between the new PC axes and the original variables. Loadings are the weighting factors used for the original variables to get the new PC axes. Since Cos(90) = 0 and Cos(0) = 1 that means that a variable with a high loading is highly correlated with the given PC axis (the angle between the original variable and the PC axis is small). A high loading value means that that variable had a high influence on the separation of the samples on that PC axis.

I ran PCA now what do I have?

- **A set of PC scores and loadings**
  - **Each PC captures the greatest amount of variation in the given direction**
  - **%var PC1 > PC2 > PC3 > PC4 ... > PCn**
- **% variance tells relative amount of information captured by a given PC, but you need to check the scores and loadings to determine if the PC contains useful information**

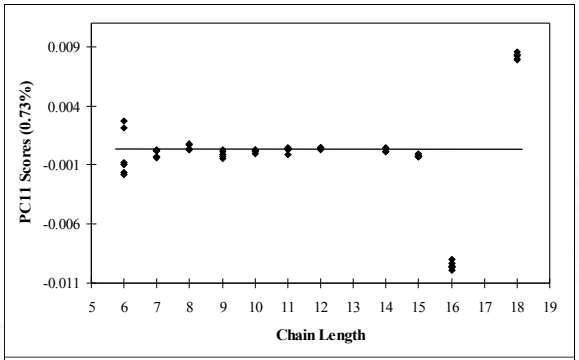
Once you have applied PCA to a given data set you will have a set of scores and loadings values. You now have a set of information that can potentially aide in interpreting your data set. The question now is where to start in digesting this information?

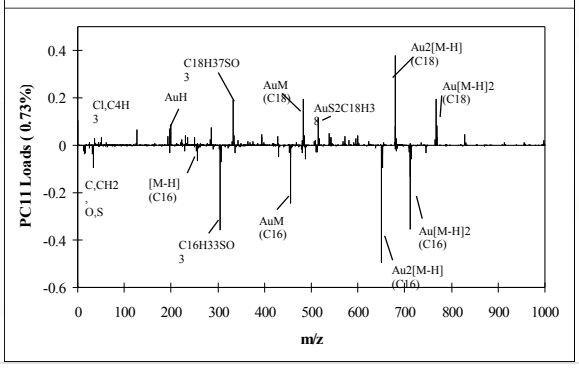
First remember that each PC captures a decreasing amount of variance from the data. The percent variance captures by a PC describes the relative amount of variance captured, but you will have to check the scores and loadings to determine if a PC contains useful information.

There are several criteria that are used in determining how many PCs to keep from PCA. The general rule is that you keep all PCs until the percent variance captured from one PC to the next does not change significantly. The number of PCs retained is most critical if you are using PCA to model a set of data that will then be used to project in new data for classification. For this type of PCA modeling retaining to many PCs can over fit the data, while not retaining enough PCs can make the model less robust. When using PCA for data exploration the number of PCs to keep depends on which PCs contain potentially useful information (i.e not just noise). This decision cannot always be made simply by looking at the percent variance captured by a given PC as will be illustrated on the next slide.

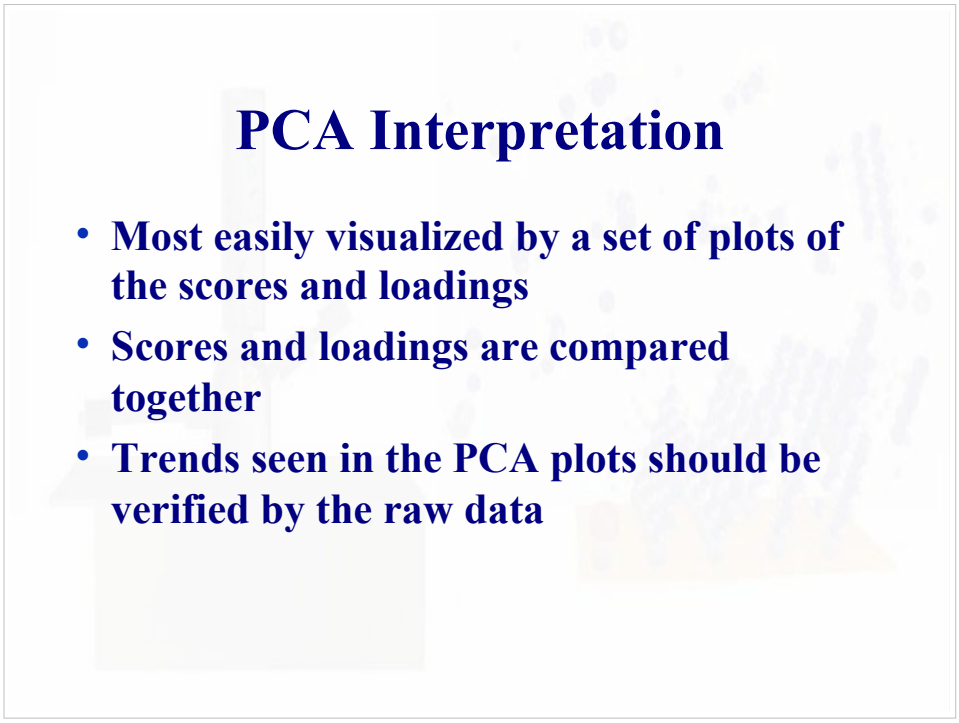
Quick Example %variance

- **Even PC11 contains useful information about the samples even though it only captures 0.73% of the variance in the data**
- **Separates C16 and C18 samples**





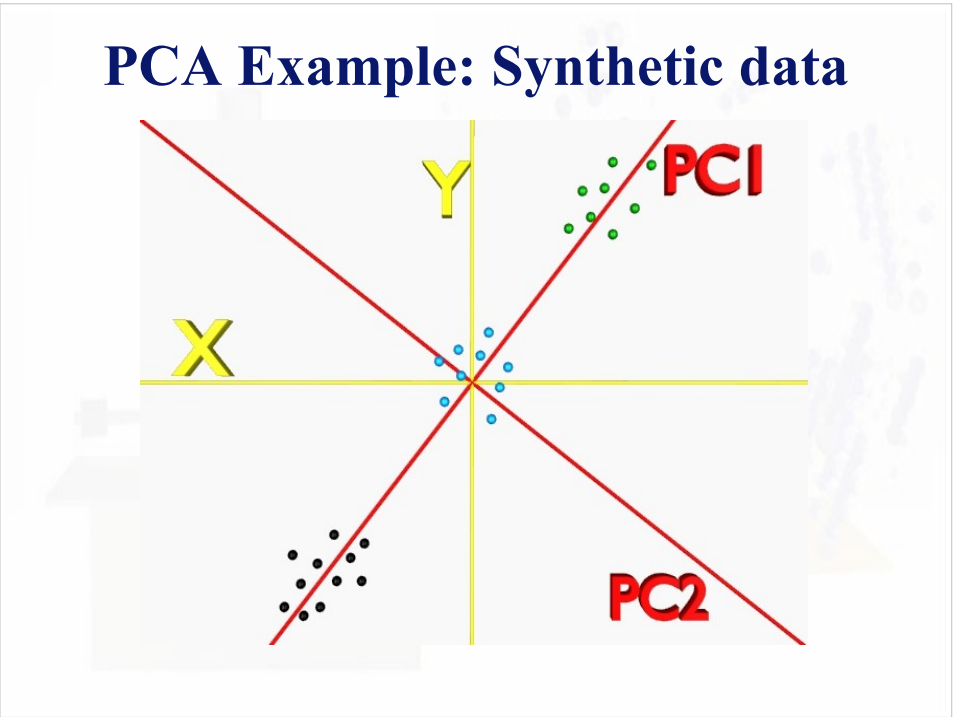
The slide above shows PC11 from a series of different chain length self-assembled monolayers on gold. Even though PC11 only captures 0.73% of the variance in the data set, it can be seen in the scores and loadings that PC11 clearly separates out the C16 and C18 monolayers (scores plot) and captures the unique molecular ion clusters for each thiol (loadings plot).



In many ways running PCA on a set of data is the simple part of the analysis. In most software packages it is as simple as pressing a few buttons. Once the program has finished running PCA the real work begins. Luckily the basics of PCA interpretation are fairly simple. The results from PCA can be interpreted by looking at the scores and loadings plots. The scores and loadings should always be looked at together since they contain complimentary information. The scores can show interesting trends between the samples, but without looking at the loadings it will not be known whether those trends make any sense or have any real meaning.

One important thing to note is that the trends seen in PCA should always be verified with the original data. This means that if you find that PCA separates out a set of samples and highlights a given set of peaks as having high influence on this separation (high loadings). You should go back to the original data and check the trends in the data for these peaks. The reason for this is that sometimes there are complex relationships going on within the scores and loadings plots, so that some peaks with high loadings will only correspond to a subset of samples in the scores plot. This will be illustrated further later in this discussion

The 'original data' refers to the data that was input into PCA. So if you normalized the data before starting PCA, you should go back to the normalized data to check the trends.



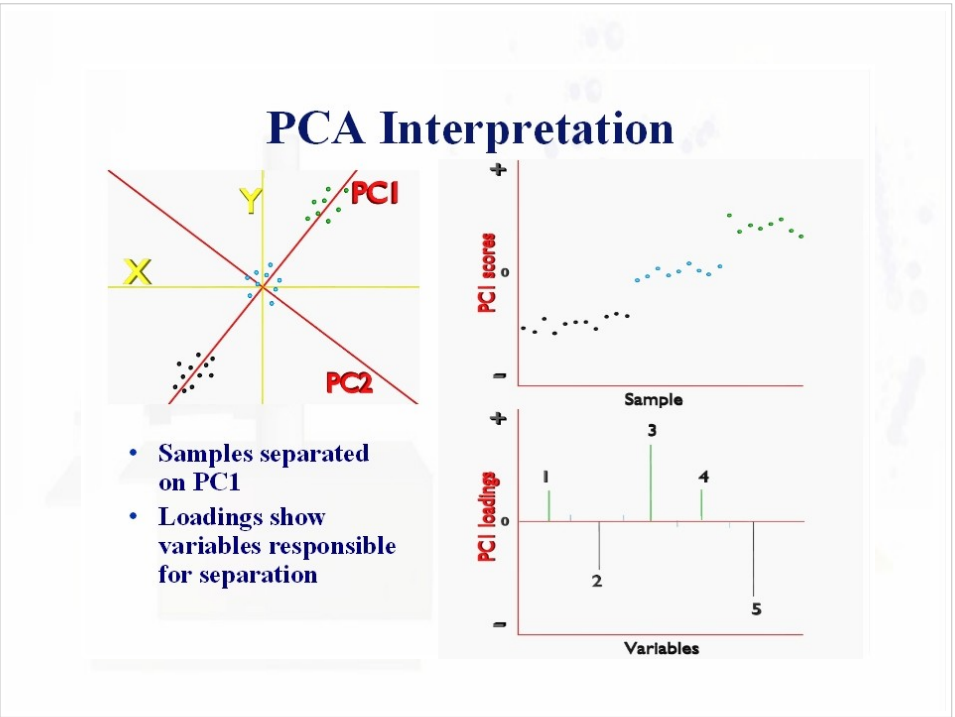
The following few slides will illustrate how to interpret PCA scores and loadings plots using results from a set of simulated data. This data set consists of 3 sets of samples (black, blue and green). The data from these samples are clearly separated into groups along the PC1 axes as seen in the figure above.

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

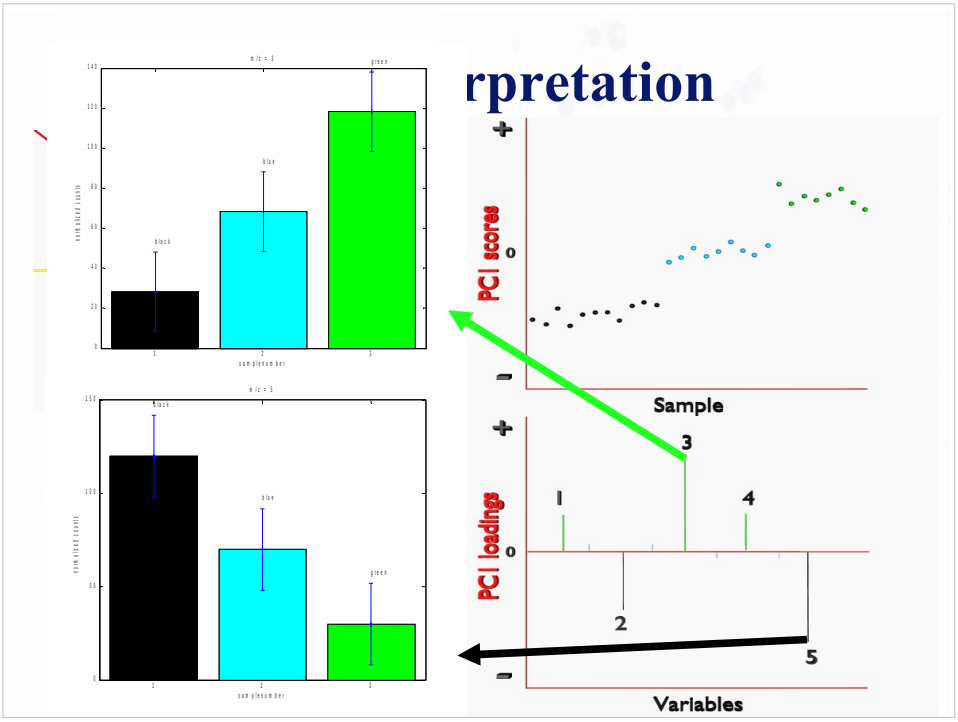
Samples with positive scores on a given PC axis are positively correlated with variables with positive loadings on the same PC axis. Samples with negative scores are positively correlated with variables with negative loadings. This means that, in general, samples with positive scores will have higher 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.

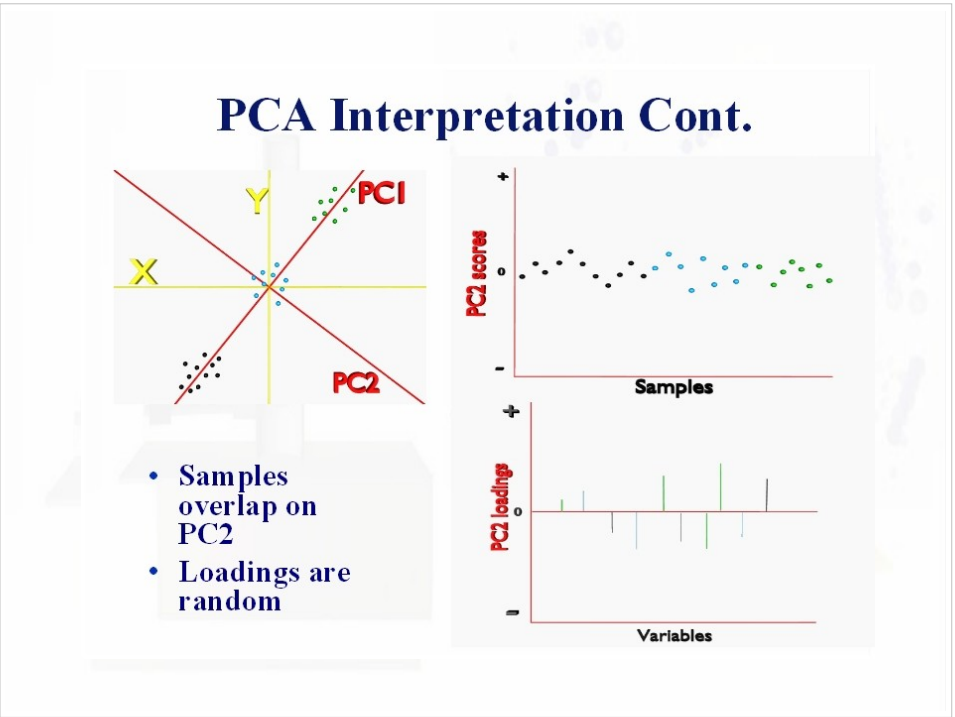
It is important to note that since PCA looks at differences in the relative intensity of variables, even if a variable is negatively correlated with a given set of samples, it does not mean that the value of that variable for those samples is necessarily zero. It just means that those samples have a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



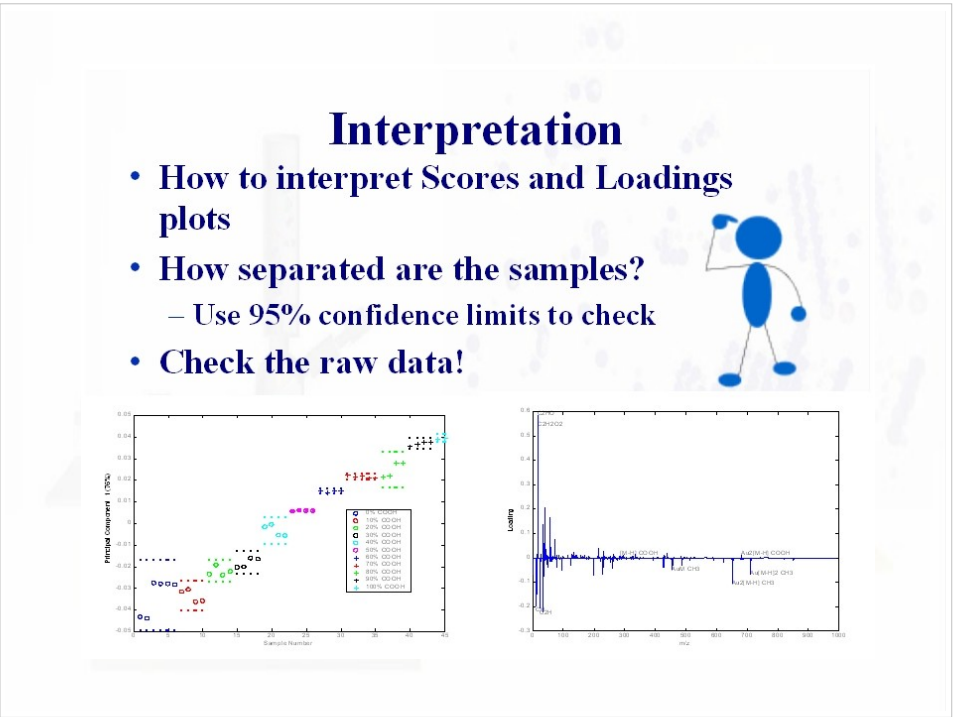
The figures on the right show the PC1 scores and loadings for this simulated data set. The scores are shown at the top of the figure and are plotted against the sample number. Plotting the scores in this way allows us to look at the PC1 scores values (the projection of the data points onto the PC1 axis). As seen in the scores plot the different colored samples are separated from each other along PC1 with the green samples having positive scores and the black samples having negative scores. The loadings plot shows that variables 1, 3 and 4 (green variables) have positive loadings, while variables 2 and 5 (black variables) have negative loadings. This means that variables 1, 3 and 4 correspond more with the green samples that have positive scores, and variables 2 and 5 correspond more with the black samples that have negative scores.



This figure shows the original normalized data for variables 3 and 5. Variable 3 has a high positive loading, meaning that it correspond with the green samples with positive scores. As would be expected the original data for variable 3 shows that the green samples have the highest relative intensity for this variable. Variable 5 has a high negative loading corresponding with the black samples that have negative scores. As would be expected the black samples show the highest relative intensity for variable 5.

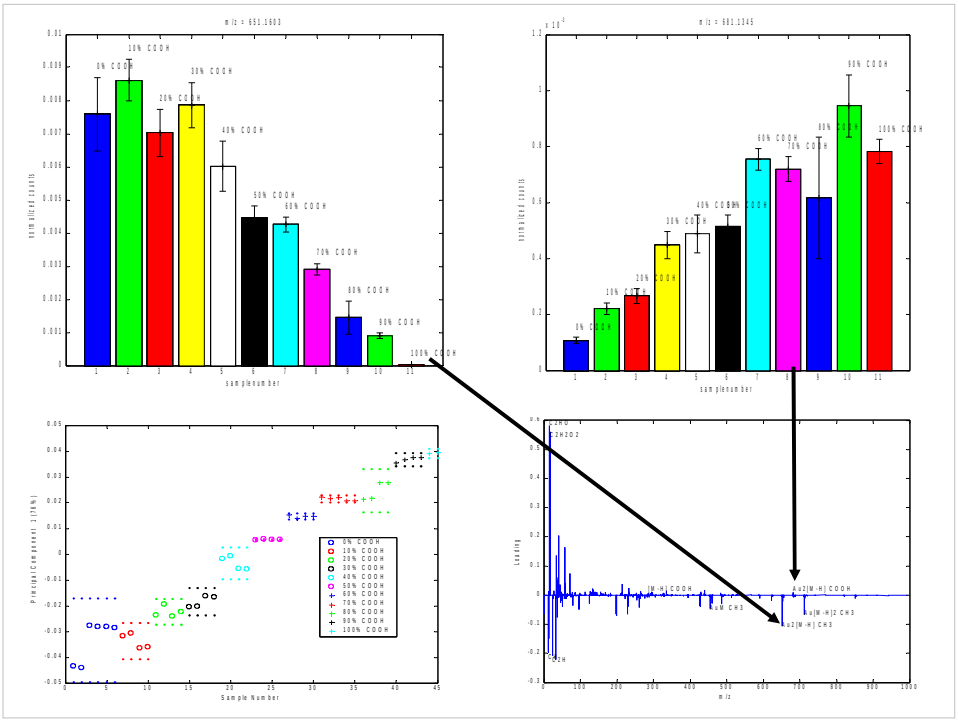


The figures on the right show the data projected onto the PC2 axis. As seen in the scores plot (top figure), the scores on PC2 overlap showing no clear separation of any samples. The loadings plot shows peaks of all colors showing both positive and negative loadings. PC2 therefore is capturing the scatter in the data within the data groups.



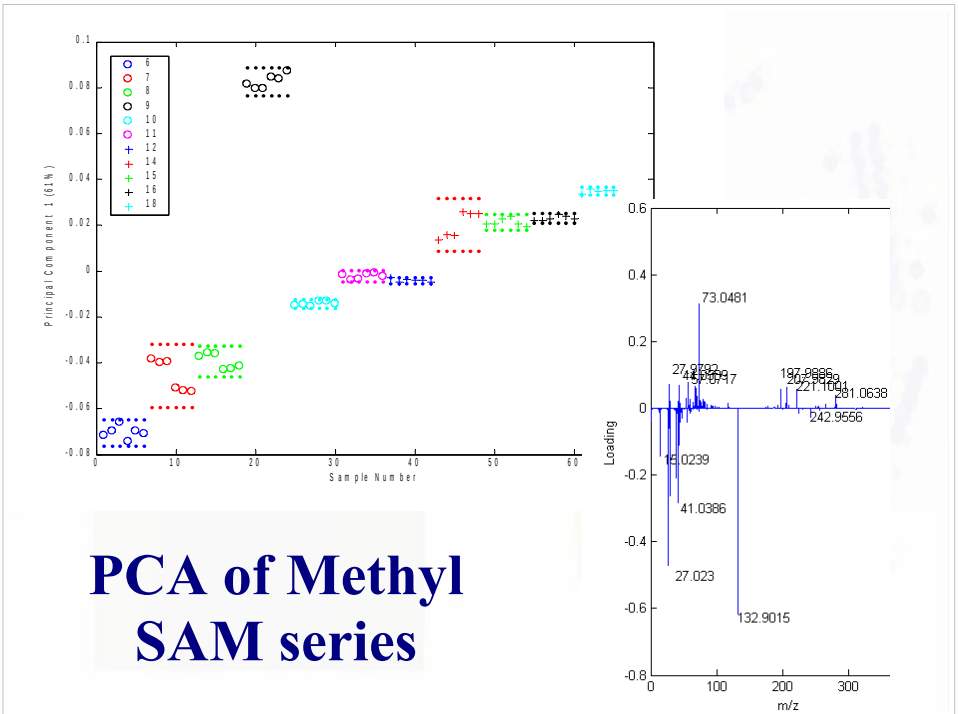
The previous slides dealt with an idealized set of samples from simulated data. So when you look at real data, how do you decide if a given set of samples is truly separated on a scores plot? One good way is to plot the 95% confidence limits for each group. Information on how to do this can be found in the paper by Wagner and Castner (Langmuir 2001, 17, 4649-4660).

As suggested before, it is always important to check the original data to verify the trends seen in the scores and loadings plots. The figures shown on this slide are the PC1 scores and loadings plots from TOF-SIMS data from a set of mixed monolayers of HS(CH2)15COOH and HS(CH2)15CH3. The scores plot is shown as the PC1 scores versus the sample number, where the samples are organized in order of increasing percent COOH thiol in solution. It can be seen that PCA is able to separate out most of the different mixed monolayer surfaces. The loadings plot is dominated by the low mass peaks, but it is noted that the high mass cluster ions show the expected trends . Peaks from the COOH thiol have positive loadings corresponding with samples with positive scores (higher concentrations of COOH thiol). Peaks from the CH3 thiol have negative loadings, corresponding with samples with negative scores (higher concentrations of CH3 thiol).



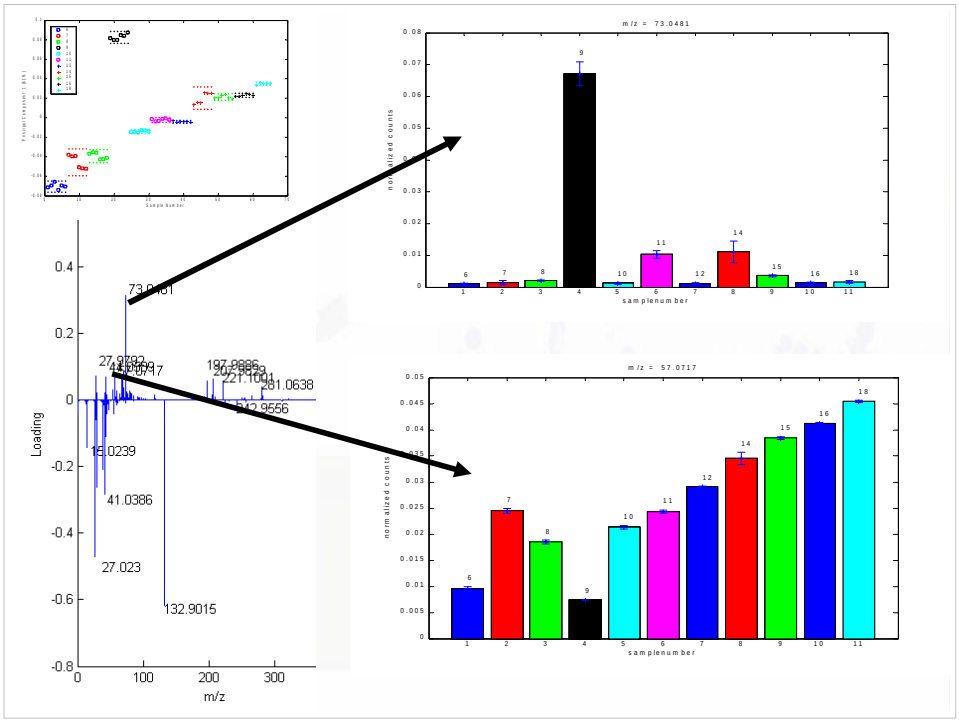
The bar charts on top of this slide show the original, normalized, data for one of the COOH thiol peaks (positive loading right chart), and one of the CH3 thiol peaks (negative loading left chart). As seen in the charts the original data follows the trend that would be expected based on the appearance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

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



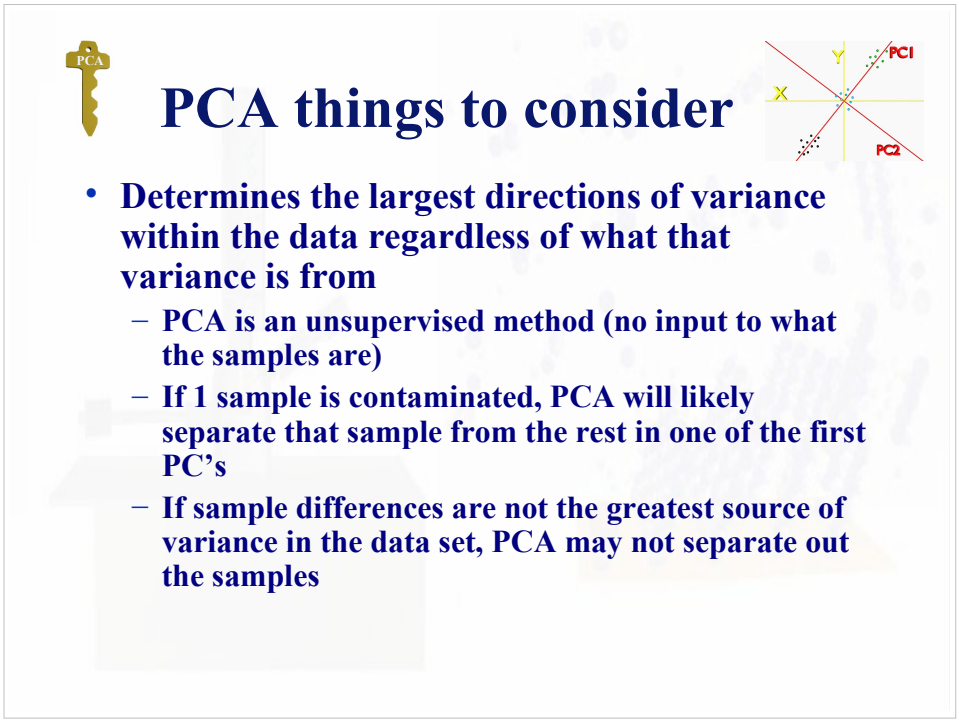
The slide above shows the PC1 scores and loadings plots from a set of methyl terminated self-assembled monolayers with varying chain lengths (from C6 to C18). It can be seen that the PC1 score values increase with increasing chain length with one clear outlier. The C9 thiol samples are seen to have significantly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

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



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

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



So to begin to summarize this general overview of PCA, it is important to remember that PCA looks for the largest directions of variance within a data set. PCA does not know anything about your data. It is an unsupervised method. This means that if one of the samples input into a data set is contaminated, PCA will likely separate out that sample from the rest in one of the first PCs.

This means that you must design your experiments carefully because if the differences in your sample chemistries are not the largest sources of variance in the data, PCA will not separate your samples as expected. It is ideal if you can design your experiments so that you only allow 1 variable to change across the sample set. Then you can have more confidence that the differences highlighted by PCA are due to changes in that variable and not random variance within the data.

PCA things to consider

- **The decision to use PCA should be part of your experimental plan, not just an afterthought after collecting your data**
- **You should understand what you are doing and how it works**
- **You should understand the assumptions made for running PCA**
- **You should check your results with the raw data**

To use PCA properly, PCA should be part of your experimental plan from the beginning and not just something you try when all else fails. As mentioned on the previous slide, using a well designed experiment can make the difference between being able to understand your PCA results and simply ending up with a series of plots that show no clear trends.

Before using PCA you should understand what PCA does, how it works, and why you are using it. The more you understand, the better your ability will be to plan, execute and interpret your results will be.

Learn and understand the assumptions being made when preprocessing your data and choosing various options when running PCA.

Always go back to your original data matrix and verify the trends you seen in the scores and loadings. Even some well designed systems can give complex results showing multiple sources of variance for a given PC.

For More Information about PCA See ...

- **JE Jackson (1980) Principal Components and Factor Analysis: Part I - Principal Components. Journal of Quality Technology 12:201-213**
- **JE Jackson (1991) A Users's Guide to Principal Components. John Wiley & Sons, Inc.,New York**
- **S Wold, K Esbensen and P Geladi (1987) Principal Components Analysis. Chemometrics and Intelligent Laboratory Systems 2:37-52**

The references here are a great starting place for learning about PCA. They present thorough, approachable introductions to PCA. Also see the reference link on the homepage of this website.