Principal Components and Exploratory Data Analysis

EIGENVECTOR RESEARCH INCORPORATED

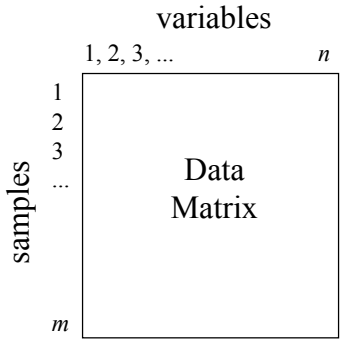
Nomenclature and Conventions

- 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 λ
  - 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 $\alpha\ e^{-1/T}$, transform with log
  - pipe flow $\alpha\ \Delta P^{4/7}$ (turbulent flow)

EIGENVECTOR RESEARCH INCORPORATED

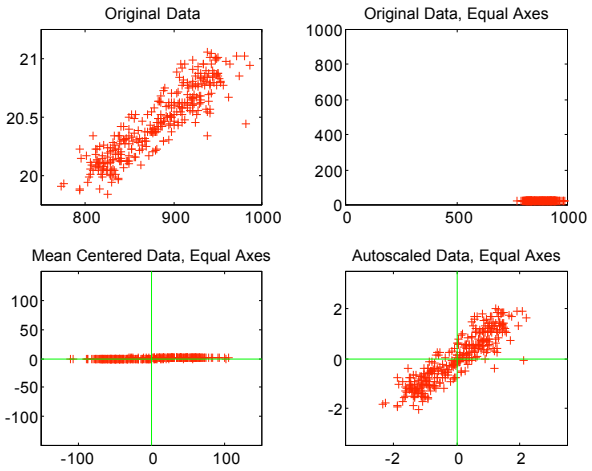
Mean Centering

- 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
  - `mncn`

Variance Scaling

- PCA is scale dependent, variance is associated with importance
- This may or may not be true
- In spectra, variance α importance (probably)
- If variables have different units, variance $\sim\alpha$ 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
  - `auto`
- Other scaling - may want to use *a priori* information, such as noise level in variables

Centering & Scaling Example



example with SFCM data
in `plsdata`

EIGENVECTOR RESEARCH INCORPORATED

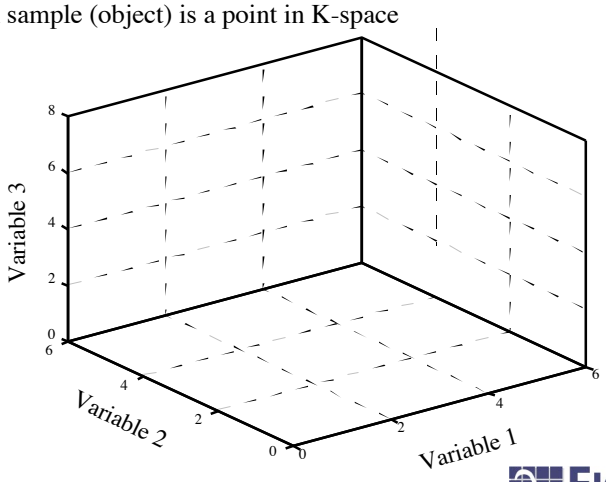
Block Scaling

- 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

EIGENVECTOR RESEARCH INCORPORATED

Principal of Projections

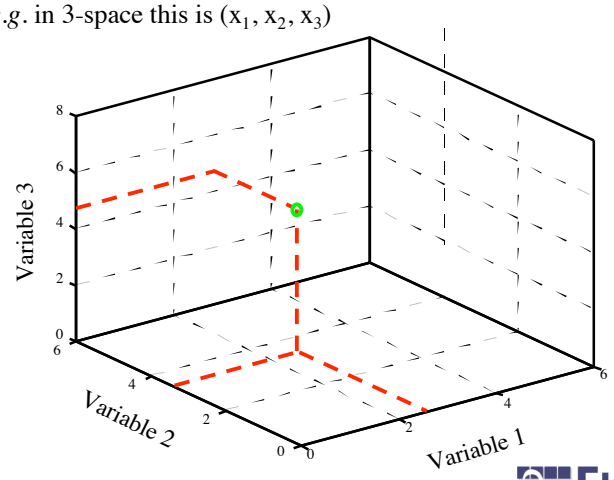
- K-space has K dimensions where each variable, or measurement on an object, is a coordinate axis
- A sample (object) is a point in K-space



EIGENVECTOR RESEARCH INCORPORATED

Projection in K-Space

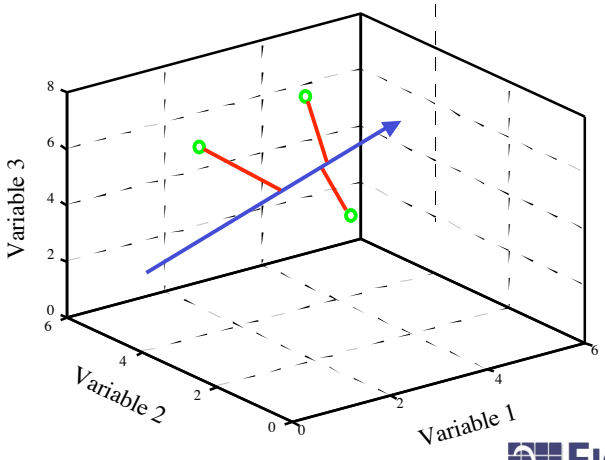
- The projection of an object onto the K-space yields the coordinates of the object in that space
- *e.g.* in 3-space this is (x_1, x_2, x_3)



EIGENVECTOR RESEARCH INCORPORATED

Projection onto a Vector

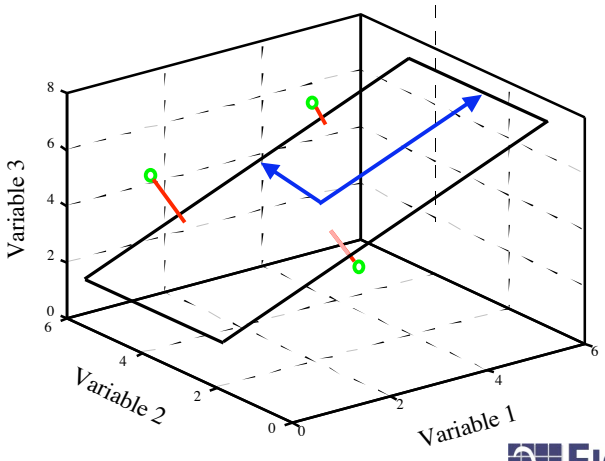
- Projection lines are perpendicular to the vector



EIGENVECTOR RESEARCH INCORPORATED

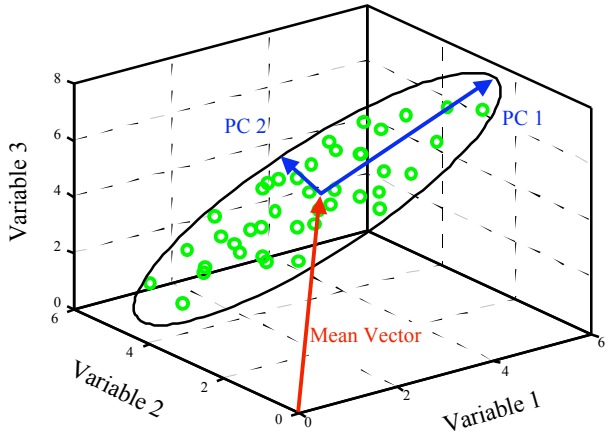
Projection onto a Plane

- Projection lines are perpendicular to the plane



EIGENVECTOR RESEARCH INCORPORATED

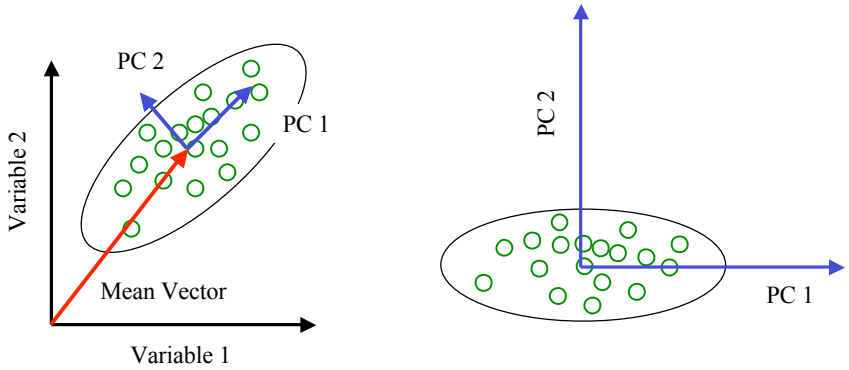
PCA



EIGENVECTOR RESEARCH INCORPORATED

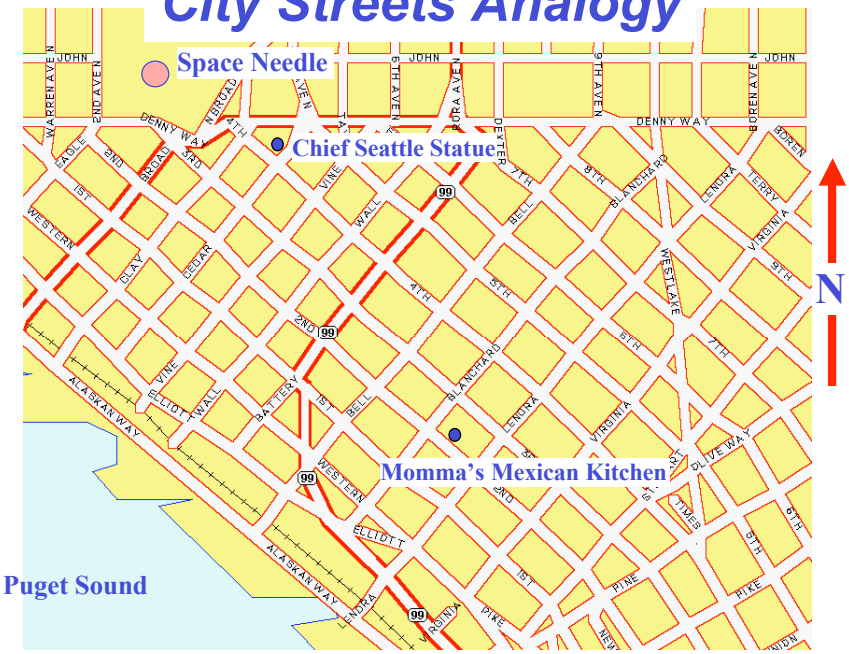
PCA

- Geometry for 2 variables



EIGENVECTOR RESEARCH INCORPORATED

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:

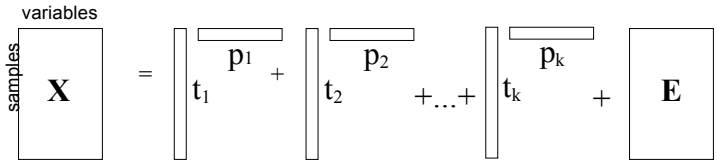
$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_k\mathbf{p}_k^T + \ldots + \mathbf{t}_q\mathbf{p}_q^T$$

Where $q \leq \min(m,n)$, and the $\mathbf{t}_i\mathbf{p}_i^T$ pairs are ordered by the amount of variance captured.

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

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

PCA Math 2 of 3



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

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

$$\text{cov}(\mathbf{X})\mathbf{p}_i = \lambda_i\mathbf{p}_i$$

and λ_i are eigenvalues.

Amount of variance captured by $\mathbf{t}_i\mathbf{p}_i^T$ proportional to λ_i.

PCA Math 3 of 3

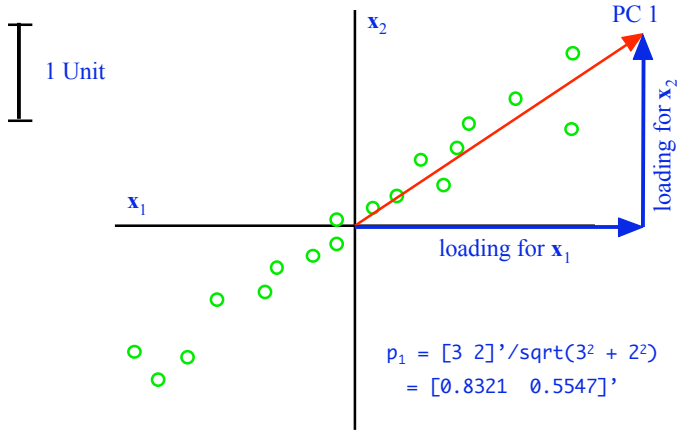
- What is PCA doing mathematically?
- For a data set $\mathbf{X}$, propose that $\mathbf{t} = \mathbf{Xp}$
  - *i.e.* $\mathbf{X}$ projected onto factor $\mathbf{p}$ yields $\mathbf{t}$
  - $\mathbf{X}$ is usually centered and scaled
  - $\max\{\mathbf{t}^T\mathbf{t} \mid \mathbf{p}^T\mathbf{p}=1\} =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

Properties of PCA

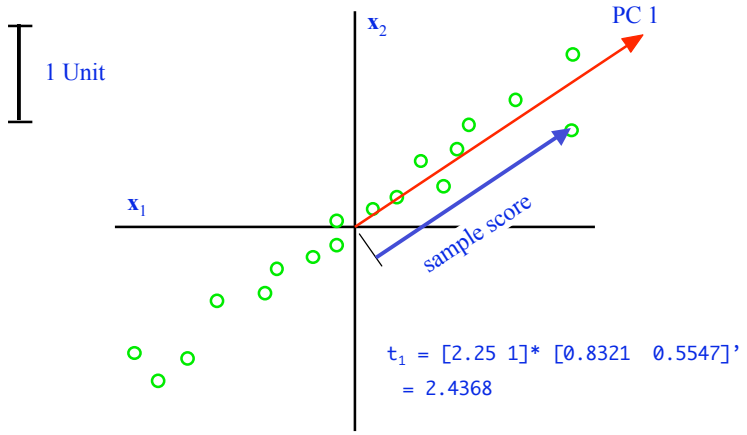
- $\mathbf{t}_i$,$\mathbf{p}_i$ ordered by amount of *variance captured*
- $\mathbf{t}_i$ or *scores* form an orthogonal set $\mathbf{T}_k$ which describe relationship between *samples*
- $\mathbf{p}_i$ or *loadings* form an orthonormal set $\mathbf{P}_k$ which describe relationship between *variables*

- scores and loadings plots are interpreted in pairs
  - *e.g.* plot $\mathbf{t}_i$ vs sample number and $\mathbf{p}_i$ vs variable number
- it is useful to plot $\mathbf{t}_{i+1}$ vs. $\mathbf{t}_i$ and $\mathbf{p}_{i+1}$ vs. $\mathbf{p}_i$

Variable Loadings, p_i



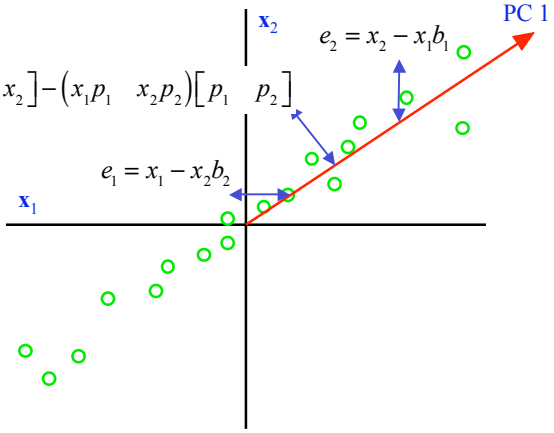
Sample Scores, t_i



EIGENVECTOR RESEARCH INCORPORATED

Minimization Criterion

$$\begin{bmatrix} e_1 & e_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} - \begin{pmatrix} x_1 p_1 & x_2 p_2 \end{pmatrix} \begin{bmatrix} p_1 & p_2 \end{bmatrix}$$



EIGENVECTOR RESEARCH INCORPORATED

Some Mathematical Relationships

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

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{X} - \mathbf{T}_k\mathbf{P}_k^T = \mathbf{X} - \mathbf{X}\mathbf{P}_k\mathbf{P}_k^T = \mathbf{X}(\mathbf{I} - \mathbf{P}_k\mathbf{P}_k^T)$$

- PCA: $\mathbf{X} = \mathbf{T}\mathbf{P}^T = \mathbf{T}_k\mathbf{P}_k^T + \mathbf{E}$
- SVD: $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$
  - $\mathbf{T} = \mathbf{U}\mathbf{S}$
  - $\mathbf{P} = \mathbf{V}$
  - $\mathbf{S}_{ii} = \sqrt{(m-1)\lambda_i}$

EIGENVECTOR RESEARCH INCORPORATED

Example: Wine Data

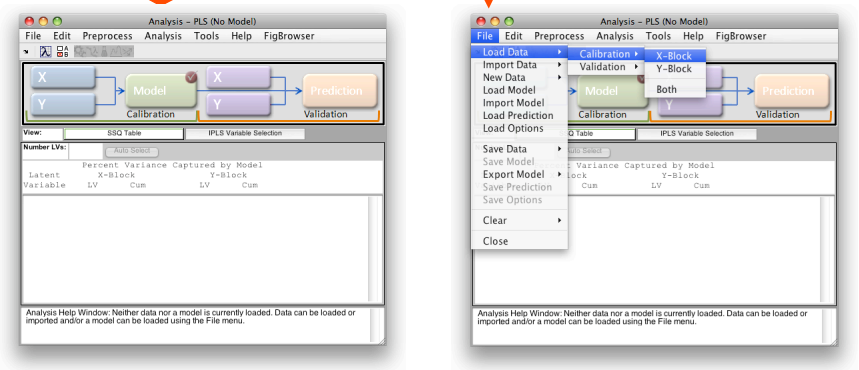
- Examine the relationship between (variables)
  - annual consumption of wine, beer, and liquor (gal/yr),
  - life expectancy (years), and
  - heart disease rate (cases/100,000)
- For 10 different countries (samples)
  - France, Italy, Switzerland, Australia, Britain, USA, Russia, Czech Republic, Japan,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

EIGENVECTOR RESEARCH INCORPORATED

Analysis of Wine Data

1 Type `pca` at the command prompt » to start the PCA program.

2 Click File:Load Data: Calibration:X-Block menu



EIGENVECTOR RESEARCH INCORPORATED

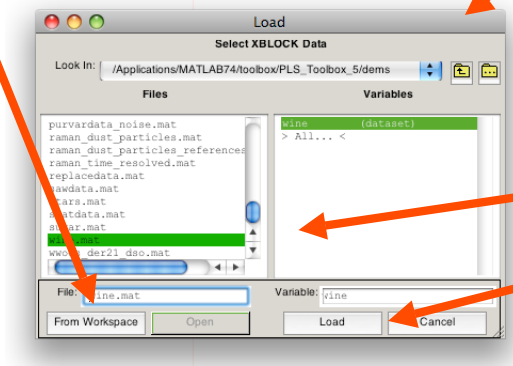
Load wine.mat

1 Click From File button to load from disk
(button will change to From Workspace)

2 Browse to desired folder

3 Highlight wine.mat and wine

4 click load



Tip: type in file name!

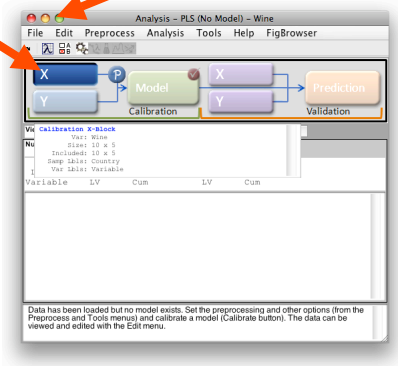
EIGENVECTOR RESEARCH INCORPORATED

Data: loaded but not analyzed

1 status window after load

2 Plot your Data: Select Edit:Calibration:Plot X-Block

Mouse over X to see status of loaded data



EIGENVECTOR RESEARCH INCORPORATED

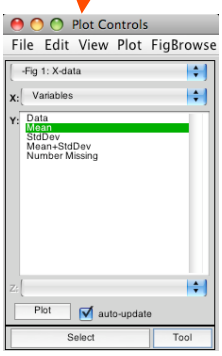
Plot Your Data

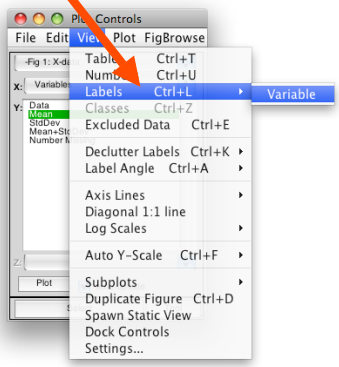
1 Plot control default
can look at summary stats

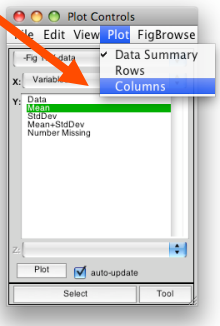
The Plot control generates plots in MATLAB figure windows

2 under view menu check labels

3 under plot menu check columns



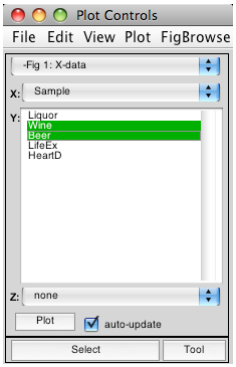


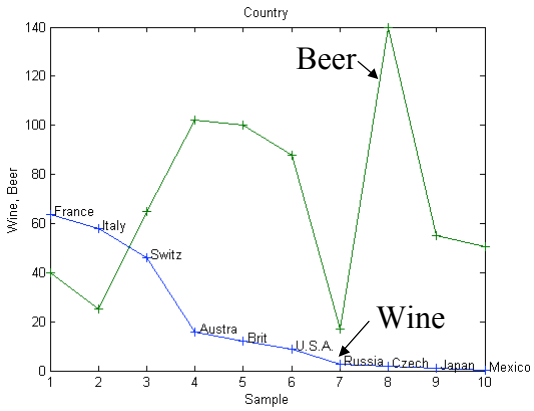


EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data

samples ordered by wine consumption



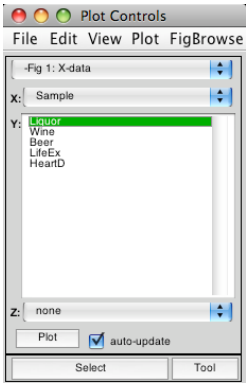


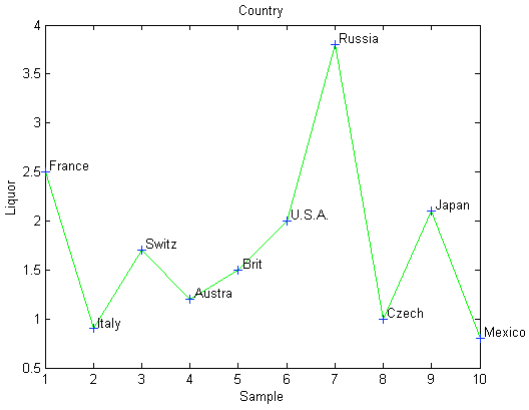
use shift key to select multiple columns

EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data

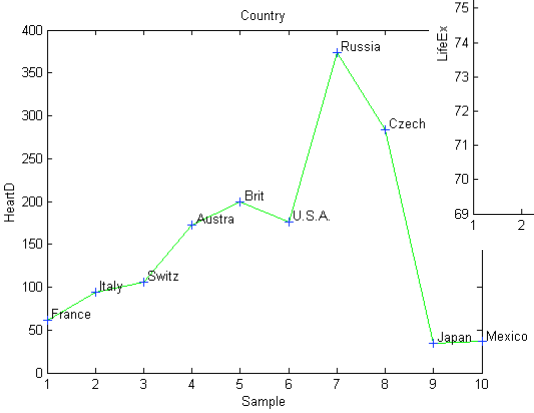
scale is ~1-2 orders of magnitude smaller than for Beer and Wine

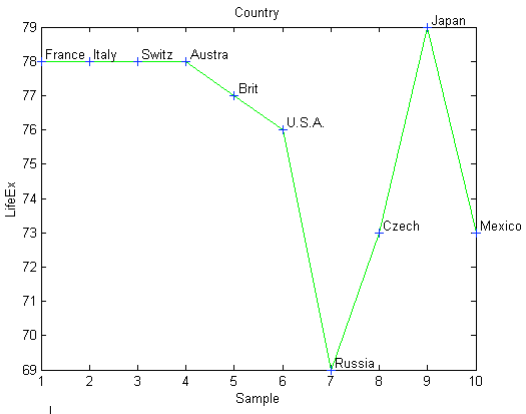




EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data





EIGENVECTOR RESEARCH INCORPORATED

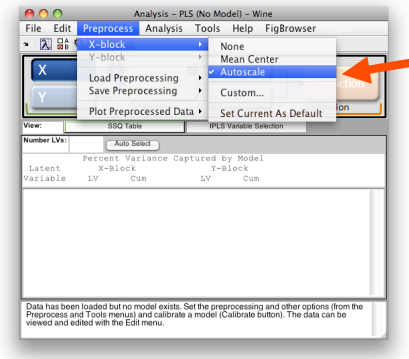
Plot Your Data Summary

- 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

EIGENVECTOR RESEARCH INCORPORATED

How should we scale the data?

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



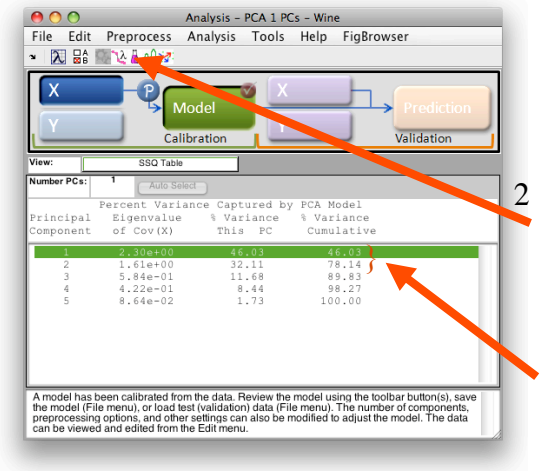
1 autoscaling is the default

2 click Calculate or Model to perform the PCA decomposition

Do the PCA Decomposition

1 After the calc button:

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

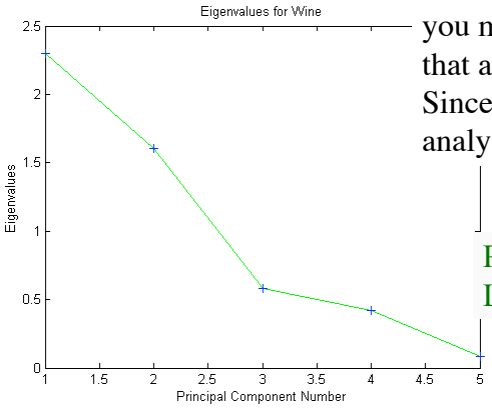


2 Click Plot Eigenvalues button to plot the eigenvalues

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)?
Leave one out CV suggests 1.

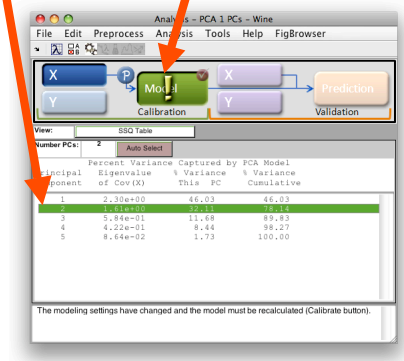
EIGENVECTOR RESEARCH INCORPORATED

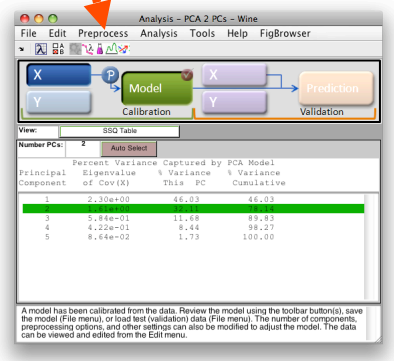
Choose Number of PCs

1 Highlight the second line to select 2 PCs

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

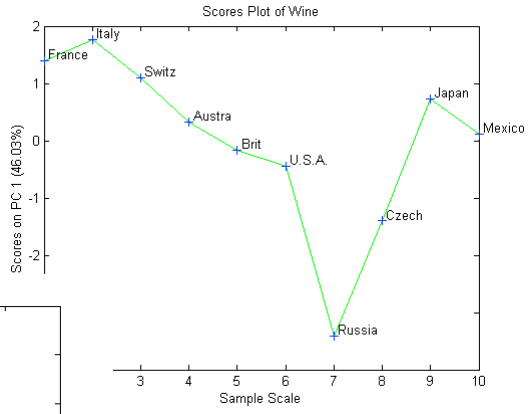
3 Click the scores button to make scores plots, loads button to for loadings plots

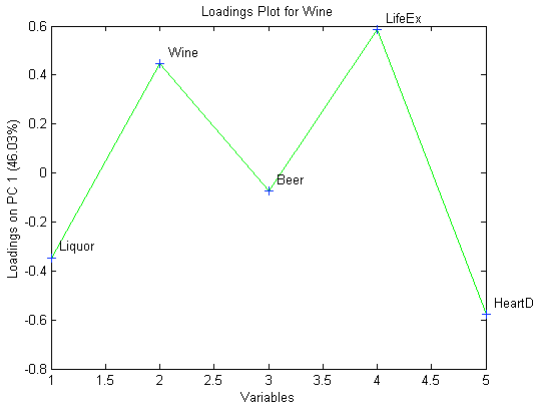




EIGENVECTOR RESEARCH INCORPORATED

Scores and Loads on PC 1





46%

EIGENVECTOR RESEARCH INCORPORATED

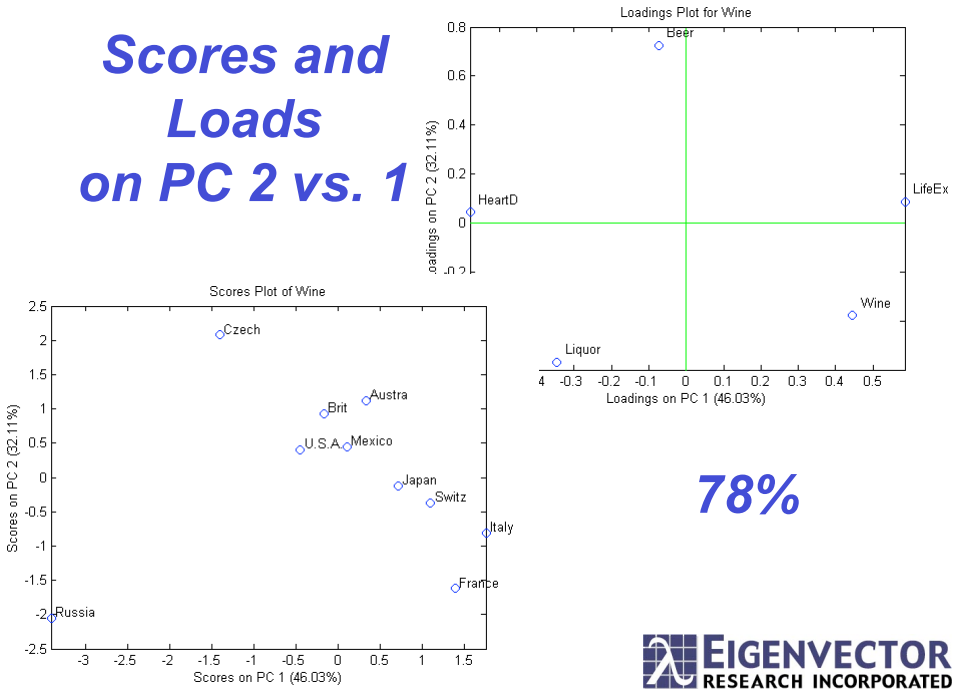
PC 1

- Heart Disease Rate and Liquor Consumption are correlated
- Wine and Life Expectancy are correlated
- Heart Disease Rate and Liquor Consumption are anti-correlated with Wine and Life Expectancy
- Russia is Low on PC 1

- But let's look at PC 2 vs 1 ...

EIGENVECTOR RESEARCH INCORPORATED

Scores and Loads on PC 2 vs. 1



78%

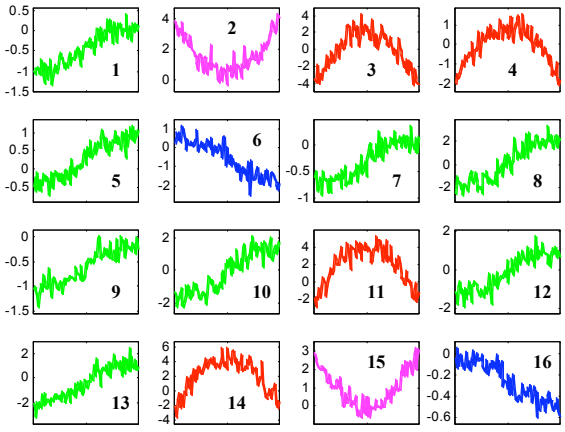
EIGENVECTOR RESEARCH INCORPORATED

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

EIGENVECTOR RESEARCH INCORPORATED

How many PC's to model this data?



EIGENVECTOR RESEARCH INCORPORATED

Variance Captured

Percent Variance Captured by PCA Model

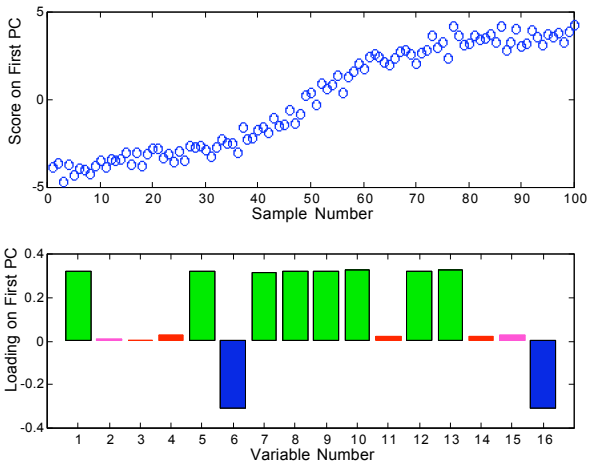
Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	8.79e+00	54.96	54.96
2	5.29e+00	33.05	88.01
3	2.49e-01	1.56	89.57
4	2.17e-01	1.35	90.92
5	1.80e-01	1.12	92.05
6	1.66e-01	1.04	93.08
7	1.51e-01	0.94	94.03
8	1.41e-01	0.88	94.91
9	1.33e-01	0.83	95.74
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?

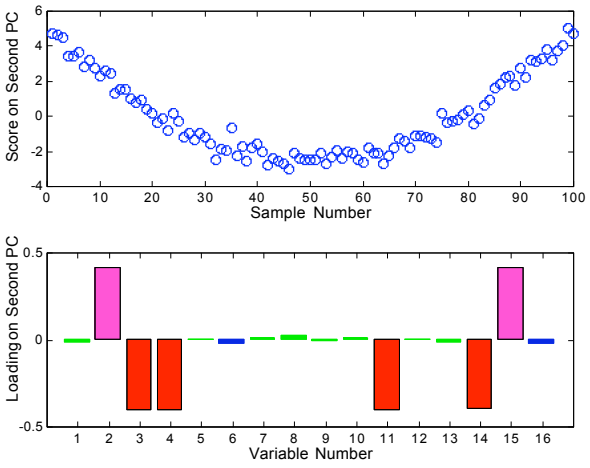
EIGENVECTOR RESEARCH INCORPORATED

PC 1: Scores and Loadings



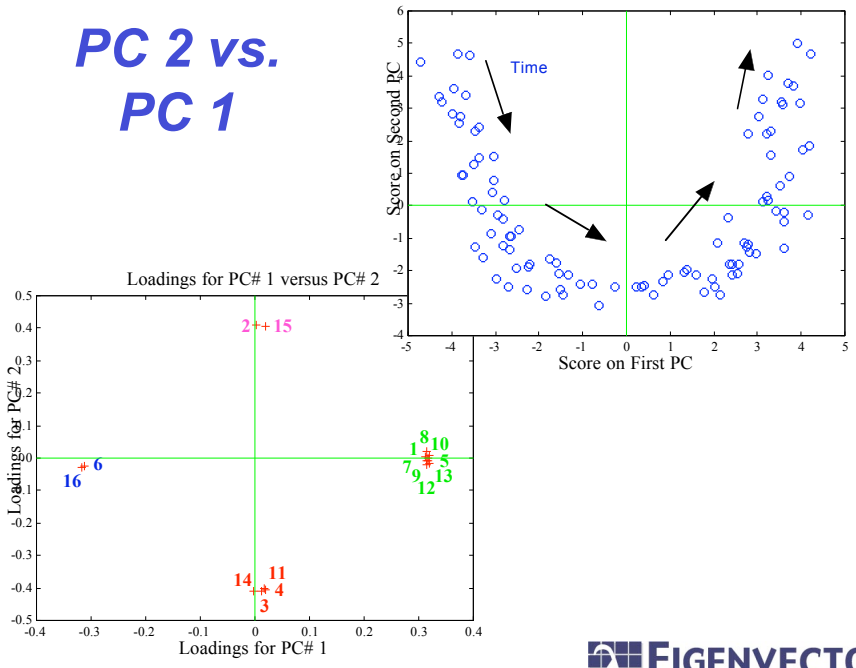
EIGENVECTOR RESEARCH INCORPORATED

PC 2: Scores and Loadings



EIGENVECTOR RESEARCH INCORPORATED

PC 2 vs. PC 1



EIGENVECTOR RESEARCH INCORPORATED

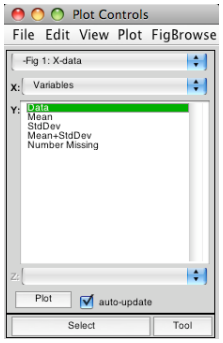
Example: ARCH

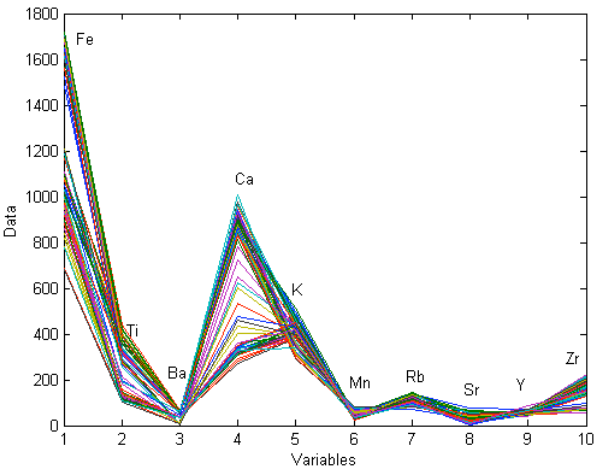
- 10 Variables: metal concentration (ppm via XRF)
- 75 Samples:
  - 63 obsidian samples from 4 quarries (known origin)
  - 12 artifacts (unknown origin)
- Data Matrix **X** is 75 by 10
- Load data from `arch.mat`
  - c:\MATLAB704\toolbox\PLS_Toolbox\dems\arch.mat

EIGENVECTOR RESEARCH INCORPORATED

Raw Data from ARCH

View:Labels
checked

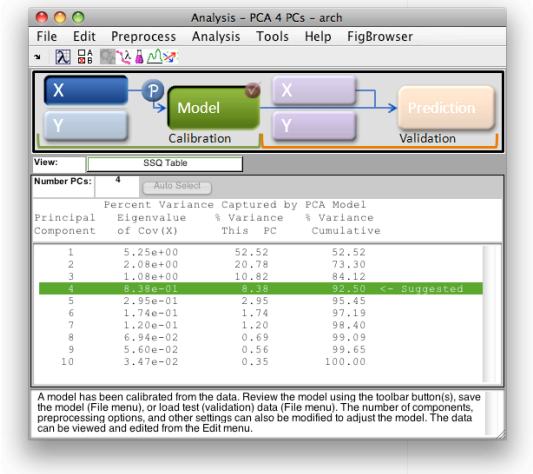




EIGENVECTOR RESEARCH INCORPORATED

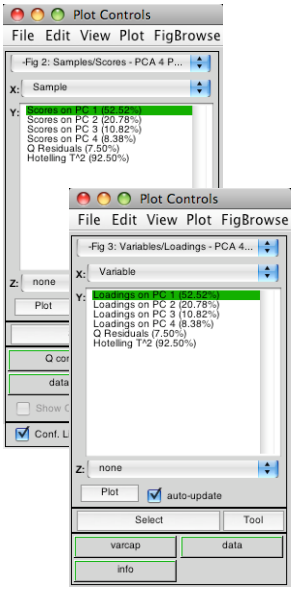
Variance Captured by PCA Model

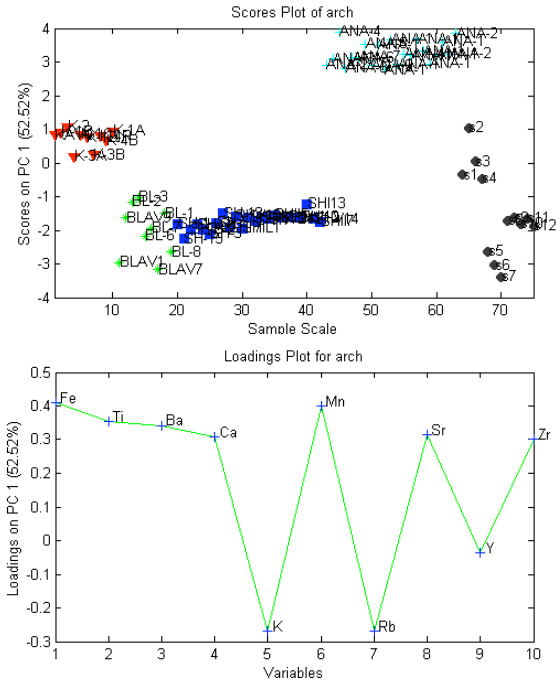
4 PCs
automatically
selected



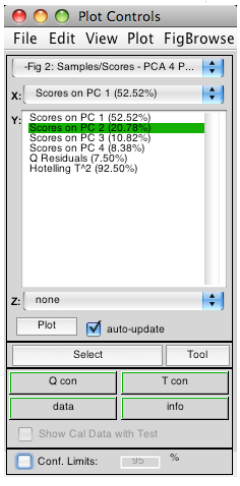
EIGENVECTOR RESEARCH INCORPORATED

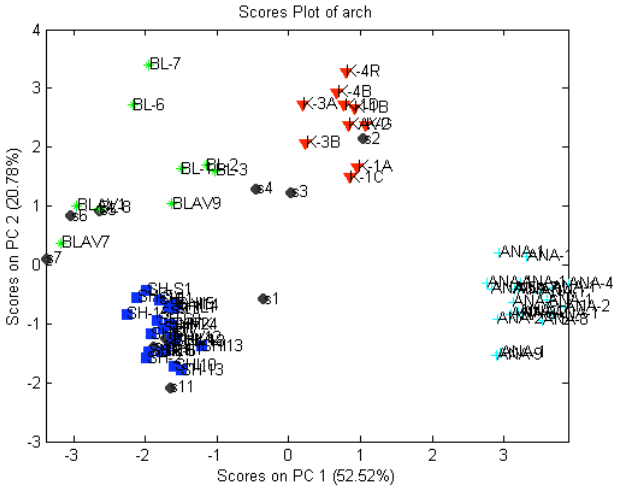
PC 1





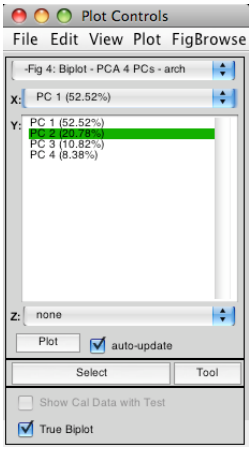
Scores on PC 2 vs 1

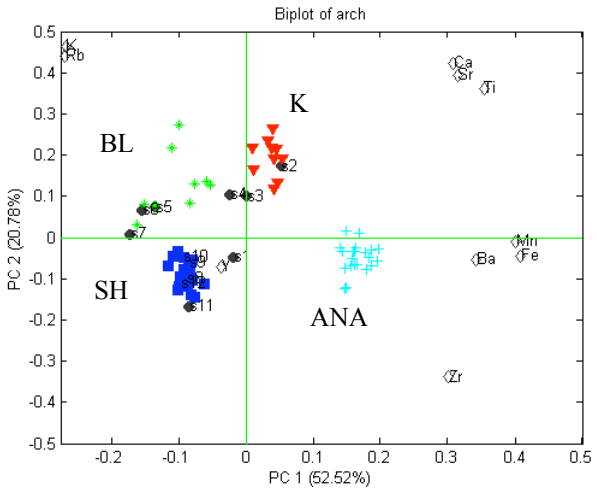




EIGENVECTOR RESEARCH INCORPORATED

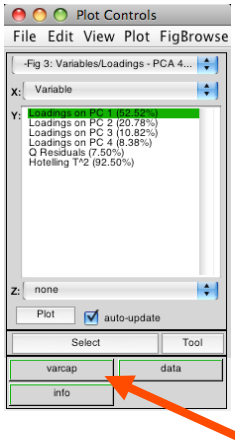
Biplot: PC 2 vs 1

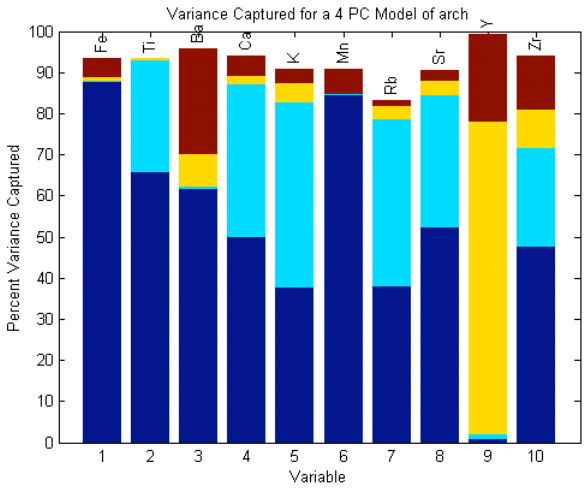




EIGENVECTOR RESEARCH INCORPORATED

Variance Captured by Variables

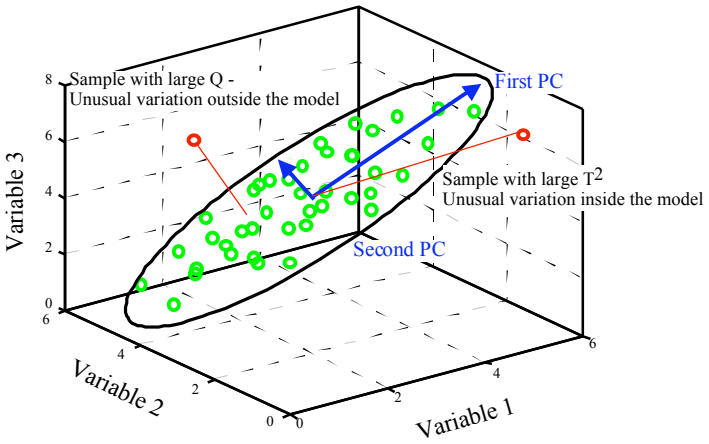




1 Click varcap

EIGENVECTOR RESEARCH INCORPORATED

Geometry of Q and T^2



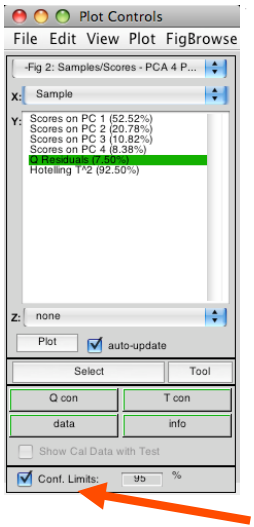
EIGENVECTOR RESEARCH INCORPORATED

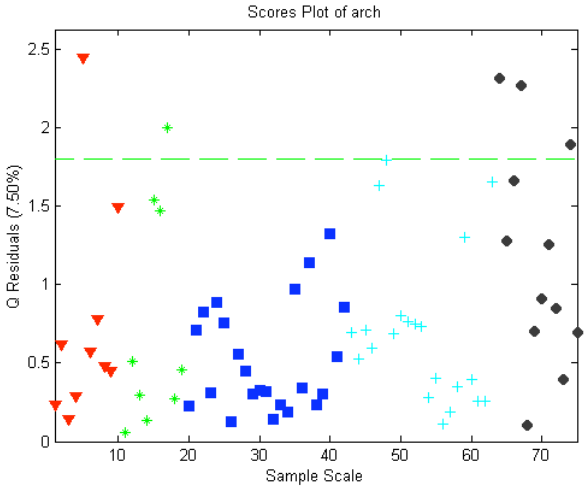
Control Limits for PCA Statistics

- Control limits can be set for
  - lack of fit statistics: for a row of $\mathbf{E}$, $\mathbf{e}_i$, and a row of $\mathbf{X}$, $\mathbf{x}_i$
    - 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 $\boldsymbol{\lambda}$
    - 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

Q Residuals for ARCH data

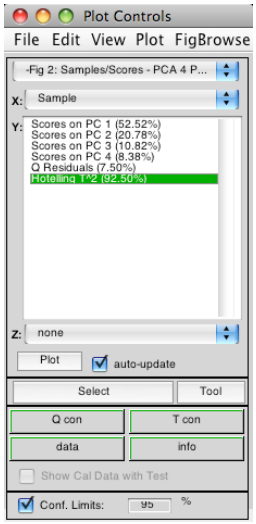


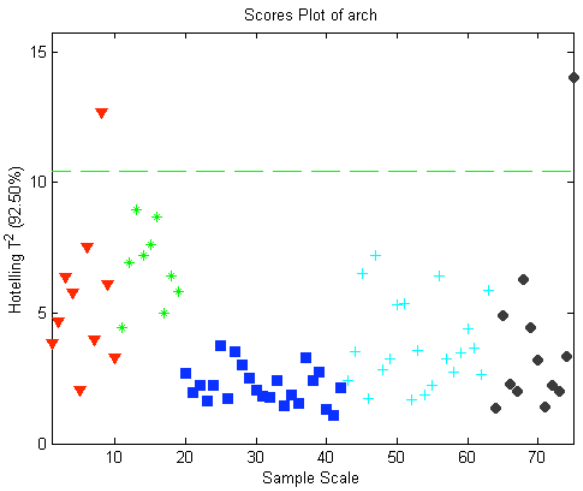


1 Check Conf. Limits

EIGENVECTOR RESEARCH INCORPORATED

T^2 for ARCH





EIGENVECTOR RESEARCH INCORPORATED

Contributions

- Contributions to Q show how samples are different from the PCA model
  - Contributions to Q are a row of **E**

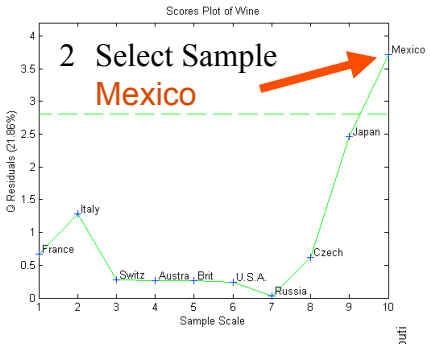
$$\mathbf{e}_i = \mathbf{x}_i(\mathbf{I}-\mathbf{P}_k\mathbf{P}_k^T)$$

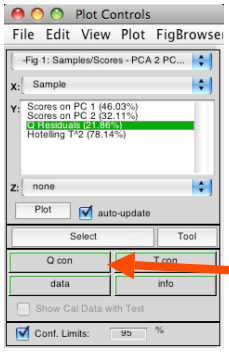
- Contributions to T^2 show how the original variables deviate from the mean within the model

$$T^2_{i,con} = \mathbf{t}_i\lambda^{-1}\mathbf{P}_k^T = \mathbf{x}_i\mathbf{P}_k\lambda^{-1}\mathbf{P}_k^T$$

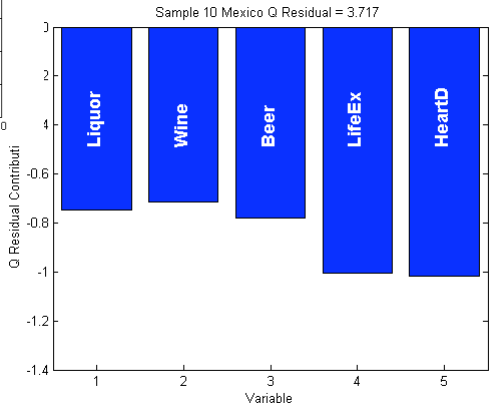
EIGENVECTOR RESEARCH INCORPORATED

Q Residuals for Wine: Q Contributions for Mexico



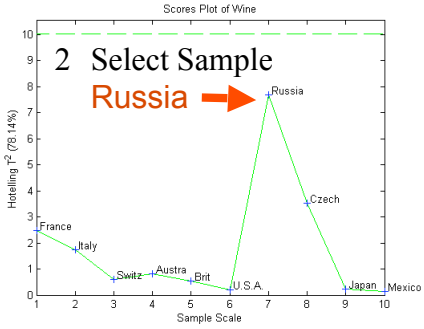


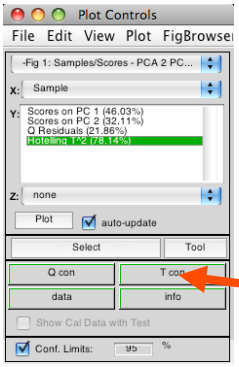
1 Click Q con

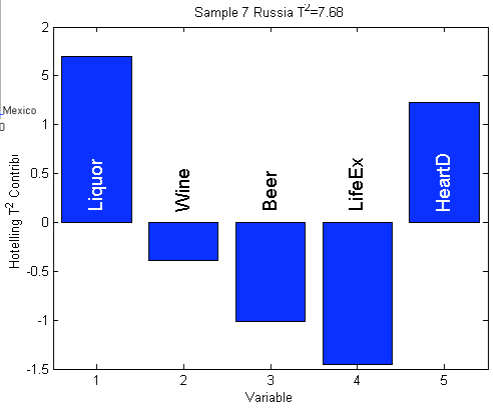


EIGENVECTOR RESEARCH INCORPORATED

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







1 Click T con

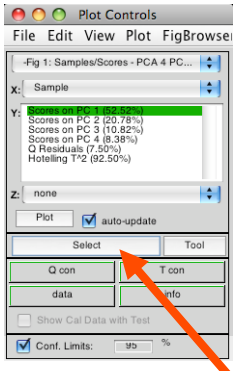
EIGENVECTOR RESEARCH INCORPORATED

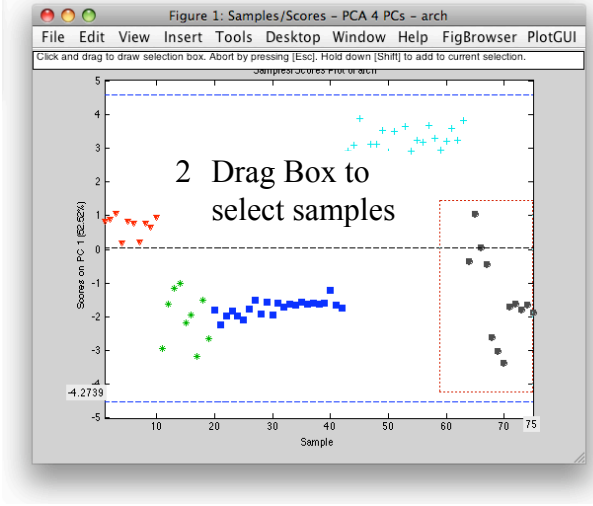
Outliers

- 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

Selecting Samples: ARCH Data

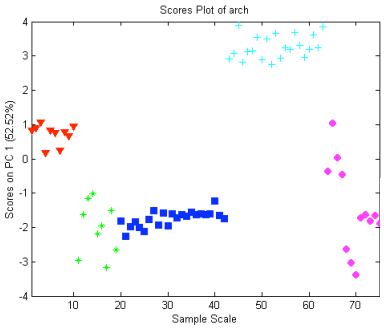


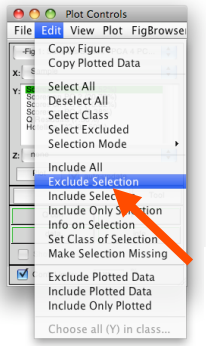


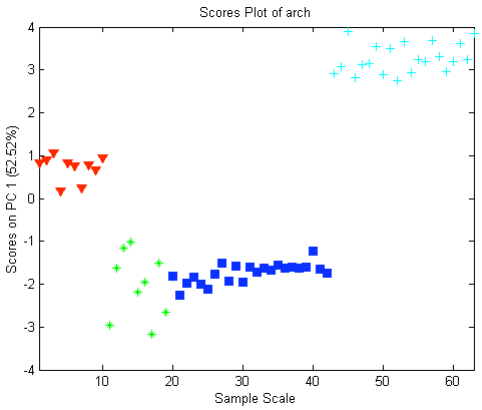
1 Click Select

EIGENVECTOR RESEARCH INCORPORATED

Deleting Samples: ARCH Data



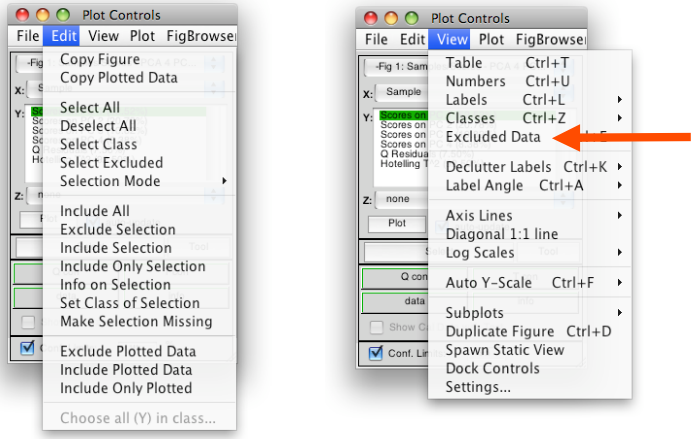




1 Edit menu highlight Exclude Selection

EIGENVECTOR RESEARCH INCORPORATED

Graphically Editing

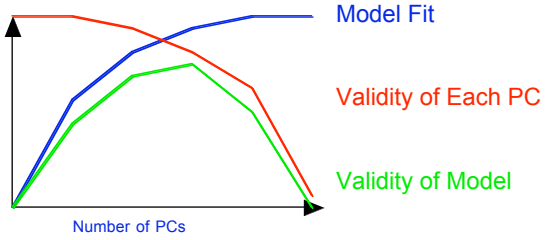


EIGENVECTOR RESEARCH INCORPORATED

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

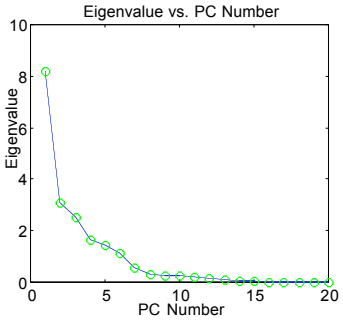
Determining the Number of Principal Components

- 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 $\lambda >$ ~1-2
  - Retain PCs with %variance > noise level
  - Omit PCs that don't make sense!
  - Use cross-validation

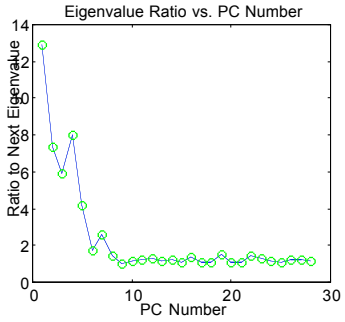
EIGENVECTOR RESEARCH INCORPORATED

Knees and Ratios

Data Set A



Data Set B

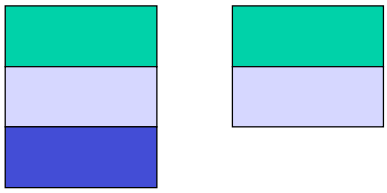


EIGENVECTOR RESEARCH INCORPORATED

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

PCA Cross-validation



Break data into *n* sets (here *n* = 3)

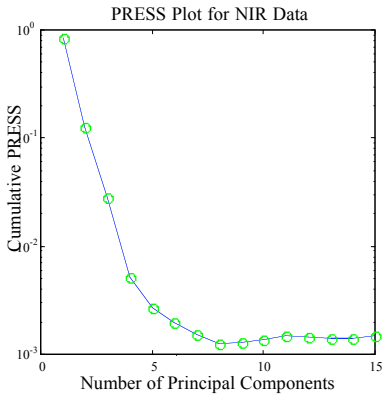
Use all but one set to build PCA model

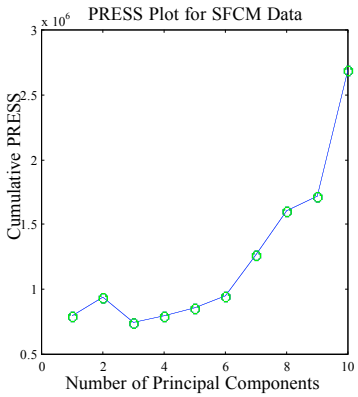
Use loadings and all but one variable to estimate remaining variable

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

Calculate estimation error as function of number of PCs used

Cross-Validation Examples

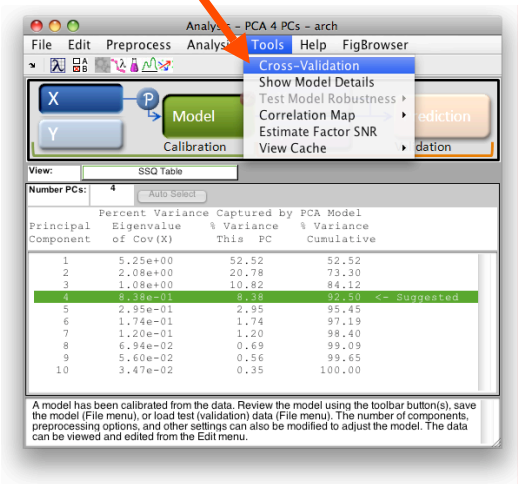




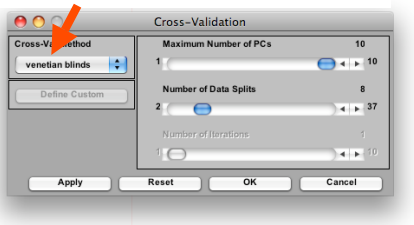
EIGENVECTOR RESEARCH INCORPORATED

Cross-Validation

1 Tools menu highlight Cross-Val



2 Select Cross-validation method



3 Click calc button to perform decomposition and Cross-Validation

4 Click Plot Eigenvalues button to plot Eigenvalues and RMSECV

EIGENVECTOR RESEARCH INCORPORATED

PCA Application to New Data

- center new data to the mean of the calibration data

  $\mathbf{X}_c = \mathbf{X} - \mathbf{1}\mathbf{x}_{mean}$
- scale the centered data using standard deviations of cal data

  $\mathbf{X}_s = \mathbf{X}_c \,./\, \mathbf{1}\mathbf{x}_{std}$
- project centered and scaled data onto loadings to get new scores

  $\mathbf{T}_{new} = \mathbf{X}_s\mathbf{P}_k$
- calculate new residuals

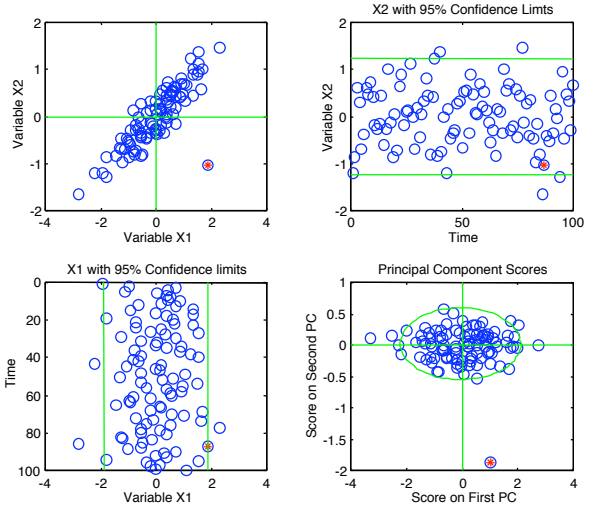
  $\mathbf{E}_{new} = \mathbf{X}_s - \mathbf{T}_{new}\mathbf{P}_k^T = \mathbf{X}_s(\mathbf{I} - \mathbf{P}\mathbf{P}^T)$
- calculate new Q residuals

  $\mathbf{Q}_{new} = \text{diag}(\mathbf{E}_{new}\mathbf{E}_{new}^T)$
- calculate new T^2 values

  $\mathbf{T}^2_{new} = \mathbf{T}_{new}\lambda^{-1}\mathbf{T}_{new}^T = \mathbf{X}_s\mathbf{P}_k\lambda^{-1}\mathbf{P}_k^T\mathbf{X}_s^T$
- compare $\mathbf{T}_{new}$, $\mathbf{E}_{new}$, $\mathbf{Q}_{new}$ and $\mathbf{T}^2_{new}$ to previously determined limits

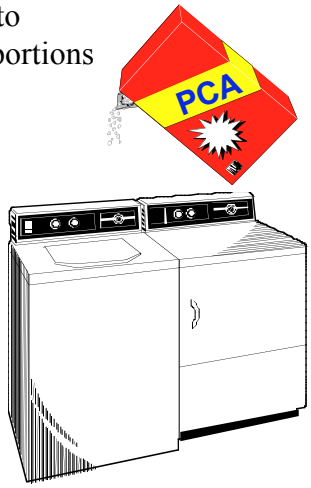
PCA Based MSPC

PCA scores can be combined with traditional statistical process control tools:
Shewart
Range
X-bar
CUSUM, etc...
Result is *Multivariate Statistical Process Control* (MSPC)



Dirty T-Shirt Analogy

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



EIGENVECTOR RESEARCH INCORPORATED