Chemometrics I: Principal Components and Exploratory Data Analysis

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Introduction
- Preprocessing-Scaling and Centering
- PCA
  - Graphically
  - Mathematically
  - Scores and Loadings
- Examples
  - Wine, Synthetic, Octene, Rain, Arch
- Q and T^2
- Application to new data
- Determining the number of components
- Exploring PCA Models

EIGENVECTOR RESEARCH INCORPORATED

Course Materials

- These slides
- PLS_Toolbox or Solo 6.7 or later
- Data sets
  - From DEMS folder (distributed with software)
    - wine.mat, arch.mat, nir_data.mat
  - From EVRIHW folder (additional data sets)
    - octene.mat, Rain.mat,

EIGENVECTOR RESEARCH INCORPORATED

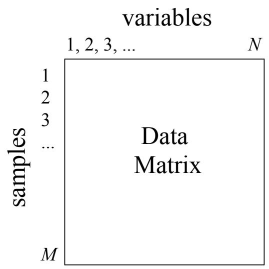
Nomenclature and Conventions

- Data is arranged in matrices where
- *rows* correspond to *samples* or *observations*, and *columns* correspond to *variables*
- Notation:
  - M = number of samples or observations
  - N = number of variables
  - K = number of Principal Components (PCs) or factors
  - $\mathbf{T}$ = scores matrix, $\mathbf{t}_1$, $\mathbf{t}_2$, ..., $\mathbf{t}_K$ score vectors
  - $\mathbf{P}$ = loadings matrix, $\mathbf{p}_1$, $\mathbf{p}_2$, ..., $\mathbf{p}_K$ loadings vectors

EIGENVECTOR RESEARCH INCORPORATED

Variables and Samples

- Examples of variables:
  - absorbance at each l
  - ion current at each m/e
  - pressure, temperature, flow
  - chromatographic peak area
- Examples of samples:
  - samples taken to lab
  - data samples at time points
  - data from specific batches
  - etc....



EIGENVECTOR RESEARCH INCORPORATED

Data Transformation

- PCA assumes that relationships between variables are linear
- If possible, non-linear data should be converted to a linear form
- Examples:
  - reaction rates proportional to $e^{-1/T}$, transform with log
  - pipe flow proportional to $P^{4/7}$ (turbulent flow)

EIGENVECTOR RESEARCH INCORPORATED

Mean Centering

- PCA is scale dependent, numerically larger variables appear more important
- Often we are most interested in how the data *varies* around the mean
  - not centering can be considered a force fit through 0
- *Mean centering* is done by subtracting the mean off each column, thus forming a matrix where each column has mean of zero
  - `[mcx,mx] = mncn(x);`

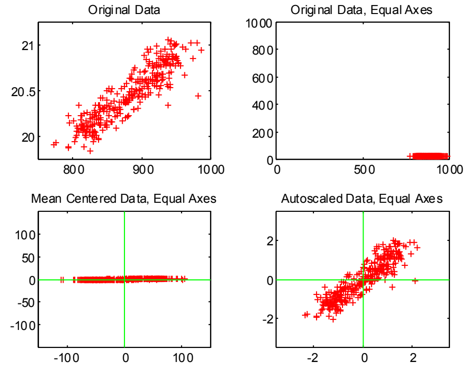
EIGENVECTOR RESEARCH INCORPORATED

Variance Scaling

- PCA is scale dependent, variance is associated with importance
- This may or may not be true
- In spectra, variance proportional to importance (probably)
- If variables have different units, variance doesn't = importance
- *Autoscaling* - divide each (mean centered) variable by its standard deviation, result is variables with unit variance
  - autoscaling implies both mean centering and scaling to unit variance
  - `[ax,mx,stdx] = auto(x);`
- Other scaling - may want to use *a priori* information, such as noise level in variables

EIGENVECTOR RESEARCH INCORPORATED

Centering & Scaling Example



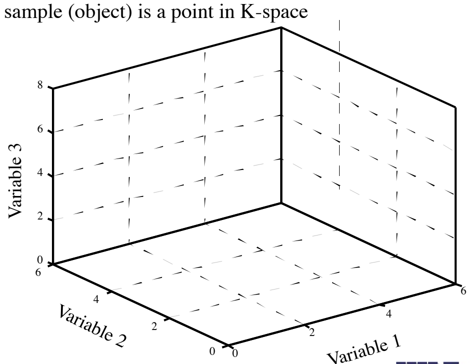
example with SFCM data in `plsdata`

Block Scaling

- With blocks of different variables, may want each block to have the same variance
  - Example: data set with NIR spectra and GC data and a collection of engineering variables, T, pH, P, Q, etc.
  - `gscale`
- Variables within blocks may be autoscaled or just mean centered
- Determine factor to multiply each block by so that total sum of squares (variance) is the same for each block

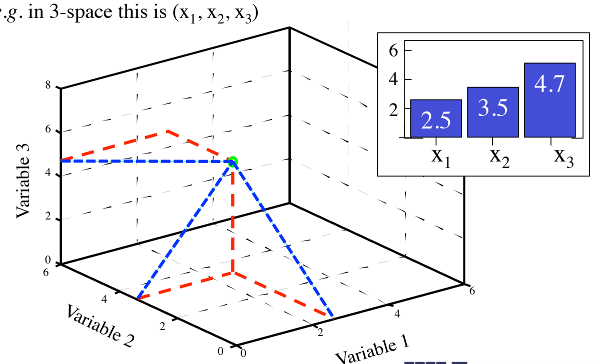
Principle of Projections

- K-space has K dimensions where each variable, or measurement on an object, is a coordinate axis
- A sample (object) is a point in K-space



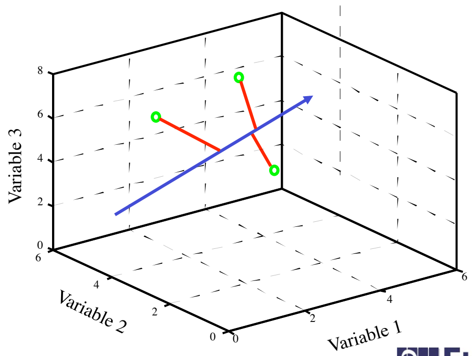
Projection in K-Space

- The projection of an object onto the K-space yields the coordinates of the object in that space
- *e.g.* in 3-space this is (x_1, x_2, x_3)



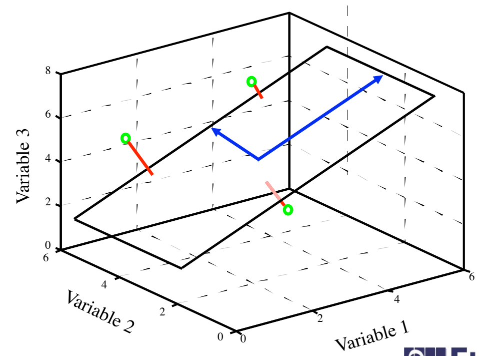
Projection onto a Vector

- Projection lines are perpendicular to the vector

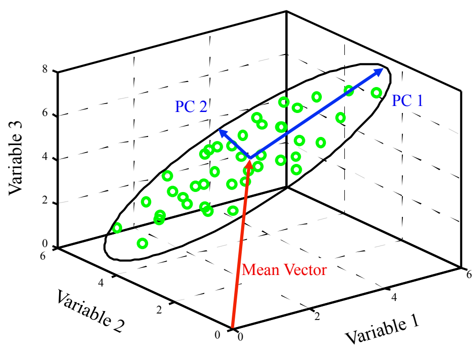


Projection onto a Plane

- Projection lines are perpendicular to the plane

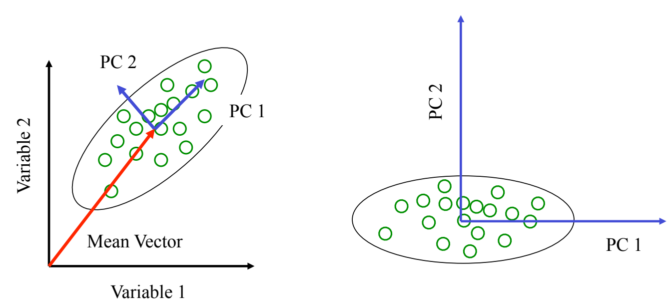


PCA

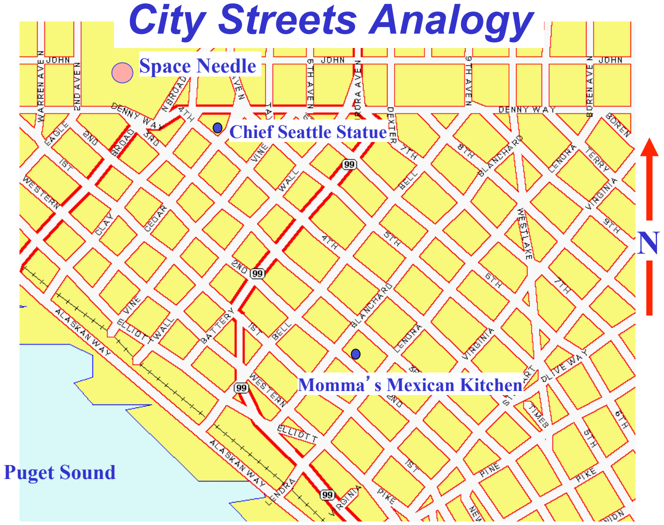


PCA

- Geometry for 2 variables



City Streets Analogy



PCA Math 1 of 3

For a data matrix **X** with *m* samples and *n* variables (generally assumed to be mean centered and properly scaled), the PCA decomposition is:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \ldots + \mathbf{t}_K\mathbf{p}_K^{\mathrm{T}} + \ldots + \mathbf{t}_R\mathbf{p}_R^{\mathrm{T}}$$

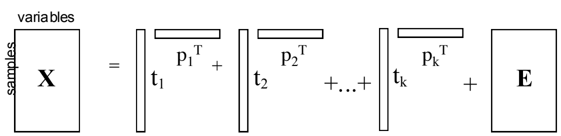
Where $R \leq \min(M,N)$, and the $\mathbf{t}_k\mathbf{p}_k^{\mathrm{T}}$ pairs are ordered by the amount of variance captured.

Generally, the model is truncated, leaving some small amount of variance in a residual matrix:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \ldots + \mathbf{t}_K\mathbf{p}_K^{\mathrm{T}} + \mathbf{E} = \mathbf{T}_K\mathbf{P}_K^{\mathrm{T}} + \mathbf{E}$$

EIGENVECTOR RESEARCH INCORPORATED

PCA Math 2 of 3



The $\mathbf{p}_k$ are eigenvectors of the covariance matrix of **X**

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

$$\mathrm{cov}(\mathbf{X})\mathbf{p}_k = \lambda_k\mathbf{p}_k$$

and the λ_i are eigenvalues.

Amount of variance captured by $\mathbf{t}_k\mathbf{p}_k^{\mathrm{T}}$ proportional to λ_k.

EIGENVECTOR RESEARCH INCORPORATED

PCA Math 3 of 3

- What is PCA doing mathematically?
- For a data set **X**, propose that $\mathbf{t} = \mathbf{X}\mathbf{p}$
  - *i.e.* **X** projected onto factor **p** yields **t**
  - **X** is usually centered and scaled
  - $\max\{\mathbf{t}^{\mathrm{T}}\mathbf{t} \mid \mathbf{p}^{\mathrm{T}}\mathbf{p}=1\} = \max\{\mathbf{p}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p} \mid \mathbf{p}^{\mathrm{T}}\mathbf{p}=1\}$
  - $L(\mathbf{p}) = \mathbf{p}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p} - \lambda(\mathbf{p}^{\mathrm{T}}\mathbf{p}-1)$ **:** take d/d**p** and set to 0
  - $\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p} = \lambda\mathbf{p}$
- Shows that the solution is an eigenvalue/ eigenvector problem

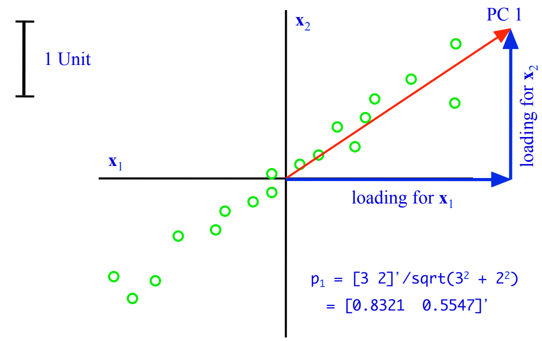
EIGENVECTOR RESEARCH INCORPORATED

Properties of PCA

- $\mathbf{t}_k$,$\mathbf{p}_k$ ordered by amount of *variance captured*
- $\mathbf{t}_k$ or *scores* form an orthogonal set $\mathbf{T}_K$ which describe relationship between *samples*
- $\mathbf{p}_k$ or *loadings* form an orthonormal set $\mathbf{P}_K$ which describe relationship between *variables*
- $k = 1,\cdots,K$ are the number of factors
- scores and loadings plots are interpreted in pairs
  - *e.g.* plot $\mathbf{t}_\mathrm{k}$ vs sample number and $\mathbf{p}_\mathrm{i}$ vs variable number
- it is useful to plot $\mathbf{t}_{k+1}$ vs. $\mathbf{t}_k$ and $\mathbf{p}_{k+1}$ vs. $\mathbf{p}_k$

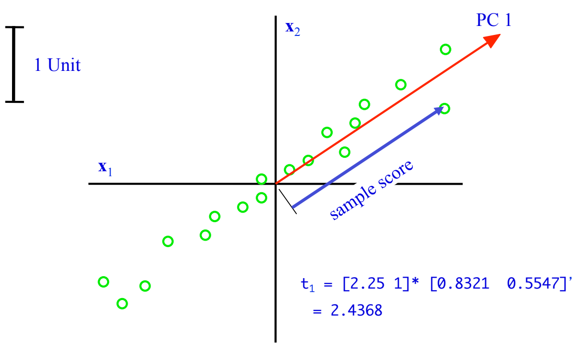
EIGENVECTOR RESEARCH INCORPORATED

Variable Loadings, p_i



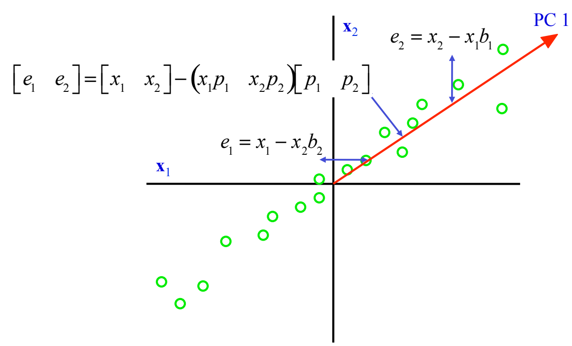
EIGENVECTOR RESEARCH INCORPORATED

Sample Scores, t_i



EIGENVECTOR RESEARCH INCORPORATED

Minimization Criterion



EIGENVECTOR RESEARCH INCORPORATED

Some Mathematical Relationships

- $\mathbf{P}$ orthonormal, so $\mathbf{PP}^{\mathrm{T}} = \mathbf{I}$, $\mathbf{P}^{\mathrm{T}} = \mathbf{P}^{-1}$, and $\mathbf{P}_K^{\mathrm{T}}\mathbf{P}_K = \mathbf{I}_K$
- Projection of $\mathbf{X}$ onto $\mathbf{P}_K$ gives the scores: $\mathbf{T}_K = \mathbf{XP}_K$
- Projection of $\mathbf{X}$ into PCA model, $\hat{\mathbf{X}}$, is equal to the scores times the loadings: $\hat{\mathbf{X}} = \mathbf{T}_K\mathbf{P}_K^{\mathrm{T}} = (-\mathbf{T}_K)(-\mathbf{P}_K^{\mathrm{T}})$
- Residual $\mathbf{E}$ is the difference between $\mathbf{X}$ and $\hat{\mathbf{X}}$, thus:

$$\mathbf{E} = \mathbf{X}-\hat{\mathbf{X}} = \mathbf{X}-\mathbf{T}_K\mathbf{P}_K^{\mathrm{T}} = \mathbf{X}-\mathbf{XP}_K\mathbf{P}_K^{\mathrm{T}} = \mathbf{X}\left(\mathbf{I}-\mathbf{P}_K\mathbf{P}_K^{\mathrm{T}}\right)$$

- PCA: $\mathbf{X} = \mathbf{TP}^{\mathrm{T}} = \mathbf{T}_K\mathbf{P}_K^{\mathrm{T}} + \mathbf{E}$
- SVD: $\mathbf{X} = \mathbf{USV}^{\mathrm{T}}$
  - $\mathbf{T} = \mathbf{US}$
  - $\mathbf{P} = \mathbf{V}$
  - $\mathbf{S}_{kk} = \sqrt{(M-1)\lambda_k}$

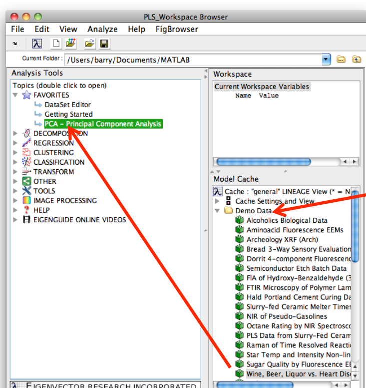
EIGENVECTOR RESEARCH INCORPORATED

Example: Wine Data

- Examine the relationship between (variables)
  - annual consumption of wine, beer, and liquor (gal/yr),
  - life expectancy (years), and
  - heart disease rate (cases/100,000)
- For 10 different countries (samples)
  - France, Italy, Switzerland, Australia, Britain, USA, Russia, Czech Republic, Japan, and Mexico
- Data from:
  Newsweek, **127**(4), 52, 1/22/1996

EIGENVECTOR RESEARCH INCORPORATED

Start from Workspace Browser in PLS_Toolbox or Solo



Browser window always open in Solo or execute
`>> browse`
in MATLAB/PLS_Toolbox

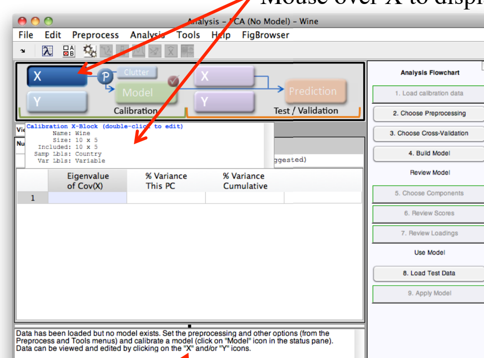
Expand "Demo Data" folder in Model Cache window

Drag "Wine, Beer...." data onto PCA in Analysis Tools window

EIGENVECTOR RESEARCH INCORPORATED

Data: loaded but not analyzed

Mouse over X to display data info

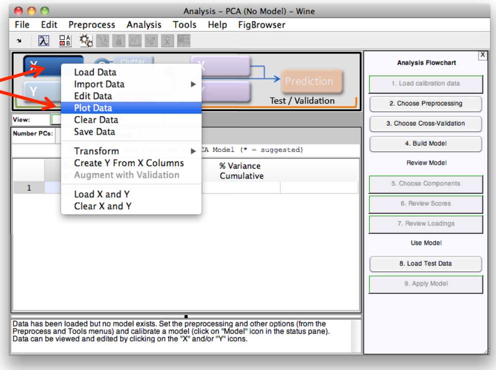


Status window after load

EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data!

Right-click or shift-click X to bring up menu, select “Plot Data”



EIGENVECTOR RESEARCH INCORPORATED

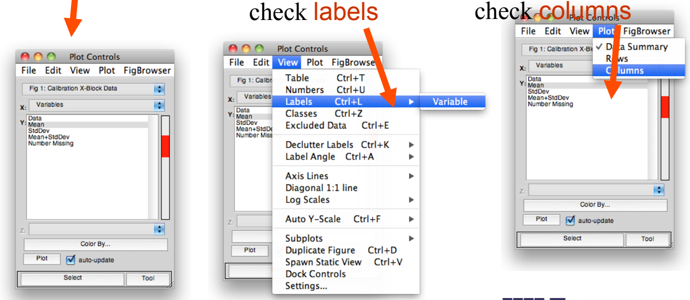
Plot Your Data

1 Plot control default can look at summary stats

The Plot control generates plots in MATLAB figure windows

2 under view menu check labels

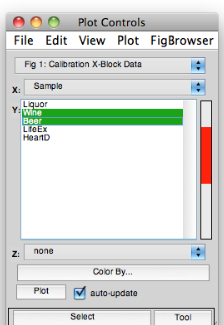
3 under plot menu check columns

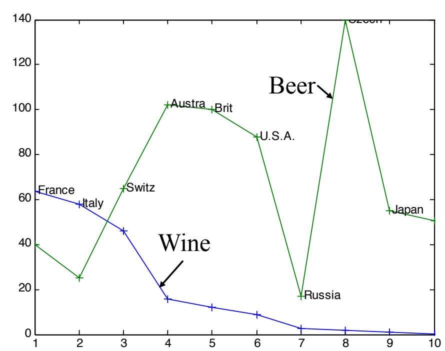


EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data

samples ordered by wine consumption



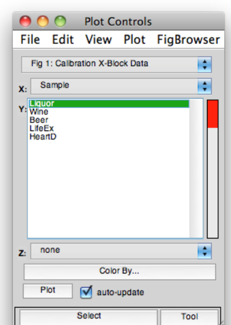


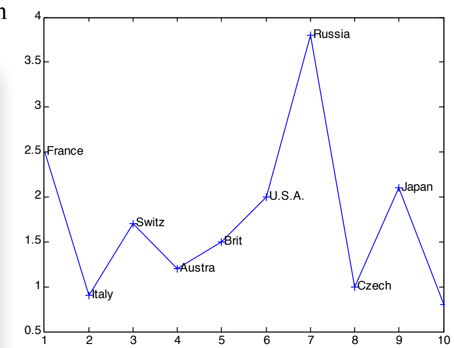
use shift key to select multiple columns

EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data

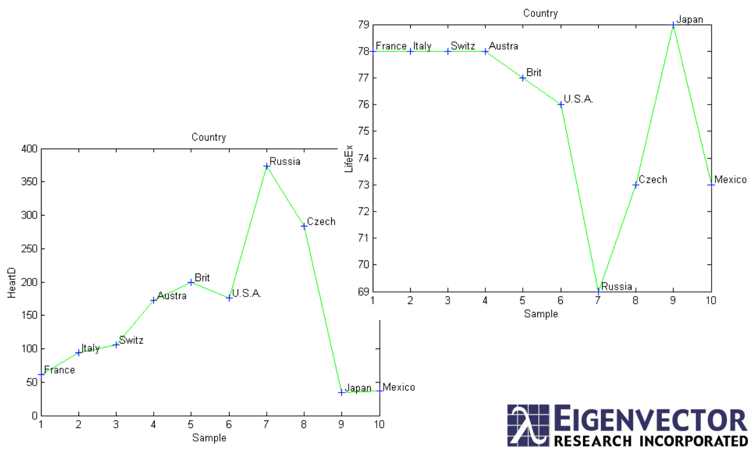
scale is ~1-2 orders of magnitude smaller than for Beer and Wine





EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data

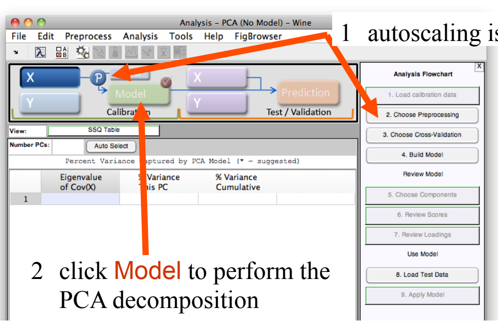


Plot Your Data Summary

- Wine consumption
  - France, Italy, Switz high
  - Rus, Czech, Jap, Mex low
- Beer consumption
  - Czech high
  - Italy, Russia low
- Liquor consumption
  - Russia high
  - Italy, Czech, Mex low
- Life Expectancy
  - Japan high
  - Russia low
- Heart Disease Rate
  - Russia high
  - Japan, Mexico low
- Some trends are apparent

How should we scale the data?

- Variables are in different units (apples and oranges): suggests autoscaling
- Variable standard deviations are of different magnitudes: suggests autoscaling



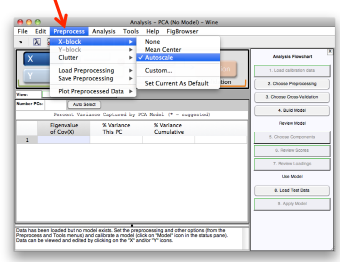
1 autoscaling is the default

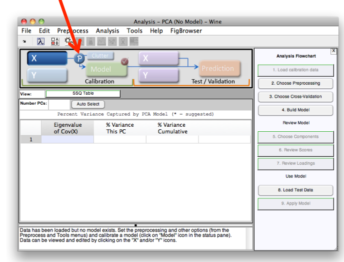
2 click Model to perform the PCA decomposition

Can Change Preprocessing...

from Preprocess menu

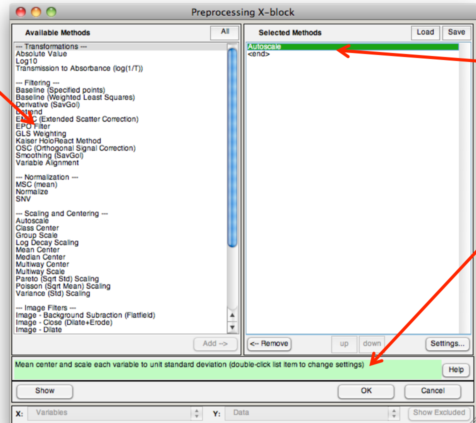
or click P icon for all preprocessing options





Preprocessing Window

choose from available methods



order selected methods

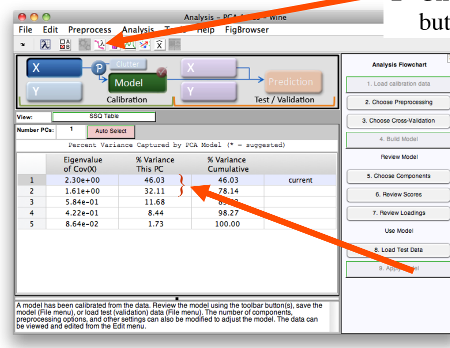
info on highlighted method

Do the PCA Decomposition

1 After the Model button:

- variance captured table: eigenvalues and % variance explained for each PC.

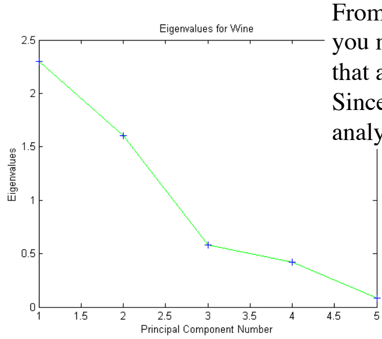
2 Click Plot Eigenvalues button to plot the eigenvalues



for autoscaled data: PCs w/ eigenvalues > 1 capture more variance than any single variable

Eigenvalue Plot

Plot the eigenvalues vs. PC.



From this and other considerations you may choose the number of PCs that are significant.
Since we're doing exploratory data analysis it doesn't really matter.

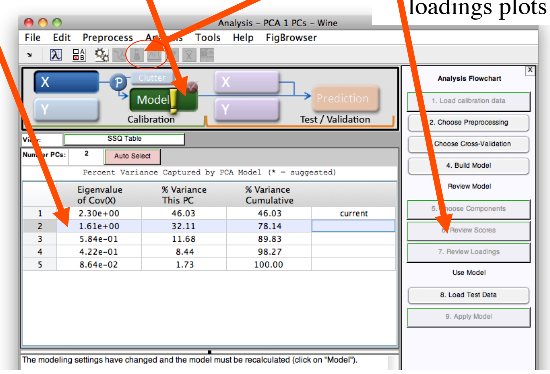
Perhaps 2 (or 4)? (later we'll show you cross-validation which suggests 1 in this case)

Choose Number of PCs

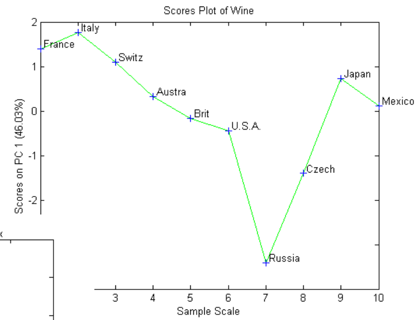
1 Highlight the second line to select 2 PCs

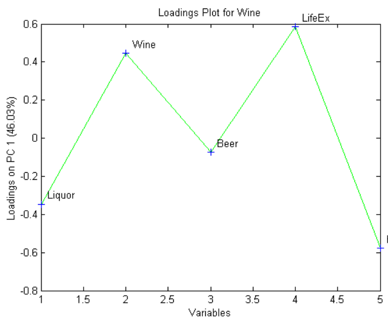
2 Click the Model! button to construct a 2 PC model

3 Click the scores button to make scores plots, loads button to for loadings plots



Scores and Loads on PC 1



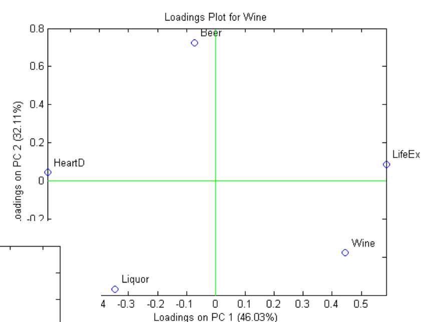


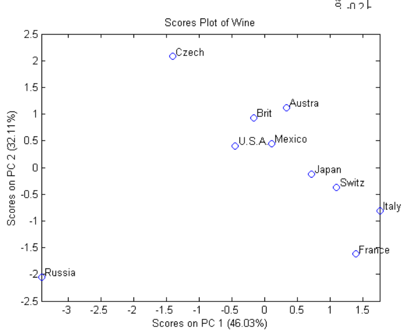
46%

PC 1

- Wine and Life Expectancy are correlated
- Heart Disease Rate and Liquor Consumption are correlated
- Heart Disease Rate and Liquor Consumption are anti-correlated with Wine and Life Expectancy
- Russia is Low on PC 1
  - but this is only 46% of the story!
- So let's look at PC 2 vs 1 ...

Scores and Loads on PC 2 vs. 1



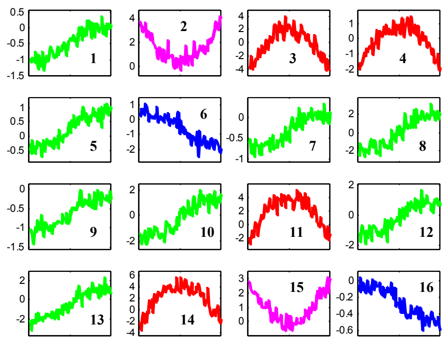


78%

PC 2 vs. 1

- HeartD and Beer: Orthogonal
- Russia is the most unusual, why?
  - tends to be high in Liquor and HeartD and low in Beer and LifeEx
- Trend from France to Czech, why?
  - France relatively high in wine and low in Beer, and HeartD
  - Czech relatively high in Beer and HeartD, and low in Wine

How many PC's to model this data?



EIGENVECTOR RESEARCH INCORPORATED

Variance Captured

Percent Variance Captured by PCA Model

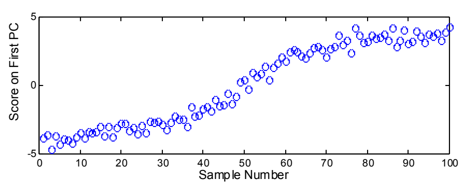
Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	8.79e+00	54.96	54.96
2	5.29e+00	33.05	88.01
3	2.49e-01	1.56	89.57
4	2.17e-01	1.35	90.92
5	1.80e-01	1.12	92.05
6	1.66e-01	1.04	93.08
7	1.51e-01	0.94	94.03
8	1.41e-01	0.88	94.91
9	1.33e-01	0.83	95.74
10	1.22e-01	0.76	96.51
11	1.19e-01	0.74	97.25
12	1.09e-01	0.68	97.93
13	1.03e-01	0.65	98.58
14	8.52e-02	0.53	99.11
15	7.36e-02	0.46	99.57

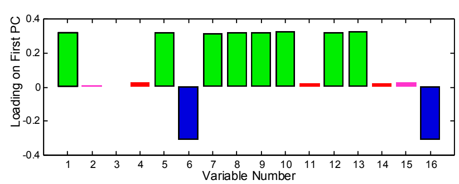
Which trend does PC 1 capture?

Which trend does PC 2 capture?

EIGENVECTOR RESEARCH INCORPORATED

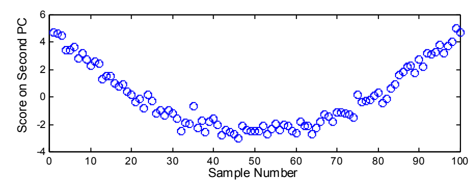
PC 1: Scores and Loadings

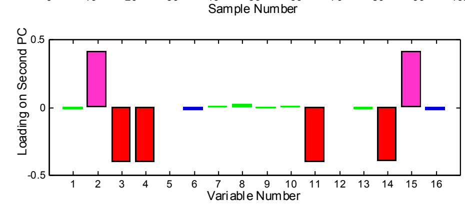




EIGENVECTOR RESEARCH INCORPORATED

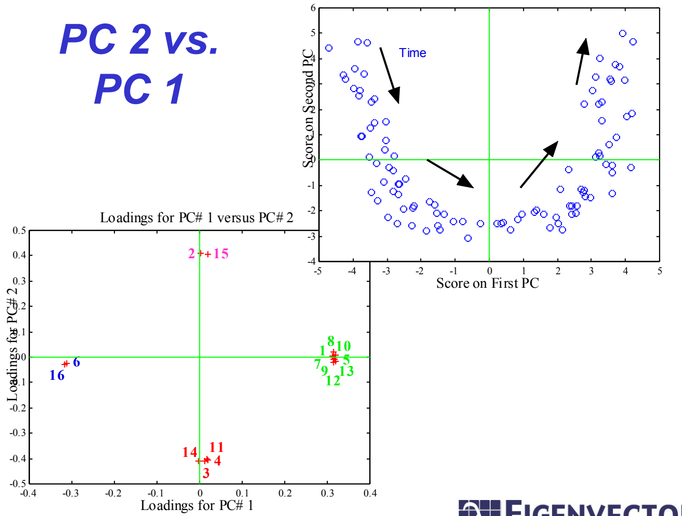
PC 2: Scores and Loadings





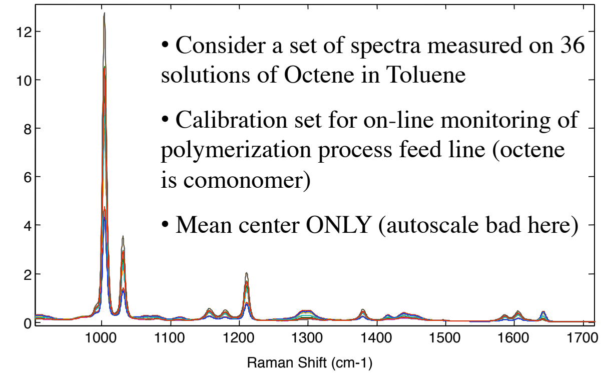
EIGENVECTOR RESEARCH INCORPORATED

PC 2 vs. PC 1



EIGENVECTOR RESEARCH INCORPORATED

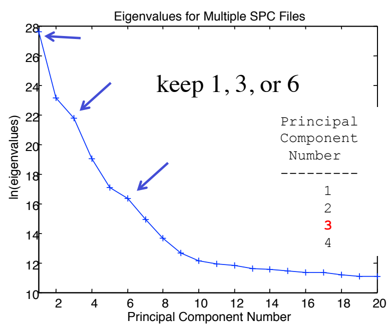
Raman Spectra of Octene in Toluene



- Consider a set of spectra measured on 36 solutions of Octene in Toluene
- Calibration set for on-line monitoring of polymerization process feed line (octene is comonomer)
- Mean center ONLY (autoscale bad here)

EIGENVECTOR RESEARCH INCORPORATED

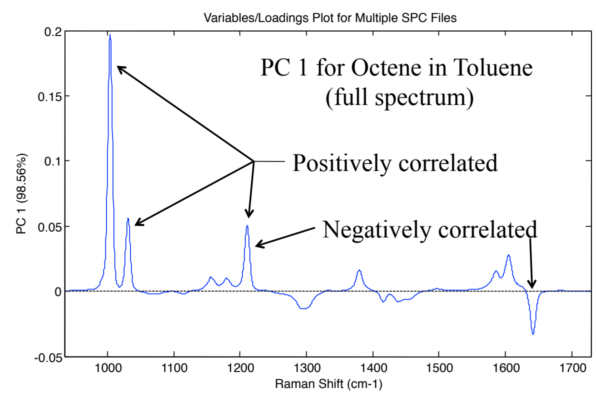
Eigenvalues for Octene in Toluene



Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC
1	9.63e+011	98.56
2	1.10e+010	1.13
3	2.83e+009	0.29
4	1.84e+008	0.02

EIGENVECTOR RESEARCH INCORPORATED

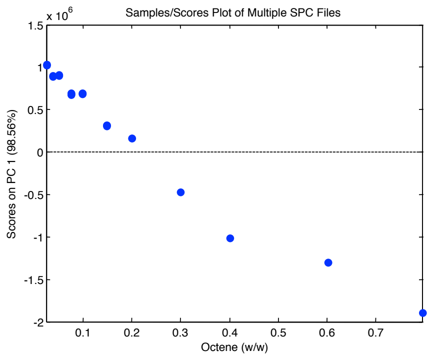
Loadings on PC1



EIGENVECTOR RESEARCH INCORPORATED

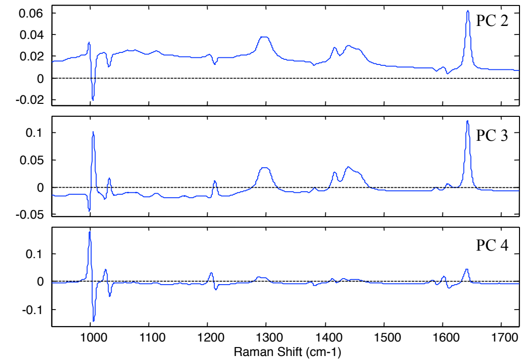
Scores on PC 1

How much of PC1 is observed in each sample?



EIGENVECTOR RESEARCH INCORPORATED

Loadings for PCs 2-4



Can be very difficult to interpret.
BE CAREFUL!

EIGENVECTOR RESEARCH INCORPORATED

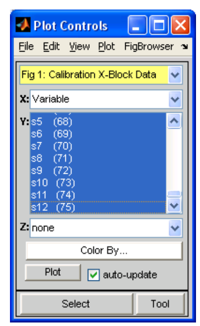
Example: ARCH

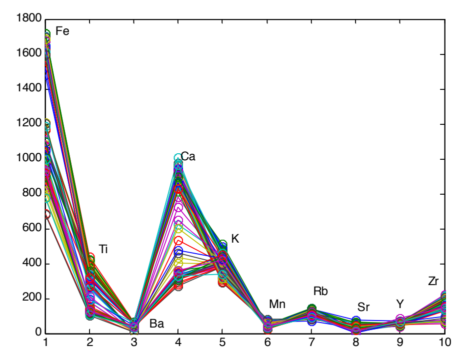
- 10 Variables: metal concentration (ppm via XRF)
- 75 Samples:
  - 63 obsidian samples from 4 quarries (known origin)
  - 12 artifacts (unknown origin)
- Data Matrix **X** is 75 by 10
- Load data from `arch.mat`

EIGENVECTOR RESEARCH INCORPORATED

Raw Data from ARCH

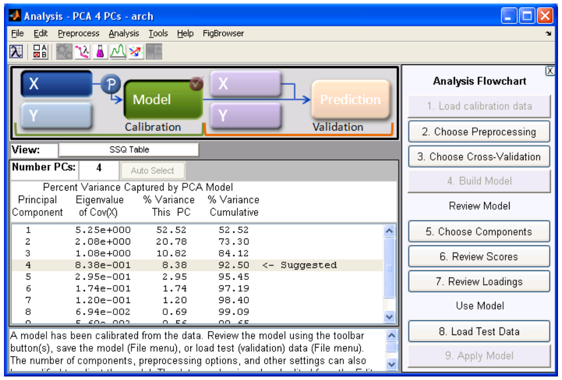
View:Labels
checked





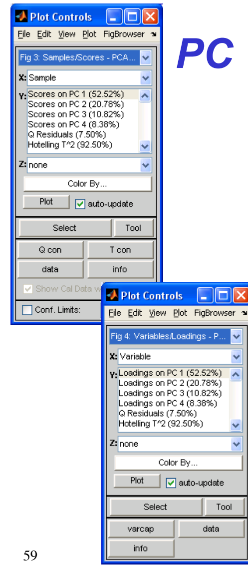
EIGENVECTOR RESEARCH INCORPORATED

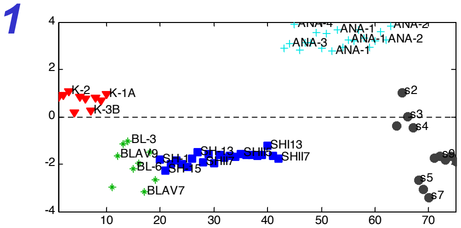
Variance Captured by PCA Model

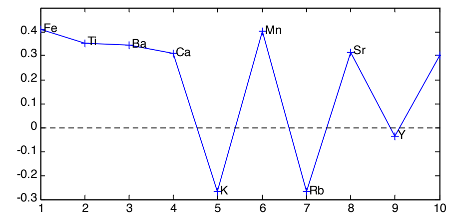


4 PCs selected

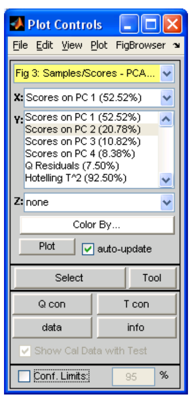
PC 1

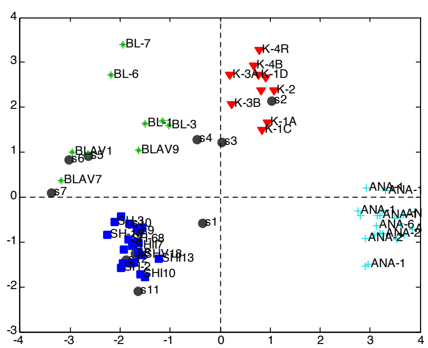




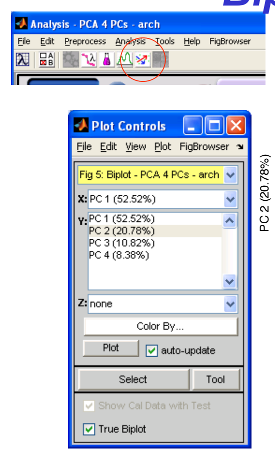


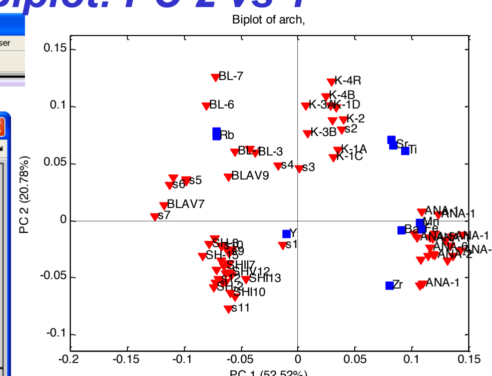
Scores on PC 2 vs 1



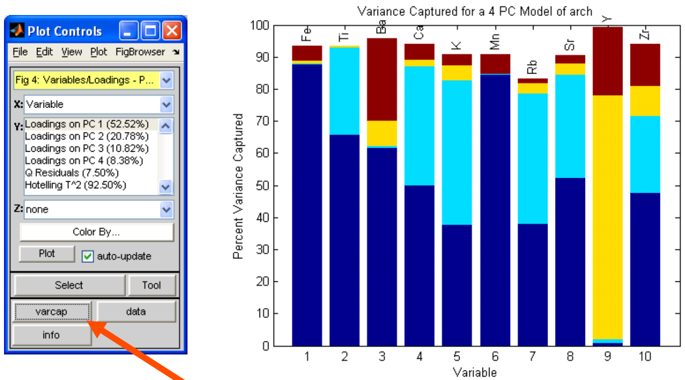


Biplot: PC 2 vs 1





Variance Captured by Variables



1 Click varcap

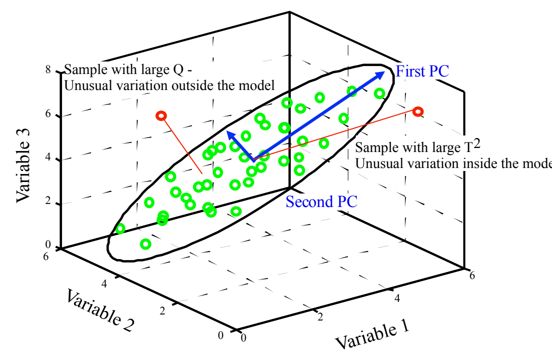
Important Diagnostics

- Q
  - portion of measurement not explained by the model
    - small Q residual => sample well explained by model
    - the converse is also true
  - residuals are orthogonal to the model space

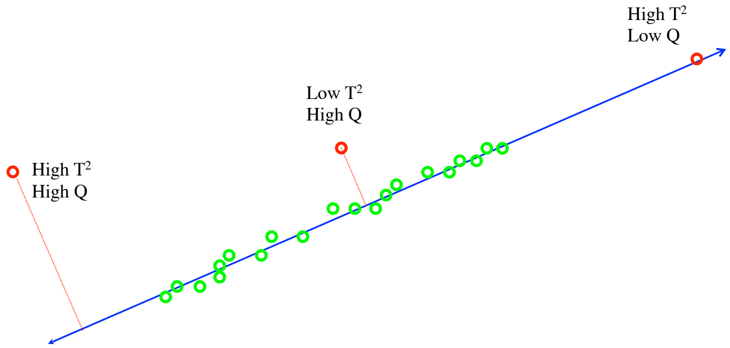
Hotelling's T^2

- Measure of distance to center of model to the point defined by the projection of the sample in the space of the model
- A sample having a large value of T^2 means that
  - the projection into the model space is unusually far away from the center of the model

Geometry of Q and T^2



Another Perspective



EIGENVECTOR RESEARCH INCORPORATED

Control Limits for PCA Statistics

- Control limits can be set for
  - lack of fit statistics: for a row of **E**, $\mathbf{e}_i$, and a row of **X**, $\mathbf{x}_i$
    - Q contributions
      $$\mathbf{e}_i = \mathbf{x}_i\left(\mathbf{I} - \mathbf{P}_k\mathbf{P}_k^T\right)$$
    - Q residual (sum of squares)
      $$Q = \mathbf{e}_i\mathbf{e}_i^T = \mathbf{x}_i\left(\mathbf{I} - \mathbf{P}_k\mathbf{P}_k^T\right)\mathbf{x}_i^T$$
  - Hotelling's T^2: for a row of $\mathbf{T}_k$, $\mathbf{t}_i$, and kxk diagonal matrix λ
    - T^2 contributions
      $$T^2_{i,con} = \mathbf{t}_i\lambda^{-1}\mathbf{P}_k^T = \mathbf{x}_i\mathbf{P}_k\lambda^{-1}\mathbf{P}_k^T$$
    - T^2
      $$T^2_i = \mathbf{t}_i\lambda^{-1}\mathbf{t}_i^T = \mathbf{x}_i\mathbf{P}_k\lambda^{-1}\mathbf{P}_k^T\mathbf{x}_i^T$$
  - also for:
    - scores, $\mathbf{t}_{ij}$
    - residuals $\mathbf{e}_{ij}$

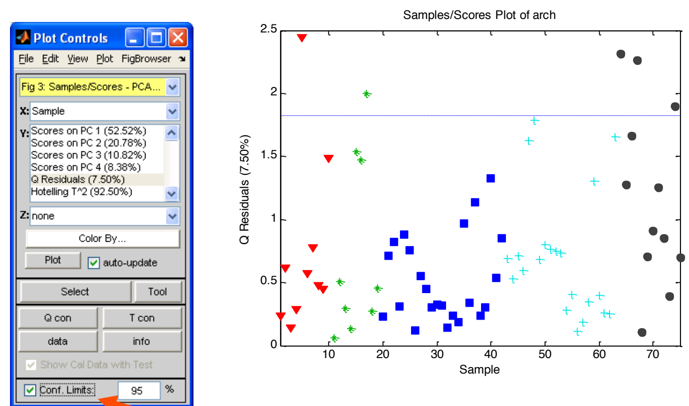
EIGENVECTOR RESEARCH INCORPORATED

Contributions

- Contributions to Q show how samples are different from the PCA model
  - Contributions to Q are a row of **E**
    $$\mathbf{e}_i = \mathbf{x}_i(\mathbf{I}\text{-}\mathbf{P}_k\mathbf{P}_k^T)$$
- Contributions to T^2 show how the original variables deviate from the mean within the model
  $$T^2_{i,con} = \mathbf{t}_i\lambda^{-1}\mathbf{P}_k^T = \mathbf{x}_i\mathbf{P}_k\lambda^{-1}\mathbf{P}_k^T$$

EIGENVECTOR RESEARCH INCORPORATED

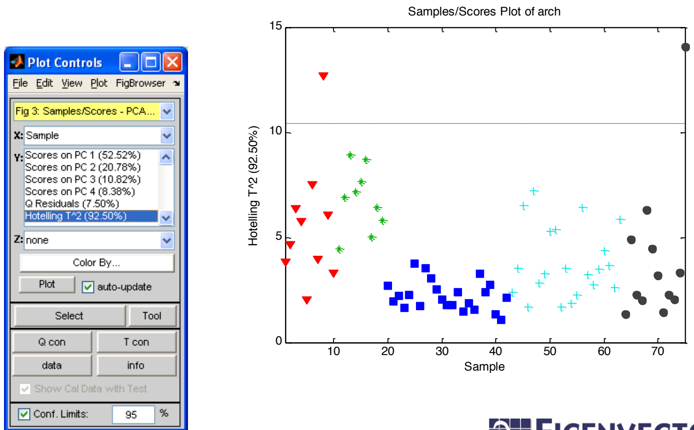
Q Residuals for ARCH data



1 Check Conf. Limits

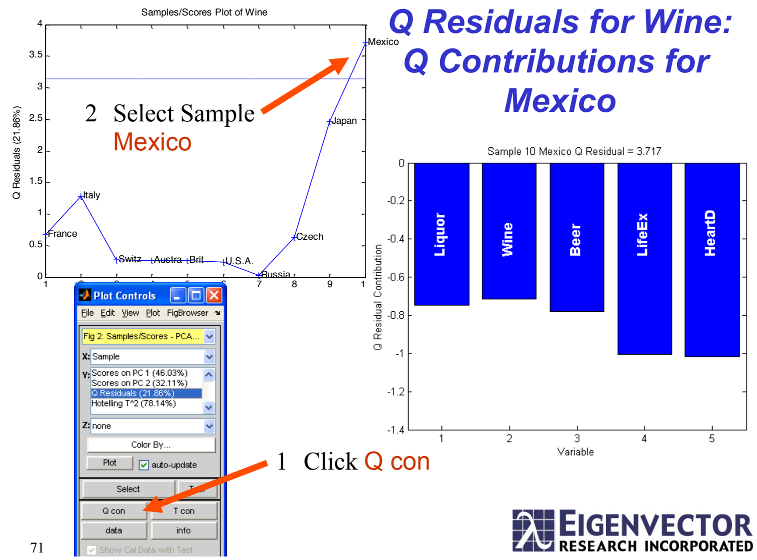
EIGENVECTOR RESEARCH INCORPORATED

T^2 for ARCH



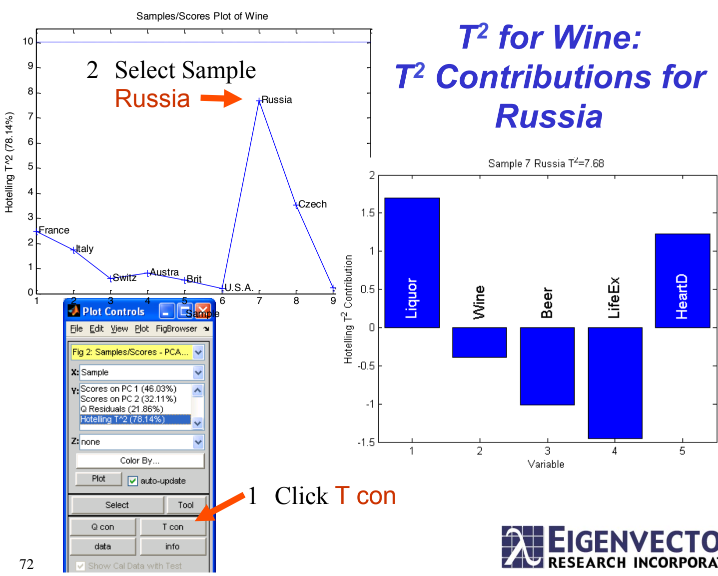
EIGENVECTOR RESEARCH INCORPORATED

Q Residuals for Wine: Q Contributions for Mexico



EIGENVECTOR RESEARCH INCORPORATED

T^2 for Wine: T^2 Contributions for Russia



EIGENVECTOR RESEARCH INCORPORATED

Outliers

- Outlier samples can have a large influence on a PCA model
- However, they are usually easily found!
- To check for outliers, look for:
  - stray samples on scores plots
  - samples with very high Q, T^2, or both

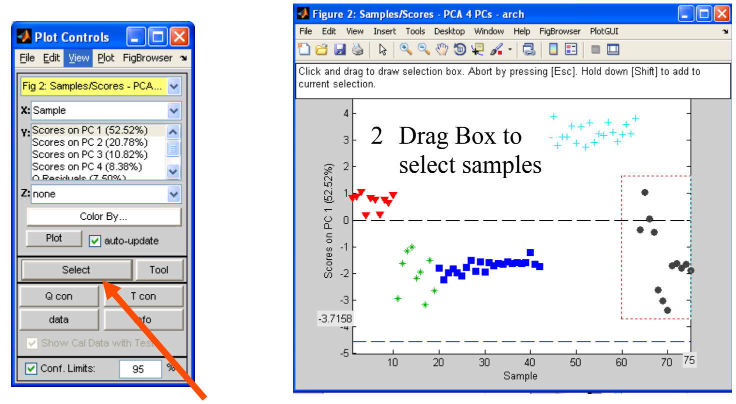
EIGENVECTOR RESEARCH INCORPORATED

PCA Application to New Data

- center new data to the mean of the calibration data
  $$\mathbf{X}_c = \mathbf{X} - \mathbf{1x}_{mean}$$
- scale the centered data using standard deviations of cal data
  $$\mathbf{X}_s = \mathbf{X}_c \,./\, \mathbf{1x}_{std}$$
- project centered and scaled data onto loadings to get new scores
  $$\mathbf{T}_{new} = \mathbf{X}_s\mathbf{P}_k$$
- calculate new residuals
  $$\mathbf{E}_{new} = \mathbf{X}_s - \mathbf{T}_{new}\mathbf{P}_k^T = \mathbf{X}_s(\mathbf{I} - \mathbf{PP}^T)$$
- calculate new Q residuals
  $$\mathbf{Q}_{new} = \mathrm{diag}(\mathbf{E}_{new}\mathbf{E}_{new}^T)$$
- calculate new T^2 values
  $$T^2_{new} = \mathbf{T}_{new}\lambda^{-1}\mathbf{T}_{new}^T = \mathbf{X}_s\mathbf{P}_k\lambda^{-1}\mathbf{P}_k^T\mathbf{X}_i^T$$
- compare $\mathbf{T}_{new}$, $\mathbf{E}_{new}$, $\mathbf{Q}_{new}$ and T^2_{new} to previously determined limits

EIGENVECTOR RESEARCH INCORPORATED

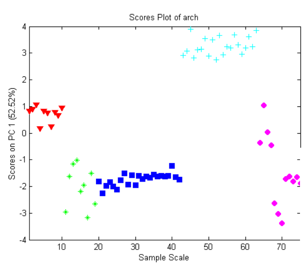
Selecting Samples: ARCH Data

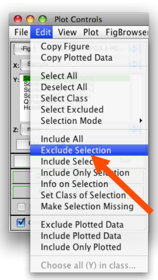


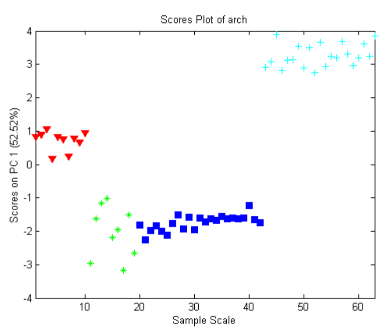
1 Click Select

EIGENVECTOR RESEARCH INCORPORATED

Deleting Samples: ARCH Data



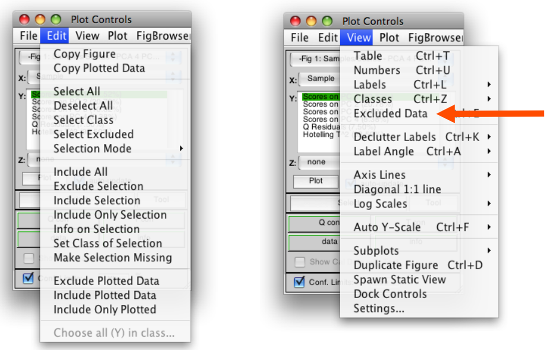




1 Edit menu highlight Exclude Selection

EIGENVECTOR RESEARCH INCORPORATED

Graphically Editing

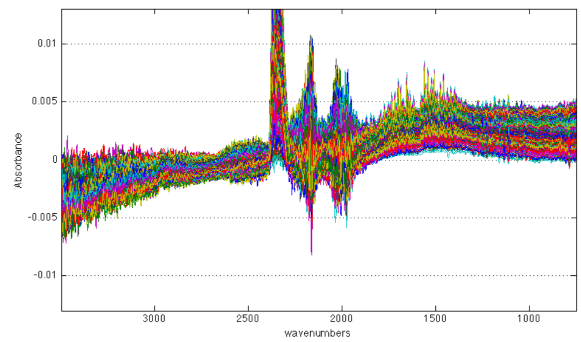


EIGENVECTOR RESEARCH INCORPORATED

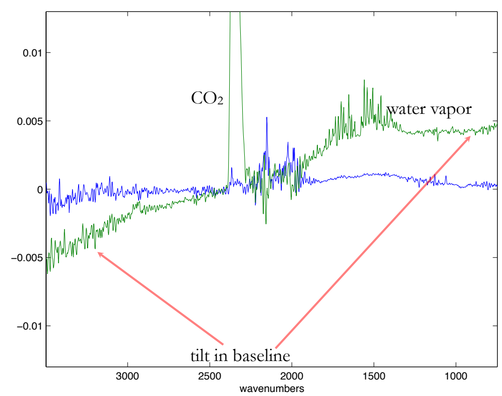
Centering of Data

- We've been consistent in stating that data should be centered (remember that autoscaling contains centering)
  - The idea is that we're looking at how data varies from this conceptual center point
  - However, if **0** (multivariate zero, [0 0 0 . . . 0] is a realistic or even idealized part of your sample space, consider **not** centering
- Example
  - Stability of 100% T lines for real-time spectral acquisition

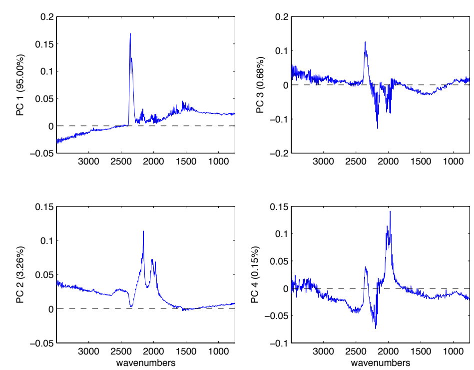
Stability Data



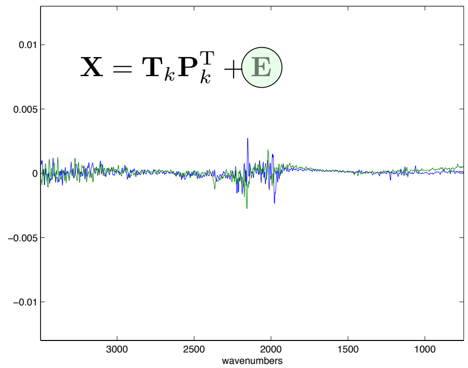
First and Last Spectra



Loadings 1-4 for PCA Model (no centering)



First and Last Spectra After Filtering

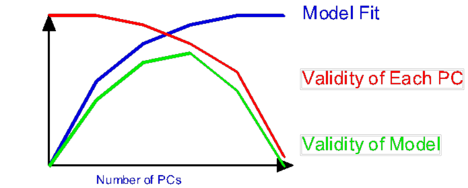


EIGENVECTOR RESEARCH INCORPORATED

How Many Principal Components?

As more PCs are kept in the model, the fit improves, but ....

The validity of the model, *when applied to new data*, eventually declines



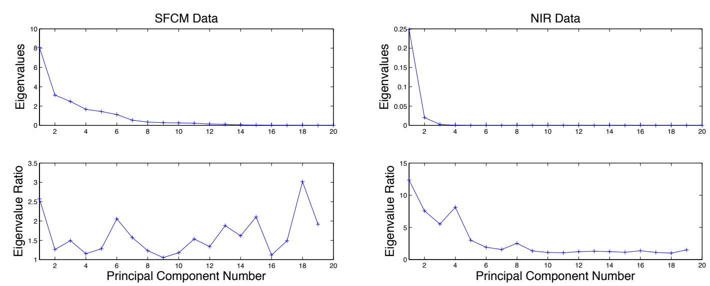
EIGENVECTOR RESEARCH INCORPORATED

Determining the Number of Principal Components

- Determination of the right number of PCs to retain in a model not always simple
- Many methods available:
  - Plot eigenvalues, look for "knee"
  - Ratios of successive eigenvalues
  - For autoscaled data, retain PCs with l > ~1-2
  - Retain PCs with %variance > noise level
  - Omit PCs that don't make sense!
  - Use cross-validation or jack-knifing

EIGENVECTOR RESEARCH INCORPORATED

Knees and Ratios



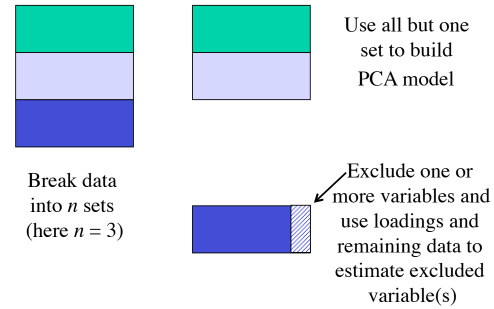
EIGENVECTOR RESEARCH INCORPORATED

Cross-Validation

- Divide data set into *j* subsets
- Build PCA model on *j*-1 subsets
- Calculate PRESS (Predictive Residual Sum of Squares) for the subset left out
  - (PCA method uses estimates of "missing")
- Repeat *j* times (until all subsets have been left out once)
- Look for minimum or knee in PRESS curve

EIGENVECTOR RESEARCH INCORPORATED

PCA Cross-validation

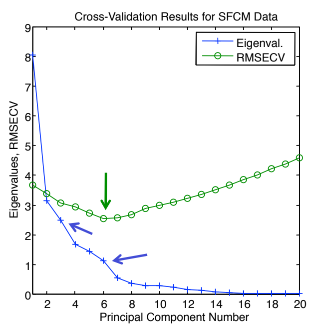


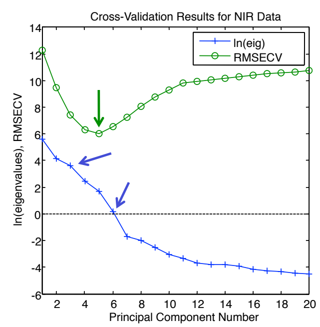
Loop over
- number of PCs
- left out variables
- left out test set

Calculate estimation error as function of number of PCs used

EIGENVECTOR RESEARCH INCORPORATED

Cross-Validation Examples





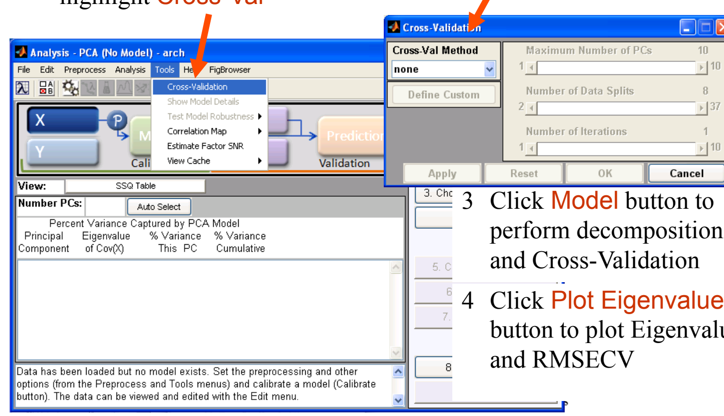
RMSECV: Look for minimum
Eigenvalues: Look for knee

EIGENVECTOR RESEARCH INCORPORATED

Cross-Validation

1 Tools menu highlight Cross-Val

2 Select Cross-validation method



3 Click Model button to perform decomposition and Cross-Validation

4 Click Plot Eigenvalues button to plot Eigenvalues and RMSECV

Example: Olive Oil Data Set

- Use FT-IR spectra and pattern recognition to distinguish authentic olive oil from counterfeit or adulterated olive oil.
- Obtain FT-IR spectra (3600 - 600 cm^{-1}) of these oils using a fixed pathlength NaCl cell
- Have a calibration set (36 samples) and a distinct test set (44 samples)
- Reference:
  D.B. Dahlberg, S.M. Lee, S.J. Wenger, J.A. Vargo "Classification of Vegetable Oils by FT-IR," Appl. Spectrosc., 51(8), 1118-1124 (1997)

Calibration Set: Details

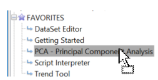
Corn Oil	9 samples	(#1-9)
Olive Oil	15 samples	(#10-24)
Safflower Oil	8 samples	(#25-32)
Corn Margarine	4 samples	(#33-36)

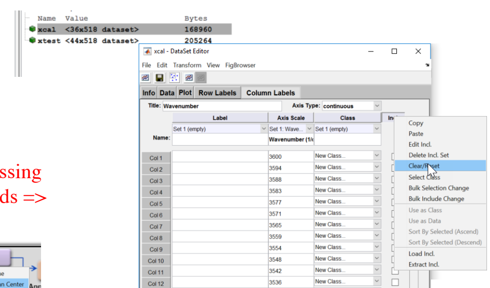
PCA: Entire Spectrum

- Drag xcal in the Browse window onto PCA
- Change the preprocessing for the x-block to mean centering
- Open the x-block in the dataset editor and include all of the variables

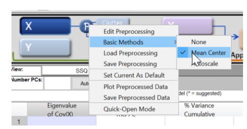
Steps

Drag xcal onto PCA



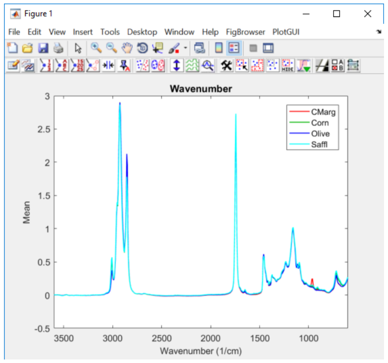


Right-click on preprocessing button, => Basic Methods => Mean Center



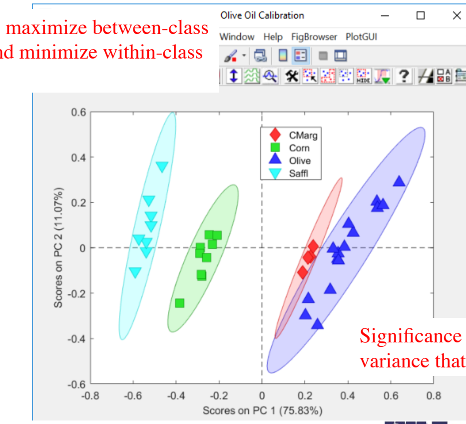
Right-click on "X", select Edit Data, click on "Column Labels", right-click on "Incl.", select Clear/Reset

Plot: Summary with Classes



PCA: Scores Plot

Objective: maximize between-class variance and minimize within-class variance



Significance within-class variance that is directional

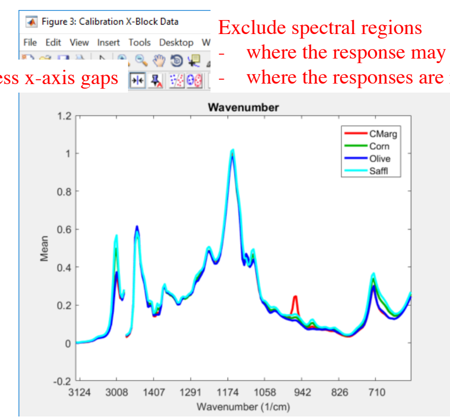
Reload the X-Block

- Right-click on the “X”
- Select “Load Data”
- Select “xcal” from workspace
- Plot the data

Included Data

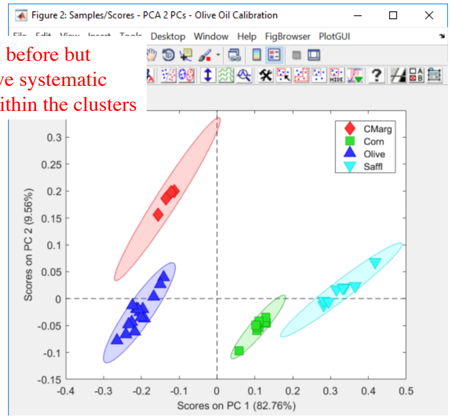
Exclude spectral regions
- where the response may be nonlinear
- where the responses are mainly similar

Compress x-axis gaps



PCA: Scores Plot Revisited

Better than before but
we still have systematic
variance within the clusters



EIGENVECTOR RESEARCH INCORPORATED

A Brief Word about Preprocessing

- To this point, we’ve focused on just mean-centering and autoscaling (which includes mean-centering)
- There’s a wide variety of preprocessing tools in the toolbox, and others can be created depending upon the nature of the data
- In general, the objective of preprocessing is to remove sources of variance that impede us from our modeling objective
  - In this case, we have significant systematic variance within the classes

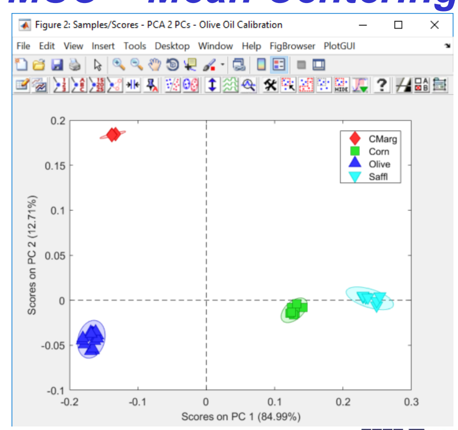
EIGENVECTOR RESEARCH INCORPORATED

Targeting of Variance Removal

- As it turns out with this data, there is a substantial variation in the effective pathlength
  - Somewhat surprising given that these are transmission measurements
- When spectroscopic data has effective pathlength indetermancy, some type of normalization can frequently help such as
  - 1- norm normalization
  - 2-norm normalization
  - SNV (single normal variate)
  - MSC (multiple scatter correction)

EIGENVECTOR RESEARCH INCORPORATED

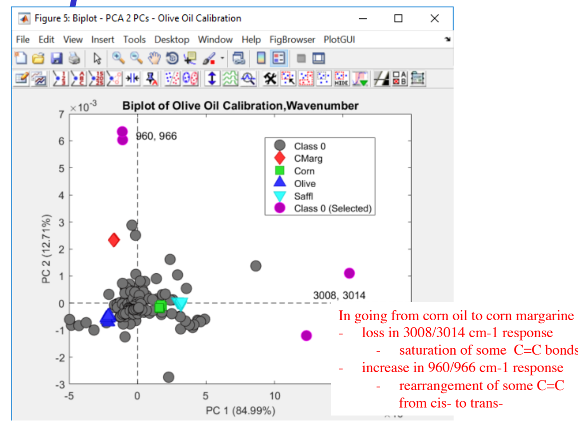
MSC + Mean Centering



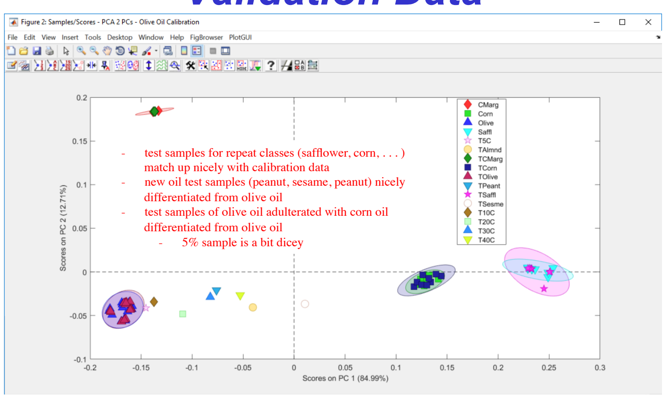
EIGENVECTOR RESEARCH INCORPORATED

Loadings: PC2 and PC1

Biplot: PC2 and PC1



Validation Data



Exploring PCA Models

- Much can be learned from considering scores and loadings plots in combination
  - scores plots show how samples are spread out or grouped
  - loadings plots show what variables are correlated, anti-correlated and uncorrelated
  - together they show what variables are responsible for the variations you see in the samples
- Can additional information be brought in?
  - have shown examples with sample classes
  - can also use “color-by” to add information

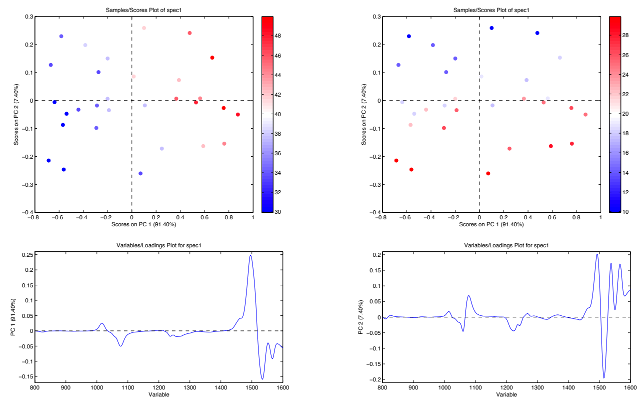
Using “color-by”

- Color points in scores or loadings plots according to any other available parameter
  - color scores by concentration or quality values, time they were measured, etc.
  - color loadings by wavelength, type of measurement, etc.

EIGENVECTOR RESEARCH INCORPORATED

Color-by on NIR data

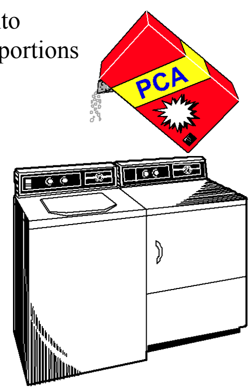
colored by toluene concentration | colored by iso-octane concentration



EIGENVECTOR RESEARCH INCORPORATED

Dirty T-Shirt Analogy

PCA attempts to partition the data into deterministic and non-deterministic portions



EIGENVECTOR RESEARCH INCORPORATED