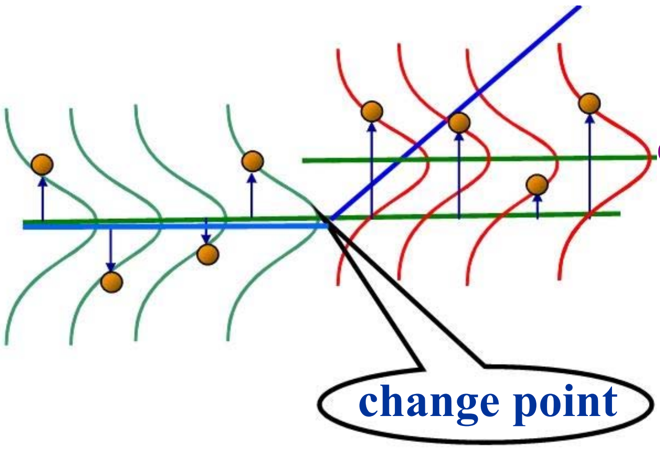
Statistical Approach to Process Monitoring



陳榮輝 台灣中原大學化工系
計算機程序系統工程研究室
e-mail : jason@wavenet.cycu.edu.tw
http://wavenet.cycu.edu.tw/

Outline

- Control in a Nonstationary World
- Univariate Charts: USPC
- Geometry of Principal Component Analysis
- Algebraic Definition of PCs
- PCA: MATLAB Codes
- Determining Number of Loading Vectors
- Batch Process Monitoring
- Probability Based PCA
- Extension Research

Control in a Nonstationary World

The stable stationary state is an unnatural one.

"If left to themselves machines do not stay adjusted, components wear out and managers and operators miscommunication and change jobs""

Box G. and Luceño A.

Process Monitoring

Process Adjusting

Left to itself the entropy of any system can never decrease.

The second law of thermodynamics

Modern Control Room

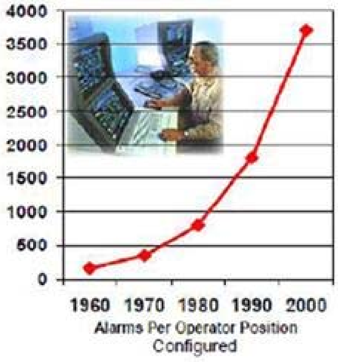
- Today, hardware and software advances has made it easy to add alarms at minimal cost
- Large increase in the quantity of alarms
- Reducing the quality and efficiency of alarms

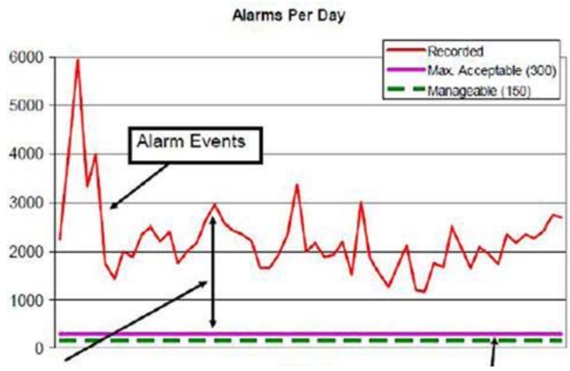
Operators Receive too Many Alarms

Industry standard and typical actual values of alarms

	EEMUA[3] standard	Oil and gas industry	Petrochemical industry
avg alarms/hr	6	36	54
avg standing alarms	9	50	100
peak alarms/hr	60	1320	1080

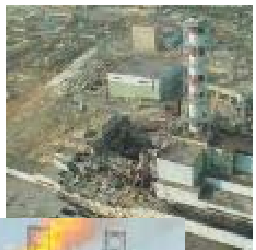
[3]Engineering Equipment and Materials User Associations





History of Repeated Accidents is Over and Again

- Chernobyl, Ukraine, 1986 (more than 4000 direct and indirect deaths)
- Piper Alpha Oil Rig, North Sea, 1988 (167 deaths)
- Phillips 66 Complex, Texas, 1989 (23 deaths)
- BP Refinery, Texas City, 2005 (15 deaths)
- Ammonium nitrate explosions, Monclove, Mexico (2007)
- Cement failure in offshore oil rig
  - » Montana rig, East Timor sea (2009)
  - » Deepwater Horizon, Gulf of Mexico (2010)
- Fertilizer Plant Explosion, Texas (2013) (14 deaths)



The repetition of accidents tells us that we need a new look into control systems in the operating plant.

What Is "Process Monitoring"?

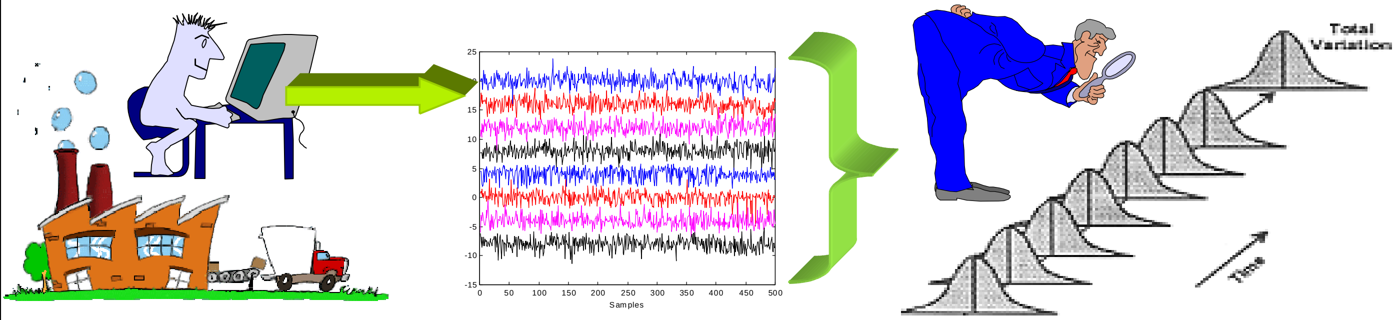
Example: Health Detection

A doctor gathers information in stages to give a diagnosis:
Var 1: Temperature
Var 2: Blood pressure
Var 3: Pain location

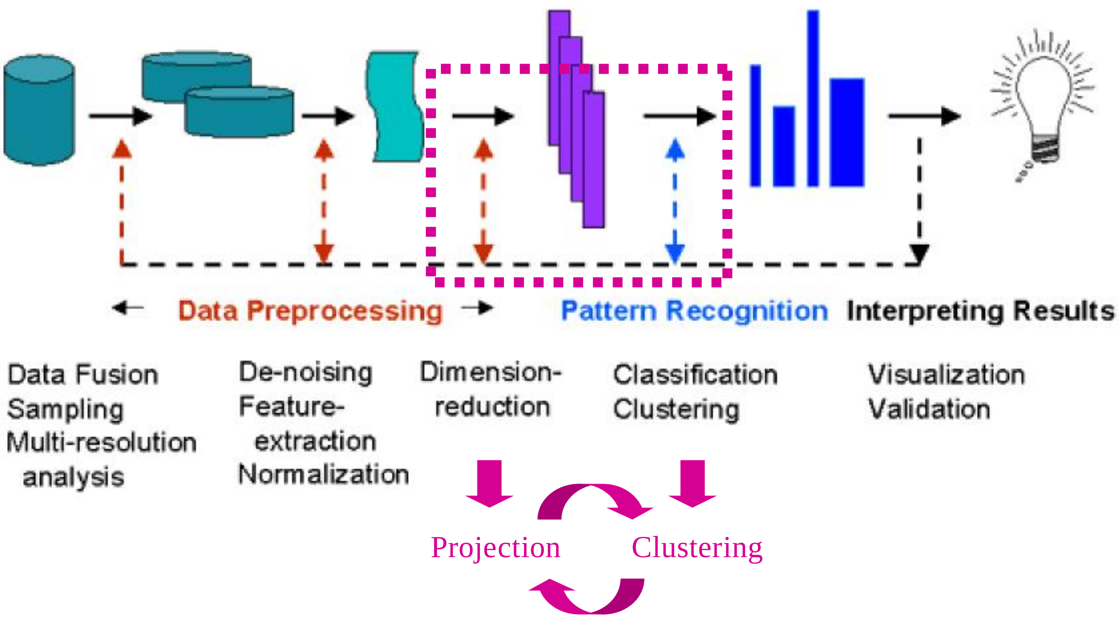
… further information might be useless

Var 4: Hair color

Let Data Talk



Big Picture of Data Mining



Quality Improvement and Statistics

- **Definitions of Quality**

  Quality means fitness for use

  - quality of design
  - quality of conformance

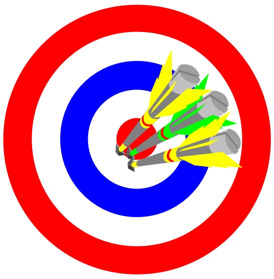
  Quality is inversely proportional to variability.

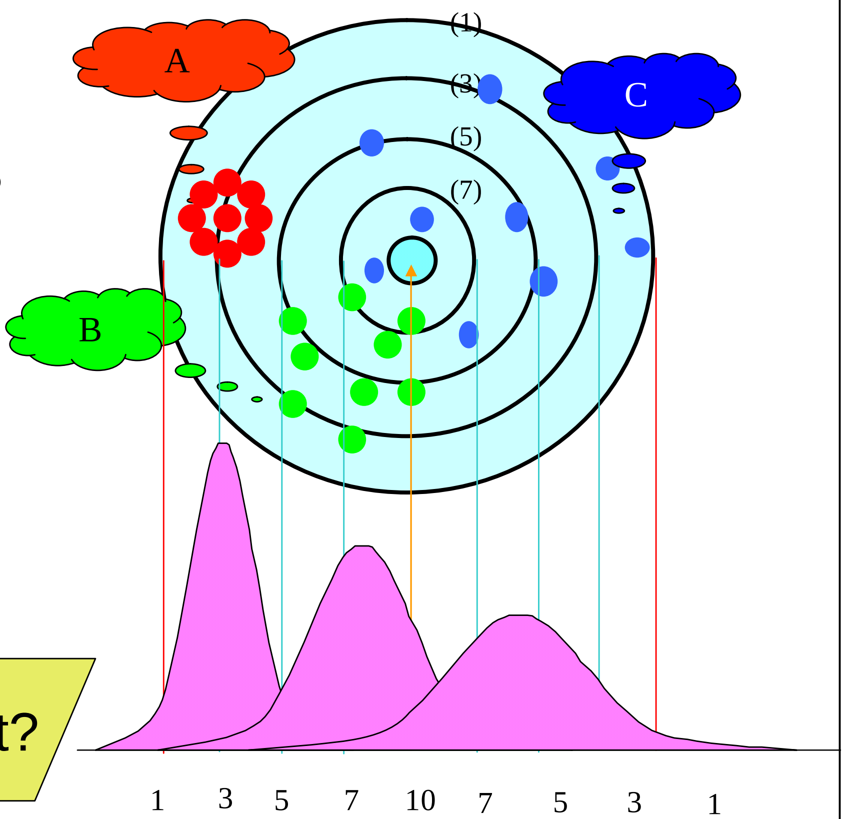
SPC has its origin in the 1920s. (Dr. Shewhart, Bell Lab.) The methodology is widely applied after World War II.

Statistical process control is a collection of tools that when used together can result in process stability and variance reduction.

Have you ever…

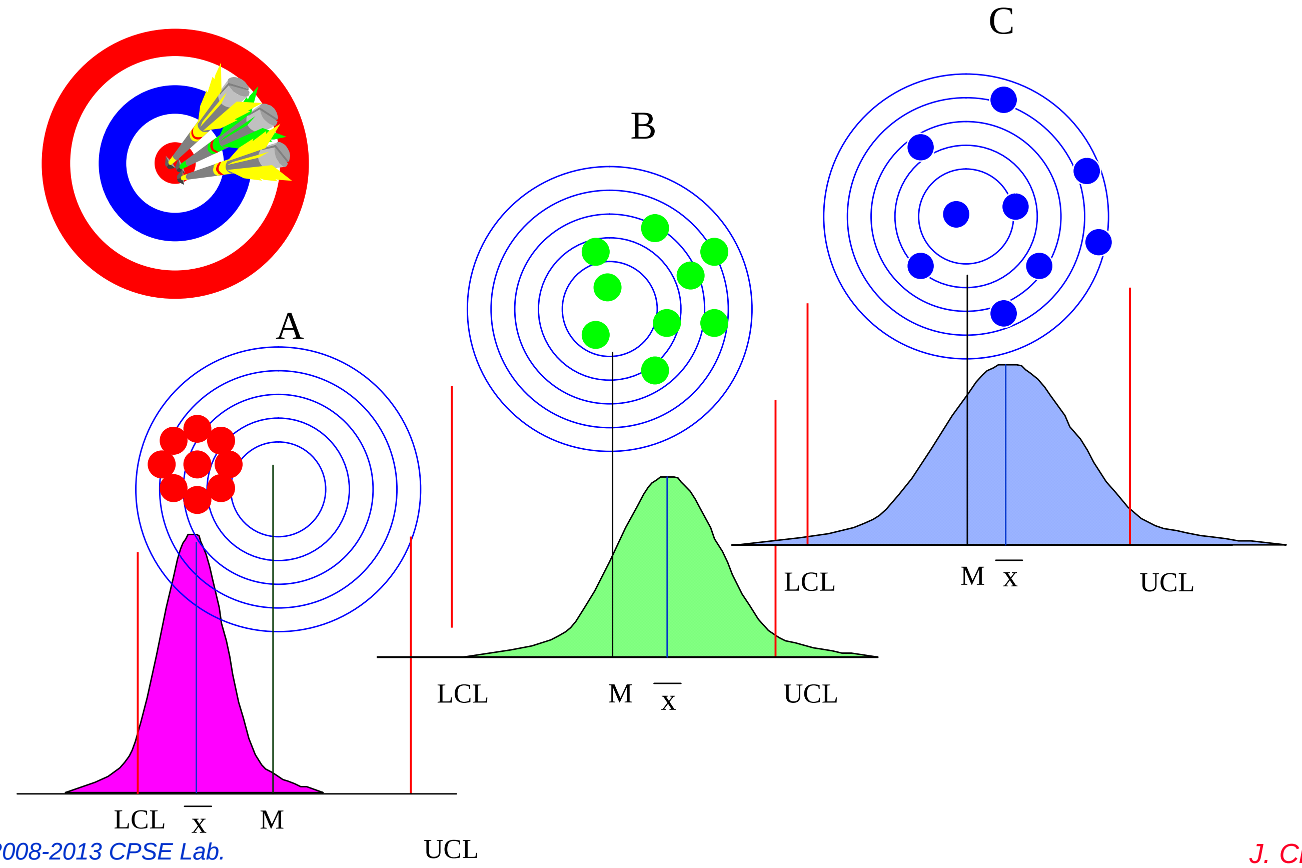
- Shot a rifle?
- Played darts?
- Shot a round of golf?
- Played basketball?





Who is the better shot?

Have you ever…

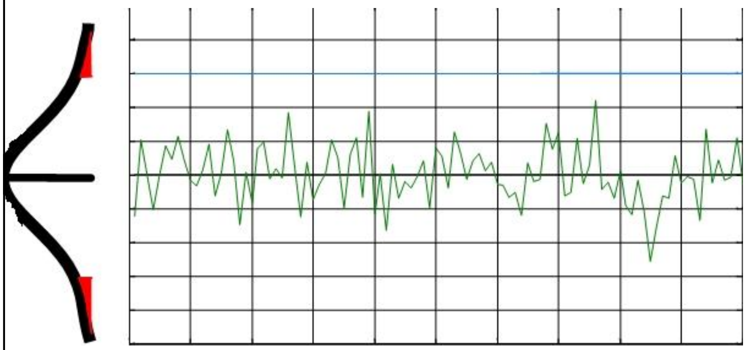


Control Charts

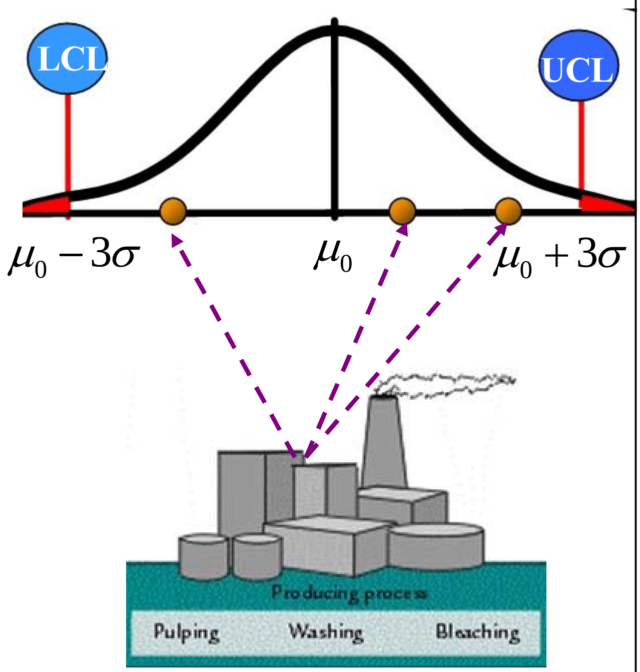
Basic Principles

- A process that is operating with only ***chance causes of variation*** present is said to be in statistical control.
- A process that is operating in the presence of ***assignable causes*** is said to be out of control.
- The eventual goal of SPC is the ***elimination of variability*** in the process.

Univariate Charts: USPC



$$P(\mu_0 - 3\sigma < x < \mu_0 + 3\sigma) = 0.9973$$

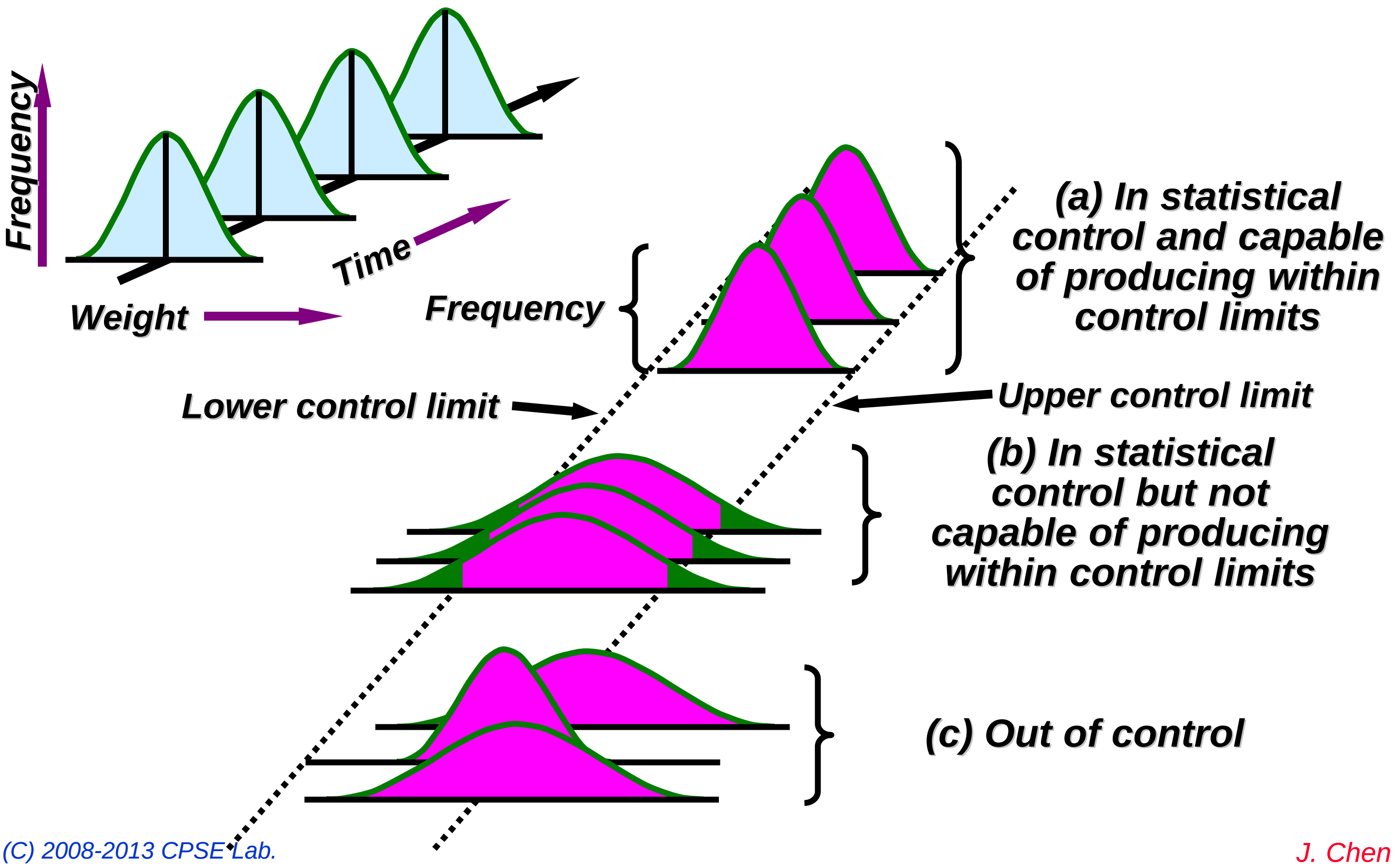


Hypothesis Test

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

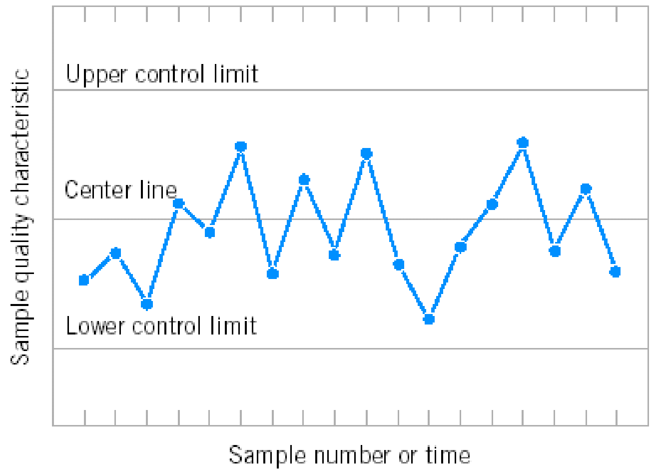
Variability



Control Charts

Basic Principles

A typical control chart has control limits set at values such that if the process is in control, nearly all points will lie within the upper control limit (UCL) and the lower control limit (LCL).



A typical control chart.

Control Charts

Basic Principles

General Model for a Control Chart

Let W be a sample statistic that measures some quality characteristic of interest, and suppose that the mean of W is μ_W and the standard deviation of W is σ_W.[3] Then the **center line** (CL) the **upper control limit** (UCL) and the **lower control limit** (LCL) become

$$UCL = \mu_W + k\sigma_W$$
$$CL = \mu_W$$
$$LCL = \mu_W - k\sigma_W \qquad (8\text{-}1)$$

where k is the "distance" of the control limits from the center line, expressed in standard deviation units.

Control Charts

Design of a Control Chart

Suppose we have a process that we assume the true process mean is μ = 74 and the process standard deviation is σ = 0.01. Samples of size 5 are taken giving a standard deviation of the sample average, average standard deviation is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.01}{\sqrt{5}} = 0.0045$$

- Control limits can be set at 3 standard deviations from the mean in both directions.
- "3-Sigma Control Limits"

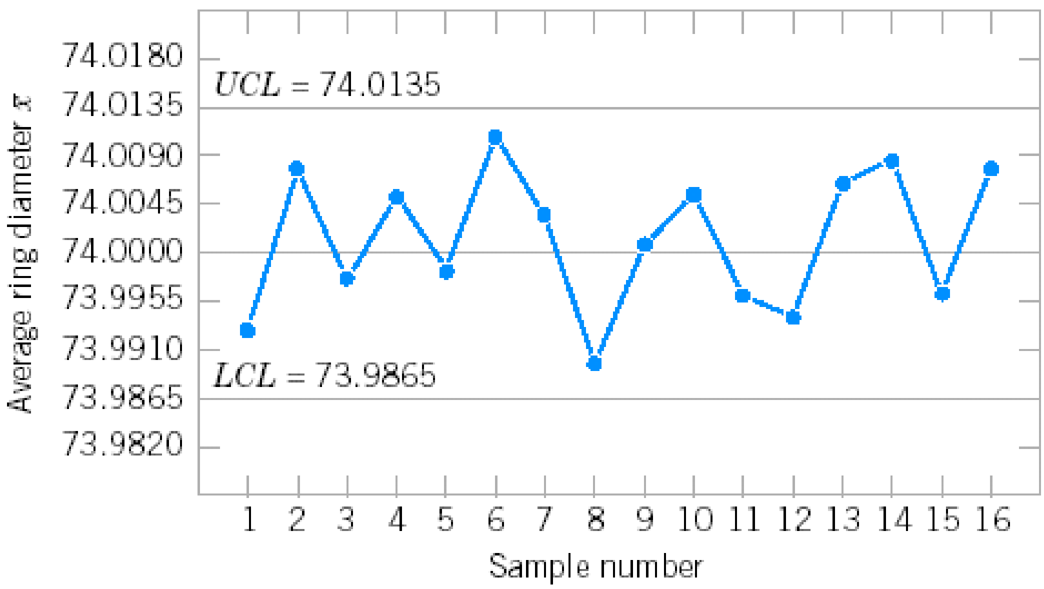
UCL = 74 + 3(0.0045) = 74.0135

CL= 74

LCL = 74 - 3(0.0045) = 73.9865

Control Charts

Design of a Control Chart

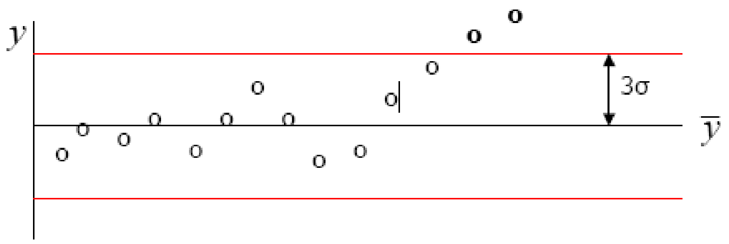


$\overline{X}$ control chart for piston ring diameter.

Univariate Charts: USPC

- Establish a permanent information system over the process evolution.
  - » **Detect** the anomalies at an early stage (special causes).
  - » Help to **identify** the causes of the anomalies.
  - » **Eliminate the anomalies** and prevent their reappearance. (Or on the contrary, incorporate them to the process if they improve its performance.)

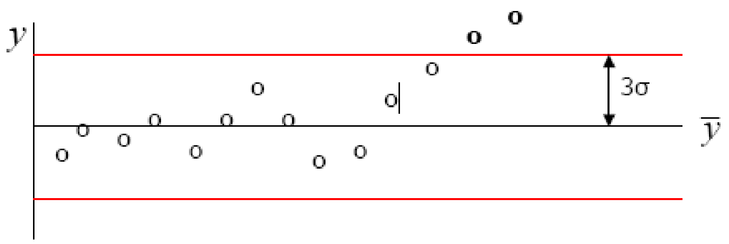
detect and take action



Control Charts: Fault Patterns

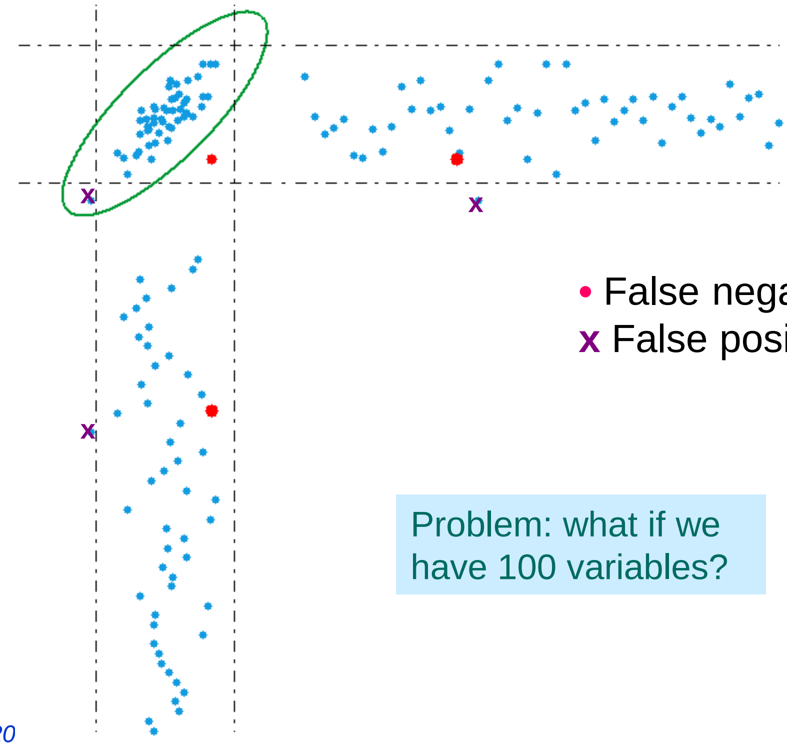
The **Western Electric rules** would signal that the process is out of control if either

1. One point plots outside three-sigma control limits.
2. Two out of three consecutive points plot beyond a two-sigma limit.
3. Four out of five consecutive points plot at a distance of one sigma or beyond from the center line.
4. Eight consecutive points plot on one side of the center line.



USPC: Limitation

- Quality is often a multivariate property.
- Univariate control charts (ignores correlation)

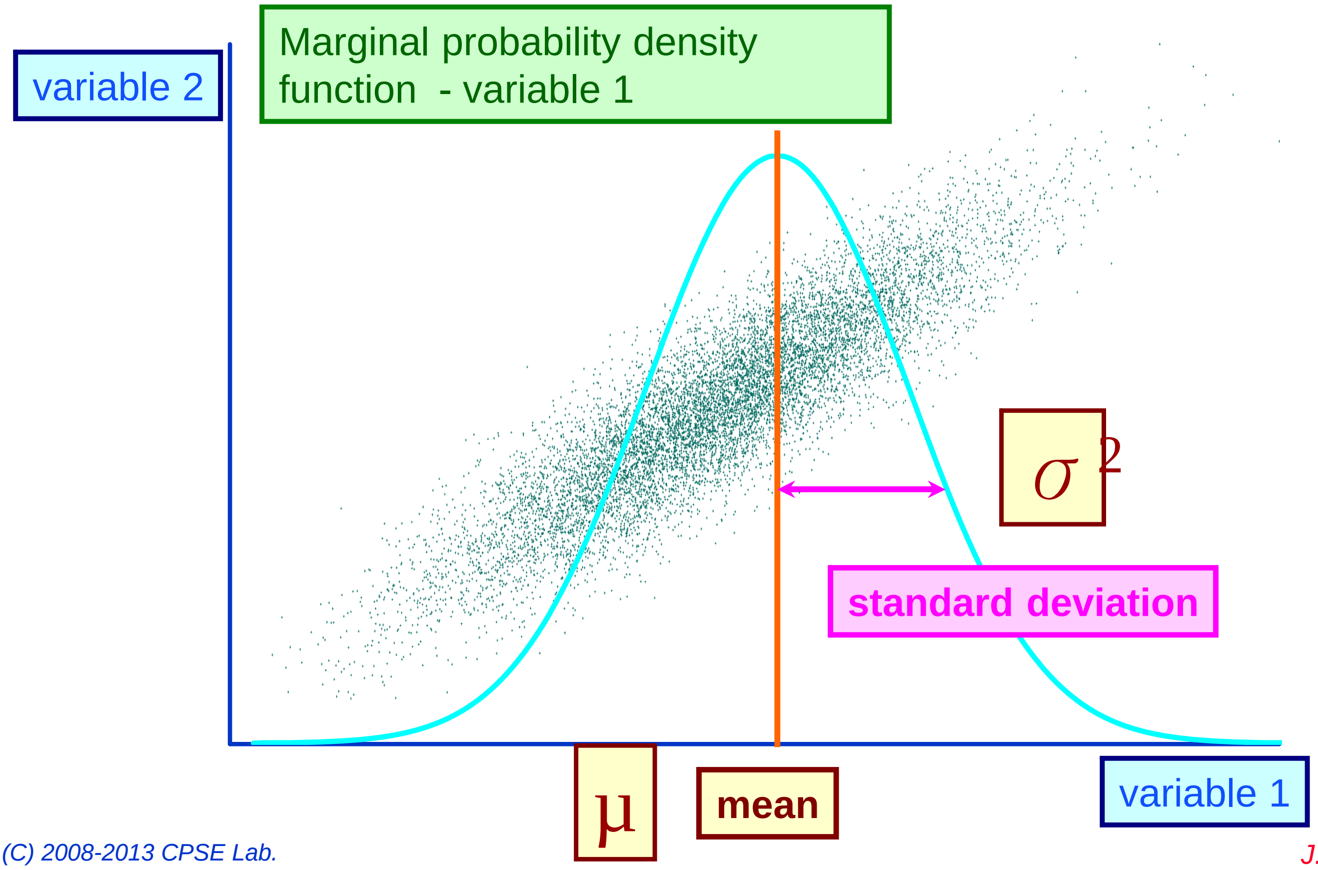


Simple example with only 2 quality variables

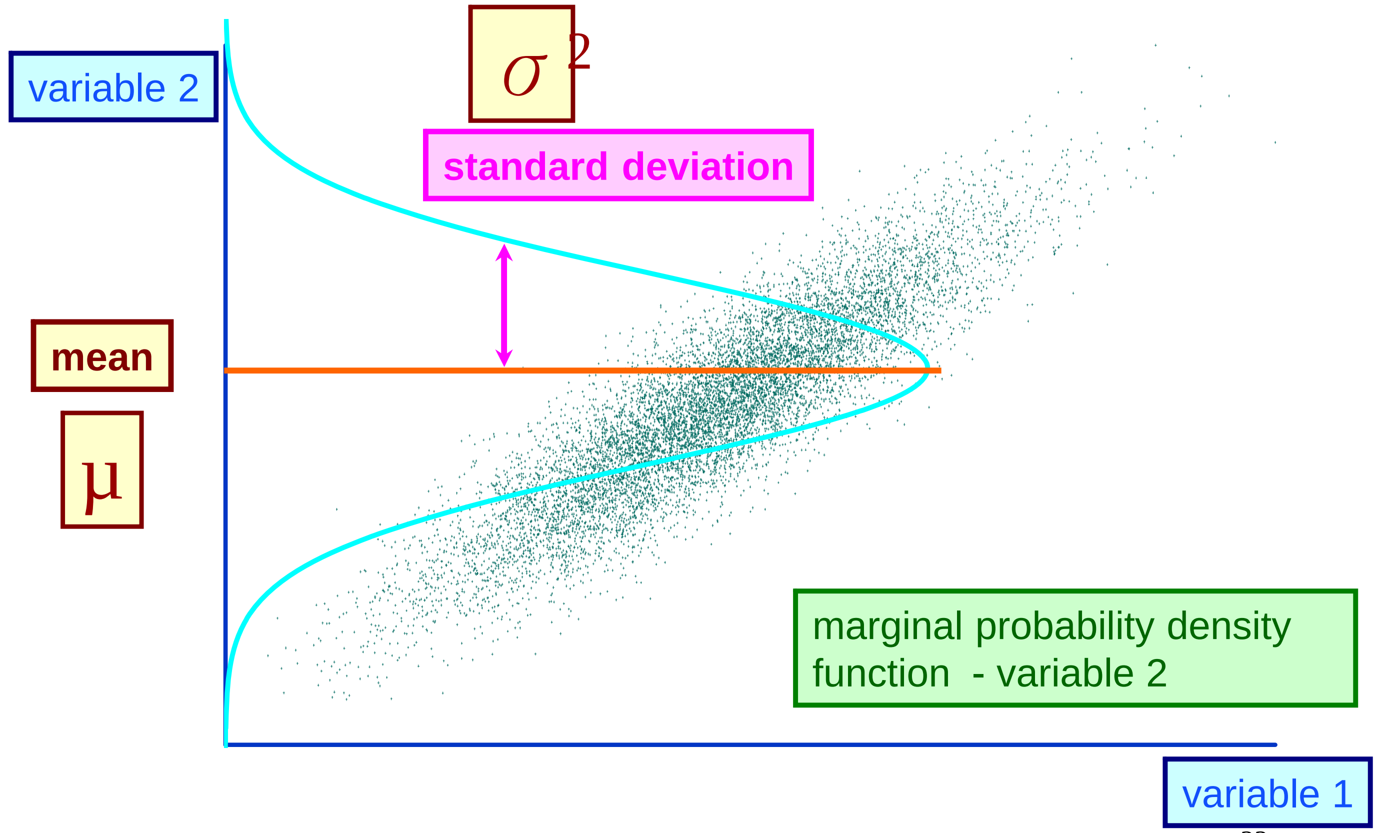
• False negative
x False positive (false alarm)

Problem: what if we have 100 variables?

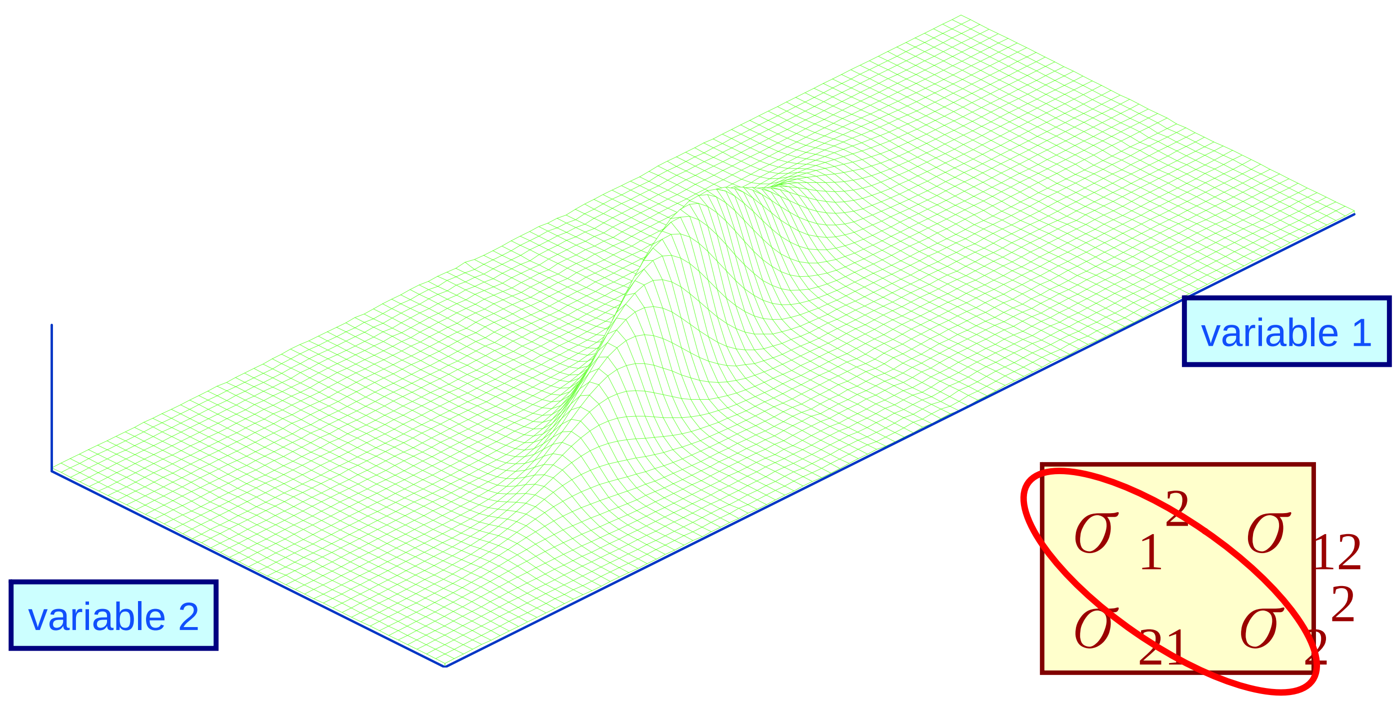
A Single Random Variable with a Multi-Normal Distribution



A Single Random Variable with a Multi-Normal Distribution

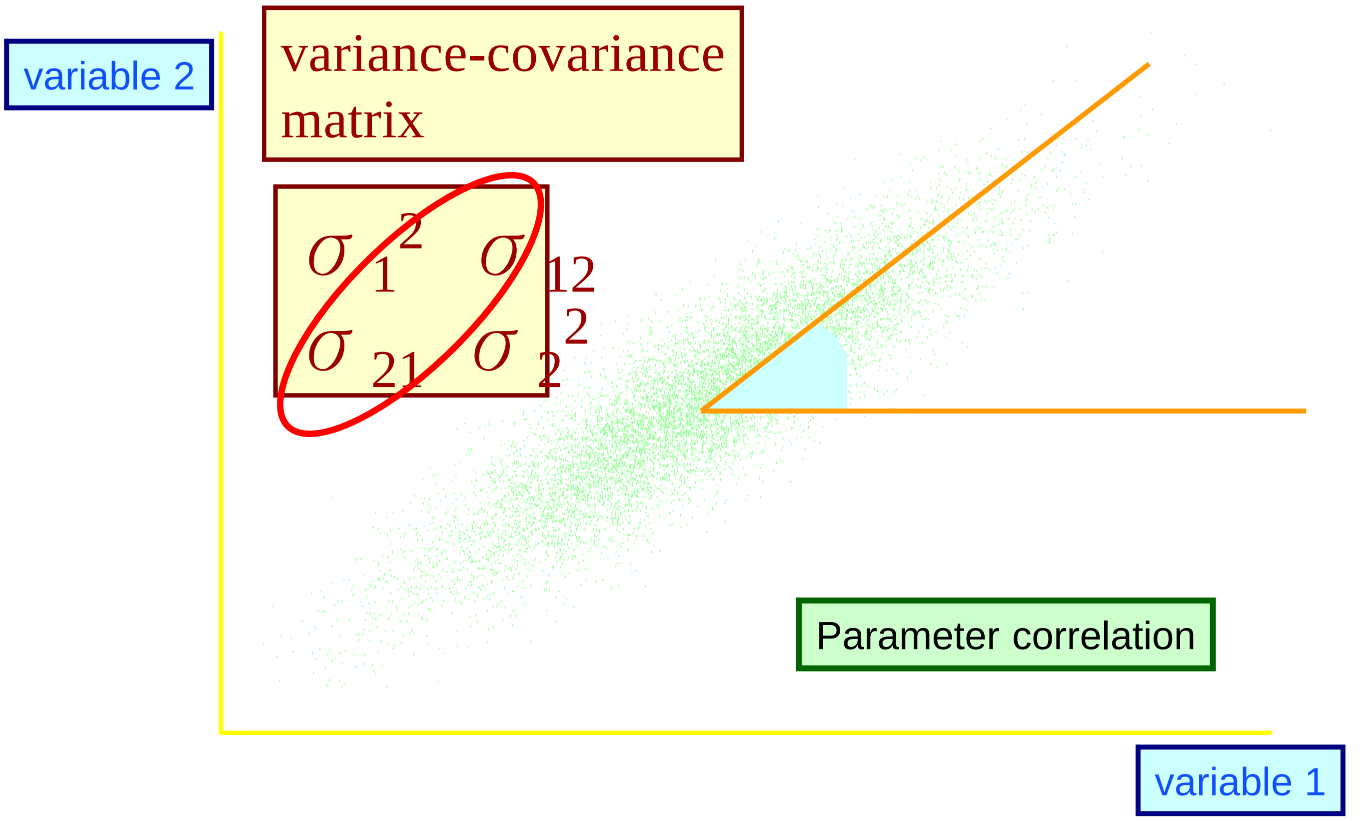


Two-Dimensional Probability Density Function



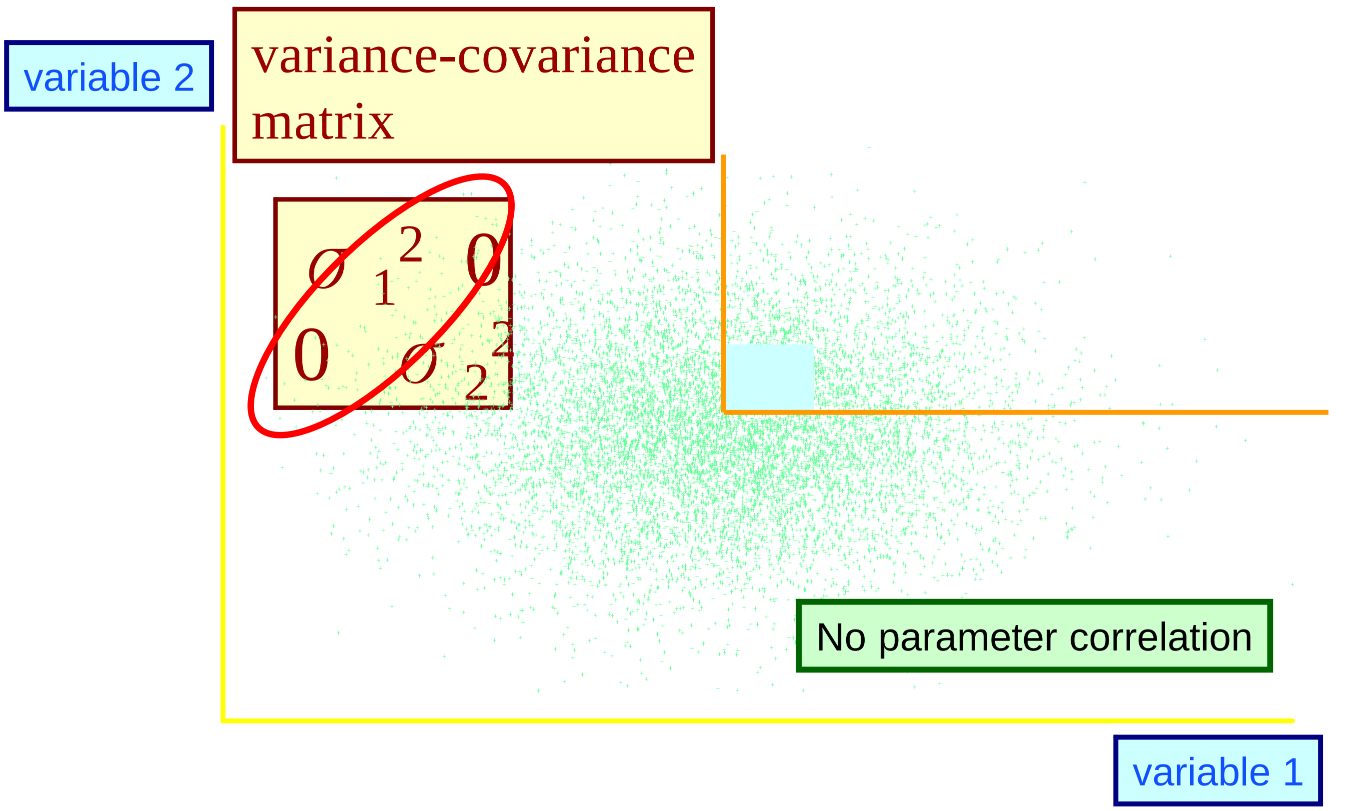
$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

A Two-Dimensional Random Variable with a Multi-normal Distribution



$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

A Two-Dimensional Random Variable with a Multi-normal Distribution



$$\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

Data Tables (Matrices)

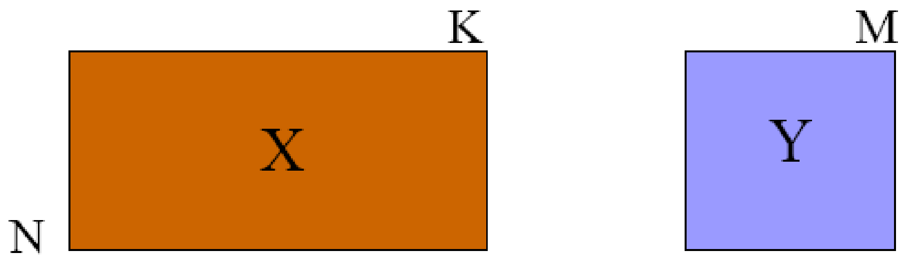
Example of a multivariate data set:
A polymerization process, N=820 observations, K=160 variables.

Data set - allied

	1	2	3	4	5	6	7	8	9	10	11	1
1			1	2	3	4	5	6	7	8	9	1
2	Nun	Name	A7-TOT-RA	FI-7524	CL21FD2H	CL21FD3C	CL21FDBS	CL21FDCL	CL21FDFR	CL21FDPH	APT40FDA	APT4
3	5	1997-01-05	108.969	141.445	1382.339	1477.49	480.507	205.915	181.471	4.91646	25.841	62
4	6	1997-01-06	107.523	132.548	1352.925	1517.79	458.897	221.667	215.75	5.01667	21.012	48
5	7	1997-01-07	101.216	124.173	1608.593	1615.98	379.442	209.667	179.25	5.0875	14.396	46
6	8	1997-01-08	102.622	133.643	1539.103	1543.91	423.319	190.792	130.292	5.40417	12.579	58
7	9	1997-01-09	99.397	126.341	1515.025	1677.23	469.441	190.042	171.25	5.09583	14.858	51
8	10	1997-01-10	105.905	127.984	1448.984	1527.06	436.526	205.875	156.875	5.3	9.55	60
9	11	1997-01-11	100.526	128.936	1554.426	1620.04	469.025	187.167	157.583	5.35833	13.012	61
10	12	1997-01-12	99.083	118.565	1357.026	1534.48	477.099	188.5	170.708	5.05	29	
11	13	1997-01-13	75.488	133.783	1312.41	1540.65	486.51	199.667	152.625	4.825	28.142	58
12	14	1997-01-14	101.859	129.431	1342.626	1570.43	453.163	183.125	173.125	5.13333	25.387	74
13	15	1997-01-15	91.129	125.117		1580.18	393.976	194.375	205.292	5.3625	8.275	66
14	16	1997-01-16	99.541	113.348	1153.555	1566	414.865	209.292	184.208	5.02917	13.004	42
15	17	1997-01-17	111.868	134.914	1135.873	1544.31	394.416	238.708	203.542	5.10417	19.817	59
16	18	1997-01-18	105.881	135.835	1752.73	1655.11	372.792	223.292	156.208	5.14583	20.133	7
17	19	1997-01-19	104.64	138.174	1271.098	1648.07	392.613	227.375	158.208	4.85833	12.813	73
18	20	1997-01-20	103.402	142.803	1219.477	1598.88	402.706	238.083	196.917	5.1	2.888	69

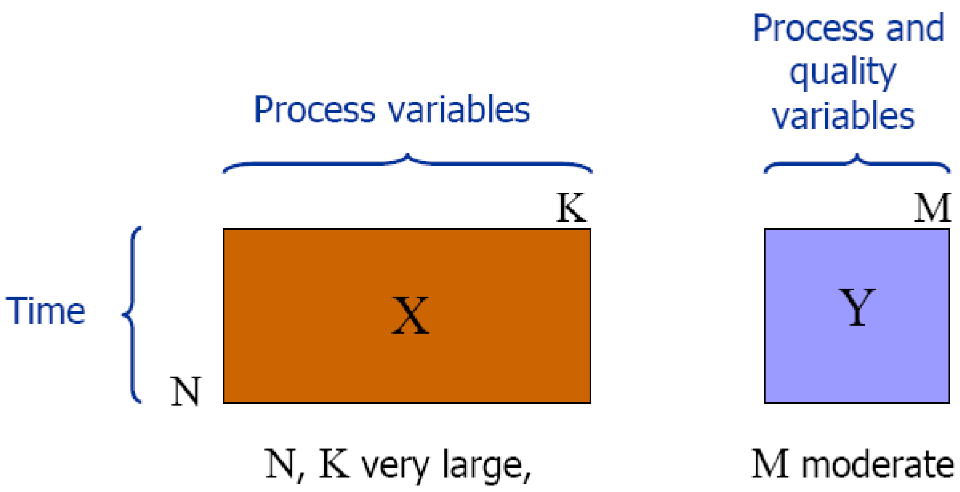
How the Nature of Data Has Changed

- Computers and automated measurement systems have lead to exponential explosion in data collected
- Large data sets (dimensions N, K, M very large)



Examples: Process Data

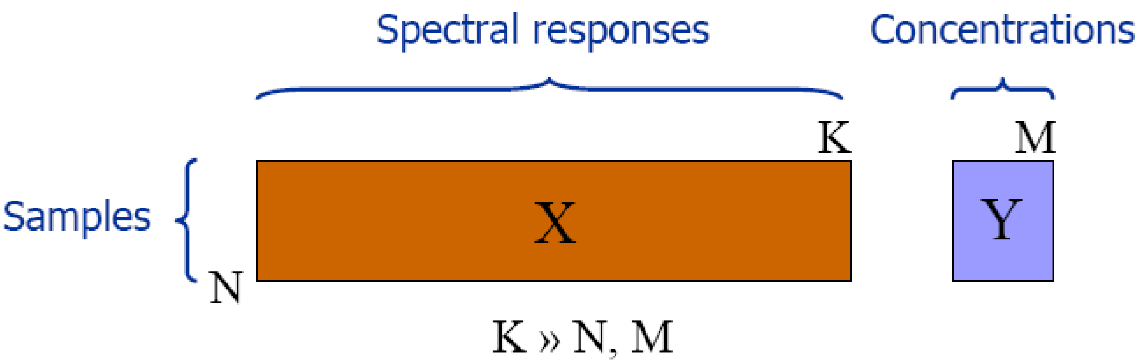
Driver: Process computers



eg: Petroleum process (one section)
N~100,000; K~500; M~20

Examples: Analytical Labs (Calibration)

Driver: New instrumentation



eg: N~30; K~2000; M~5

More Problems ... and More Data

- Traditional SPC chart monitors single variables, often just the quality variables, Y.
- For SPC why not use process variables X?
- Why use the X-variables ?
  - » Many more X variables available than Y
    - – Use easily available process measurements to build a soft sensor. Temperatures, pressures, flows, levels, etc.
  - » X's are on-line (real-time), Y's are often off-line (lab)
  - » X's are more frequent, and often more precise
  - » Fingerprints of faults are in the X's
  - » More faults may be detected with the X's, than with Y's

Geometry of Principal Component Analysis

	x1		x2	
Obs	Original	Corrected	Original	Corrected
1	16	8	8	5
2	12	4	10	7
3	13	5	6	3
4	11	3	2	-1
5	10	2	8	5
6	9	1	-1	-4
7	8	0	4	1
8	7	-1	6	3
9	5	-3	-3	-6
10	3	-5	-1	-4
11	2	-6	-3	-6
12	0	-8	0	-3
Mean	8	0	3	0
Var	23.091	23.091	21.091	21.091

$$\frac{1}{12}\sum_{m=1}^{12}\left(x_{m,1}-\bar{x}_1\right)^2 \qquad \frac{1}{12}\sum_{m=1}^{12}\left(x_{m,2}-\bar{x}_2\right)^2$$

- 12 observations and 2 variables
- Center and scale the variables to have equal basis.
- Total variances of variables are 44.182 (i.e. 23.091+21.091).
- The percentages of the total variance accounted for x1 and x2 are 52.26% and 47.74%.
- Correlation coefficient is 0.746.

$$r_{i,j=}\frac{1}{12}\sum_{m=1}^{12}\left(x_{m,i}-\bar{x}_i\right)\left(x_{m,j}-\bar{x}_j\right)$$

$$\begin{bmatrix} 1 & 0.746 \\ 0.746 & 1 \end{bmatrix}$$

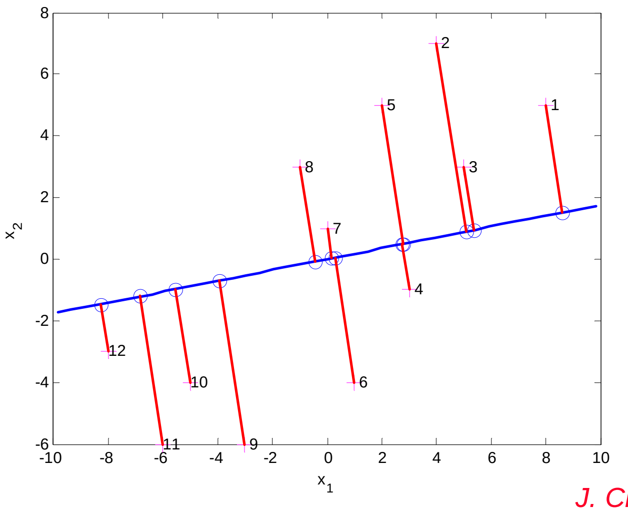
Geometry of Principal Component Analysis

Obs	x1	x2	x1*
1	8	5	8.747
2	4	7	5.155
3	5	3	5.445
4	3	-1	2.781
5	2	5	2.838
6	1	-4	0.29
7	0	1	0.174
8	-1	3	-0.464
9	-3	-6	-3.996
10	-5	-4	-5.619
11	-6	-6	-6.951
12	-8	-3	-8.399
Mean	0	0	0
Var	23.091	21.091	28.659

»Lec5PC1

- New variable x1* for a rotation of 10 degree
- The projection of the observations onto x1 gives the coordinate of the observation with respect to x1*.

$$x_1^* = x_1 \cos\theta + x_2 \sin\theta$$



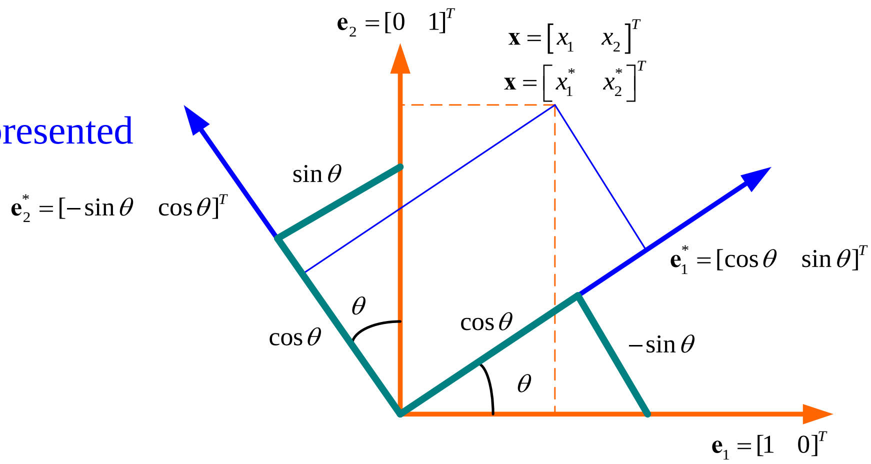
Representation Points with Respect to Different Axes

$$\mathbf{x} = x_1\mathbf{e}_1^T + x_2\mathbf{e}_2^T$$

The vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ can be represented with respect to $\mathbf{e}_1^*$ and $\mathbf{e}_2^*$ as

$$\mathbf{e}_1^T = \cos\theta\mathbf{e}_1^{*T} - \sin\theta\mathbf{e}_2^{*T}$$

$$\mathbf{e}_2^T = \sin\theta\mathbf{e}_1^{*T} + \cos\theta\mathbf{e}_2^{*T}$$



$$\mathbf{x} = (x_1\cos\theta + x_2\sin\theta)\mathbf{e}_1^{*T} + (-x_1\sin\theta + x_2\cos\theta)\mathbf{e}_2^{*T}$$

$$\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

The coordinates of **x** with respect to the new axes are linear combinations of the coordinates with respect to the old axes.

Variance Accounted for New Variables x1 for Various New Axes

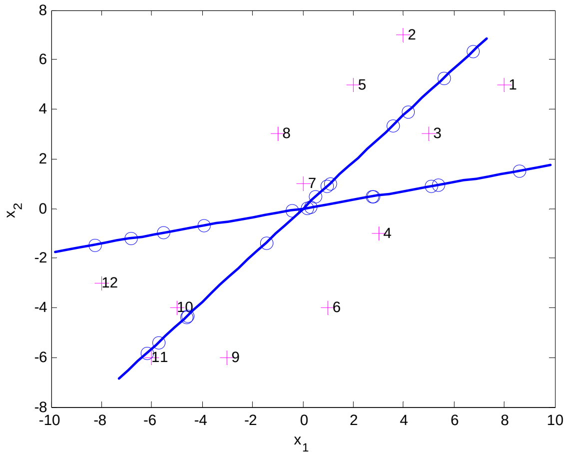
Angle with x1	Total Var.	Var. of x1*	%
0	44.182	23.091	52.263
10	44.182	28.659	64.866
20	44.182	33.434	75.676
30	44.182	36.841	83.387
40	44.182	38.469	87.072
43.261	44.182	38.576	87.312
50	44.182	38.122	86.282
60	44.182	35.841	81.117
70	44.182	31.902	72.195
80	44.182	26.779	60.597
90	44.182	21.091	47.772

- The percentage of the total variance accounted for x1 increases as the angle between x1* and x1 increases and then, after a certain maximum value, the variance accounted for x1* begins to decrease.
- There is one and only one new axis that results in a new variable accounting for the maximum variance in the data.

Variance Accounted for New Variables x1 for New Axes Making an Angle of 43.261

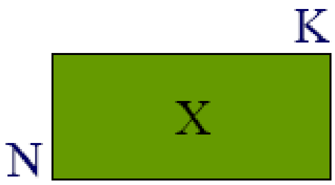
OBS	x1	x2	x1*	x2*
1	8	5	9.253	-1.841
2	4	7	7.71	2.356
3	5	3	5.697	-1.242
4	3	-1	1.499	-2.784
5	2	5	4.883	2.271
6	1	-4	-2.013	-3.598
7	0	1	0.685	0.728
8	-1	3	1.328	2.87
9	-3	-6	-6.297	-2.313
10	-5	-4	-6.382	0.514
11	-6	-6	-8.481	-0.257
12	-8	-3	-7.882	3.298
Mean	0	0	0	0
Var.	23.091	21.091	38.572	5.606

- The percentage of the total variance accounted for x1* is about 87.31% (38.576/44.182) of the total variance in the data.
- The second axis accounts for the maximum of the variance that is not accounted for x1*.



»Lec5PC2

Starting Point: Problem ⇒ Data Table X (N x K)



- Data set = table (matrix)
  N objects and K variables
- Often many variables - large K
- Often few observations (K>>N) or many of both (N and K large)
- Missing data
- Poor data: clusters and collinearity

- **Objects (rows):**
  - » Analytical samples
  - » Process time points
  - » Trials (experim. runs)
  - » Chemical compounds, …
- **Variables (columns):**
  - » Sensors (T, P, flow, pH, conc.,…)
  - » Chromatographic Peaks (HPLC, GC, Electrophoresis, …)
  - » Laboratory assays

Data Tables (Matrices)

Example of a multivariate data set:
A polymerization process, N=820 observations, K=160 variables.

Data set - allied

	1	2	3	4	5	6	7	8	9	10	11	12
1			1	2	3	4	5	6	7	8	9	1(
2	Num	Name	A7-TOT-RA	FI-7524	CL21FD2H	CL21FD3C	CL21FDBS	CL21FDCL	CL21FDFR	CL21FDPH	APT40FDA	APT4
3	5	1997-01-05	108.969	141.445	1382.339	1477.49	480.507	205.915	181.471	4.91646	25.841	62
4	6	1997-01-06	107.523	132.548	1352.925	1517.79	458.897	221.667	215.75	5.01667	21.012	48
5	7	1997-01-07	101.216	124.173	1608.593	1615.98	379.442	209.667	179.25	5.0875	14.396	46
6	8	1997-01-08	102.622	133.643	1539.103	1543.91	423.319	190.792	130.292	5.40417	12.579	58
7	9	1997-01-09	99.397	126.341	1515.025	1677.23	469.441	190.042	171.25	5.09583	14.858	51
8	10	1997-01-10	105.905	127.984	1448.984	1527.06	436.526	205.875	156.875	5.3	9.55	60
9	11	1997-01-11	100.526	128.936	1554.426	1620.04	469.025	187.167	157.583	5.35833	13.012	61
10	12	1997-01-12	99.083	118.565	1357.026	1534.48	477.099	188.5	170.708	5.05	29	
11	13	1997-01-13	75.488	133.783	1312.41	1540.65	486.51	199.667	152.625	4.825	28.142	58
12	14	1997-01-14	101.859	129.431	1342.626	1570.43	453.163	183.125	173.125	5.13333	25.387	74
13	15	1997-01-15	91.129	125.117		1580.18	393.976	194.375	205.292	5.3625	8.275	66
14	16	1997-01-16	99.541	113.348	1153.555	1566	414.865	209.292	184.208	5.02917	13.004	42
15	17	1997-01-17	111.868	134.914	1135.873	1544.31	394.416	238.708	203.542	5.10417	19.817	59
16	18	1997-01-18	105.881	135.835	1752.73	1655.11	372.792	223.292	156.208	5.14583	20.133	7(
17	19	1997-01-19	104.64	138.174	1271.098	1648.07	392.613	227.375	158.208	4.85833	12.813	73
18	20	1997-01-20	103.402	142.803	1219.477	1598.88	402.706	238.083	196.917	5.1	2.888	69

Data Preprocessing

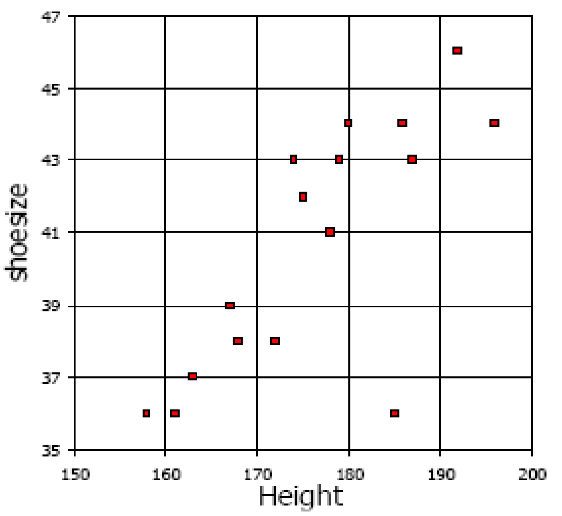
- **Centering** = subtracting column averages,
  ⇨ columns that vary around zero

- **Scaling the variables** = usually dividing columns by their standard deviation (i.e. scaling all variables to unit variance).

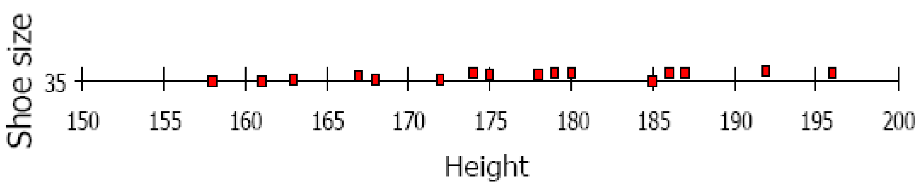
Example of Scatter Plots with Two Scalings

i	Height	Shoe size
1	180	44
2	161	36
3	158	36
4	175	42
5	174	43
6	167	39
7	172	38
8	196	44
9	192	46
10	163	37
11	185	36
12	187	43
13	186	44
14	178	41
15	179	43
16	168	38

(a) Equal importance (length of axes)

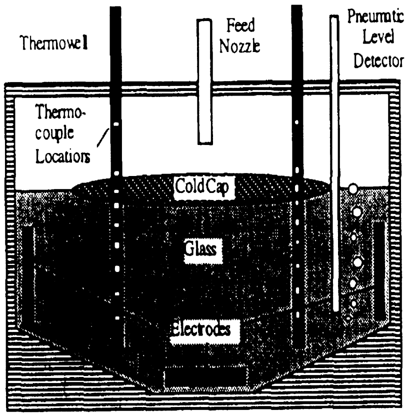


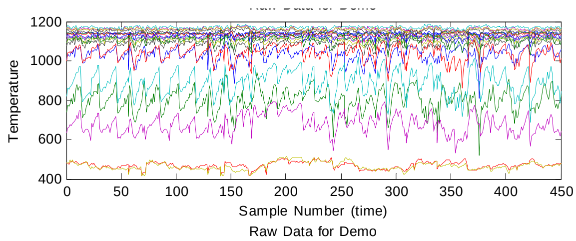
(b) unequal importance

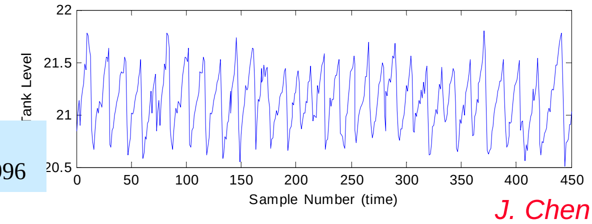


Illustrated Example: Slurry-Fed Ceramic Melter (SFCM)

- Nuclear waste from fuel reprocessing is combined with glass forming materials.
- The slurry is fed into a high temperature glass melter, producing a stable vitrified product for disposal in a long term repository.
- Temperatures are measured at 10 locations in the melter.
- Many of the variables with a great deal of correlation appear to follow a saw-tooth pattern. **»pcademo**



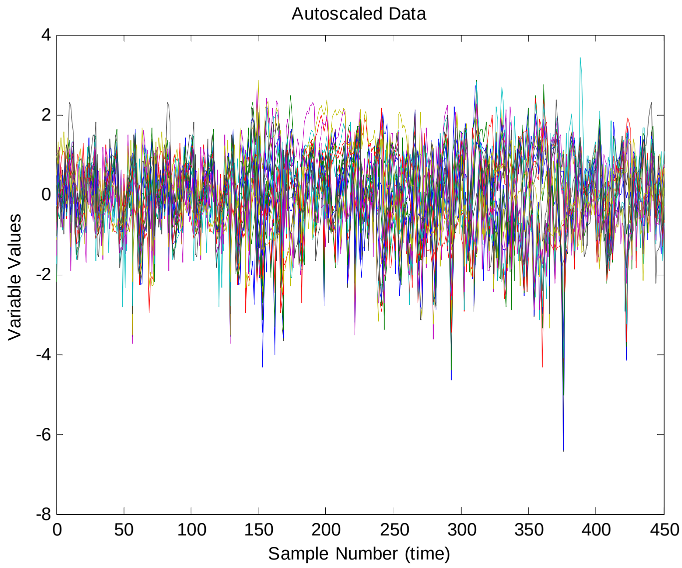




Ref: Wise, B. M. and Gallagher, N. B. The process chemometrics approach to process monitoring and fault detection, J. Proc. Cont. 6(8) 329-348, 1996

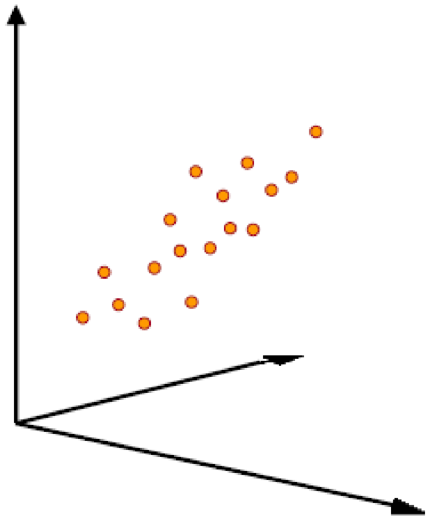
Illustrated Example: Scaled Data (SFCM)

- The first thing we want to do is to scale data before the data apply to PCA.
- If there are only temperature data, an argument can be made for mean centering of the data. Now the inclusion of a level measurement argues for autoscaling.
- Now, the data distributions look better.



»pcademo

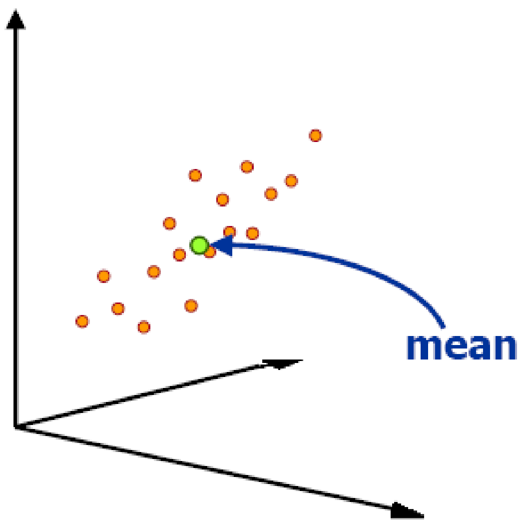
PCA - Geometric interpretation : Objects/Points



- We construct a space, with K dimensions for the matrix of data, X.

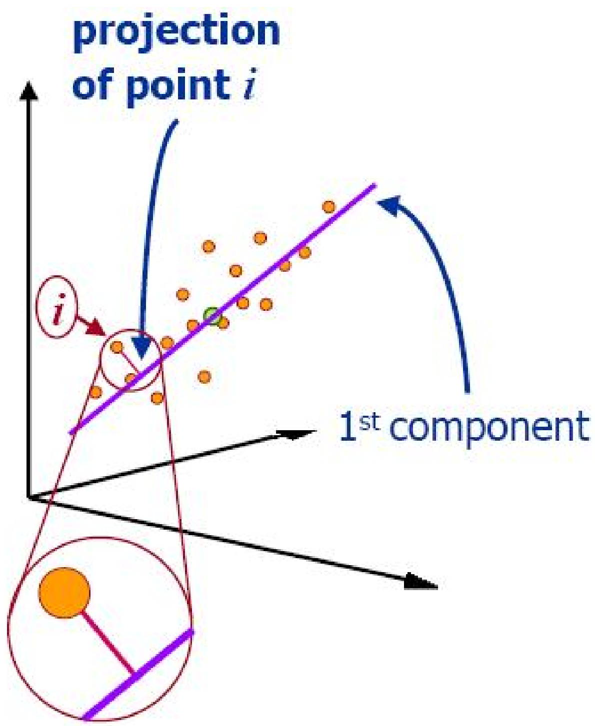
  This is called the “X-space”.
- Each variable has one coordinate axis, with the length determined by its scaling, usually unit variance.
- Each row or object in X is represented by one point in X-space.
- The data matrix X represents a swarm of points in this space.

PCA - Geometric Interpretation: Average



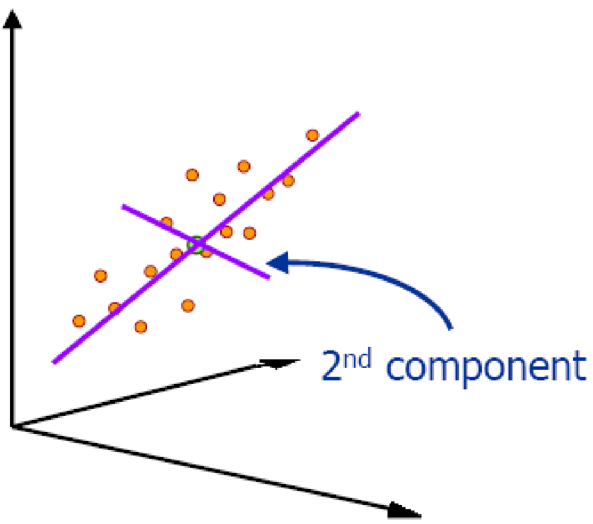
- First, we calculate the average of each variable.
- The vector of variable averages is also a point in X-space.
- This average is subtracted from the data matrix. This corresponds to moving the origin of the coordinate system to the middle (“center-of-mass”) of the data swarm.

PCA - Geometric Interpretation: First Component



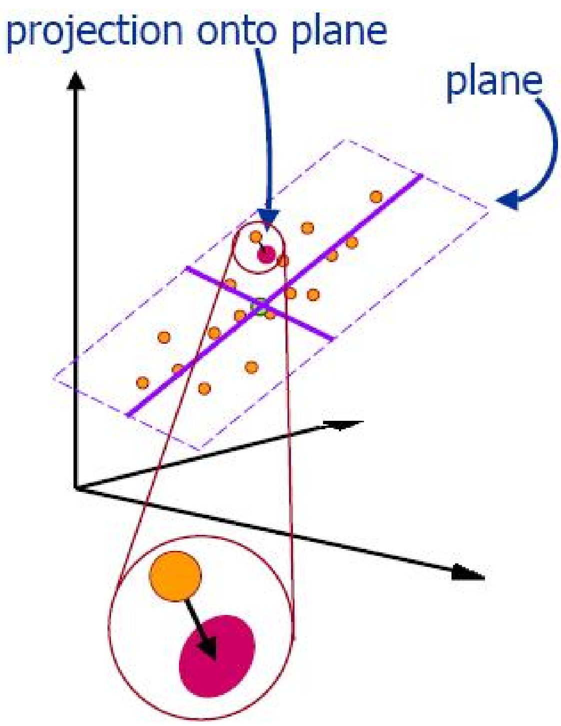
- The first principal component (PC) is a line in X-space that best approximates the data (in the least squares sense).
- It explains the greatest possible amount of variation.
- The line goes through the average point.
- The direction of the line is determined by the loading vector $\mathbf{p}_1$ (elements p_{1k}).
- The position of each point, *i*, on the line is t_{i1}.

PCA - Geometric Interpretation: Second Component



- The second PC is a line in X-space orthogonal to the line of the first component.
- It also goes through the average point.
- This line improves the approximation of the data points as much as possible.
- It explains the next greatest amount of variation.

PCA - Geometric Interpretation: PC-Plane



- The principal components together form a plane (hyper-plane) in X-space.
- Projection of points onto this plane provides a low dimensional window into our process.

Algebraic Definition of 1st PC

- Given a sample of *M* observations on a vector of *N* variables

$$\mathbf{x}_m^T = \begin{bmatrix} x_{m1} & x_{m2} & \cdots & x_{mN} \end{bmatrix}$$

- Define the first principal component of the sample by the chosen linear transformation

$$t_{m1} = \mathbf{x}_m^T \mathbf{p}_1 = \begin{bmatrix} x_{m1} & \cdots & x_{mN} \end{bmatrix} \begin{bmatrix} p_{11} \\ \vdots \\ p_{N1} \end{bmatrix} = \sum_{n=1}^{N} x_{mn} p_{n1}$$

such that

$$\max_{\mathbf{p}_1} Var(t_1) \quad \text{and} \quad \mathbf{p}_1^T \mathbf{p}_1 = 1$$

Algebraic Derivation of $\mathbf{p}_1$

$$Var(t_1) = E\left[\mathbf{p}_1^T \mathbf{x}_m \mathbf{x}_m^T \mathbf{p}_1\right]$$

$$= E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} p_{i1} x_{mi} x_{mj} p_{j1}\right]$$

$$= \sum_{n_1=1}^{N}\sum_{n_2=1}^{N} p_{n_1 1} s_{n_1 n_2} p_{n_2 1} = \mathbf{p}_1^T \mathbf{S} \mathbf{p}_1$$

where $s_{ij} \equiv \sigma_{ij} = E\left[x_{mi} x_{mj}\right]$

and $\mathbf{p}_1^T \mathbf{p}_1 = 1$

Lagrange multiplier $\max\limits_{\mathbf{p}_1, \lambda_1} L_1 = \mathbf{p}_1^T \mathbf{S} \mathbf{p}_1 - \lambda_1 \left(\mathbf{p}_1^T \mathbf{p}_1 - 1\right)$

By differentiating $2\left(\mathbf{S}\mathbf{p}_1 - \lambda_1 \mathbf{p}_1\right) = 0 \Rightarrow \left(\mathbf{S} - \lambda_1\right)\mathbf{p}_1 = 0$

p_1 is an eigenvector of S and
λ_1 is the corresponding eigenvalue.

Algebraic Definition of r-th PC

- Given a sample of *M* observations on a vector of *N* variables

$$\mathbf{x}_m^T = \begin{bmatrix} x_{m1} & x_{m2} & \cdots & x_{mN} \end{bmatrix}$$

- Define the *r*-th PC of the sample by the chosen linear transformation

$$t_r = \mathbf{x}_m^T \mathbf{p}_r = \begin{bmatrix} x_{m1} & \cdots & x_{mN} \end{bmatrix} \begin{bmatrix} p_{1r} \\ \vdots \\ p_{Nr} \end{bmatrix} = \sum_{n=1}^{N} x_{mn} p_{nr}$$

such that

$$\max_{\mathbf{p}_r} Var(t_r) \quad \text{and} \quad \mathbf{p}_r^T \mathbf{p}_r = 1$$

$$\operatorname{cov}\left(t_r, t_l\right) = 0 \qquad r > l \geq 1$$

Algebraic Derivation of p_2

$$Var(t_2) = \mathbf{p}_2^T \mathbf{S} \mathbf{p}_2 \quad Cov(t_2, t_1) = \mathbf{p}_2^T \mathbf{S} \mathbf{p}_1 = \lambda_1 \mathbf{p}_2^T \mathbf{p}_1 = 0 \quad \mathbf{p}_2^T \mathbf{p}_2 = 1$$

Lagrange multiplier

$$\max_{\mathbf{p}_2, \lambda_2, \phi} L_2 = \mathbf{p}_2^T \mathbf{S} \mathbf{p}_2 - \lambda_2 \left(\mathbf{p}_2^T \mathbf{p}_2 - 1 \right) - \phi \mathbf{p}_2^T \mathbf{p}_1$$

By differentiating $2\left(\mathbf{S}\mathbf{p}_2 - \lambda_2 \mathbf{p}_2 \right) - \phi \mathbf{p}_1 = 0$

$$2\mathbf{p}_1^T \mathbf{S} \mathbf{p}_2 - 2\lambda_2 \mathbf{p}_1^T \mathbf{p}_2 - \phi \mathbf{p}_1^T \mathbf{p}_1 = 0 \Rightarrow \phi = 0$$

$$\Rightarrow \left(\mathbf{S} - \lambda_2 \right) \mathbf{p}_2 = 0$$

p_2 is an eigenvector of S and
λ_2 is the corresponding eigenvalue.

Algebraic Derivation of *r*-th PC

- In general

$$\max_{\mathbf{p}_r} Var(t_r) = \mathbf{p}_r^T \mathbf{S} \mathbf{p}_r = \lambda_r$$

- The *r*-th largest eigenvalue of S is the variance of the *r*-th PC.
- The *r*-th PC retains the *r*-th greatest fraction of the variation in the sample.

Algebraic Formulation of PCA

- Given a sample of *M* observations on a vector of *N* variables

$$\mathbf{x}_m^T = \begin{bmatrix} x_{m1} & x_{m2} & \cdots & x_{mN} \end{bmatrix}$$

- Define a vector of *R* PCs $\mathbf{t} = \begin{bmatrix} t_1 & t_2 & \cdots & t_R \end{bmatrix}$

  according to $\mathbf{t} = \mathbf{x}_m^T \mathbf{P}$

  where **P** is an orthogonal *N* X *R* matrix
  whose *r*-th column is the *r*-th eigenvector $\mathbf{p}_r$ of S

- Then $\Lambda = \mathbf{P}^T\mathbf{S}\mathbf{P}$ is the covariance matrix of the PCs, being diagonal with elements.

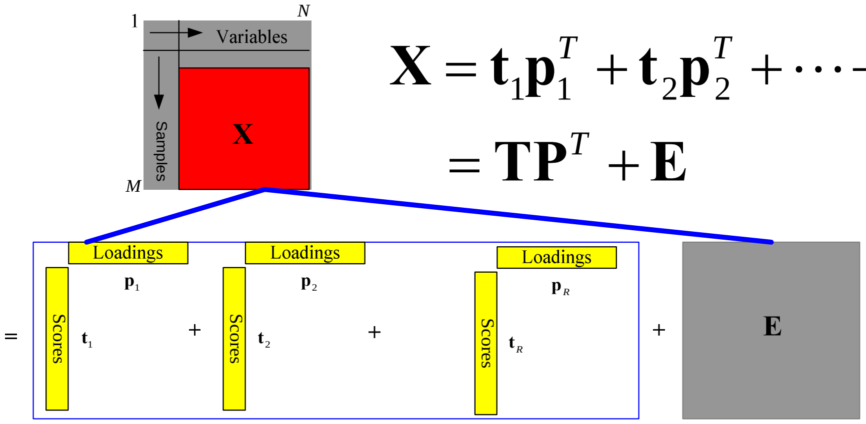
Probability Distribution for Sample PCs

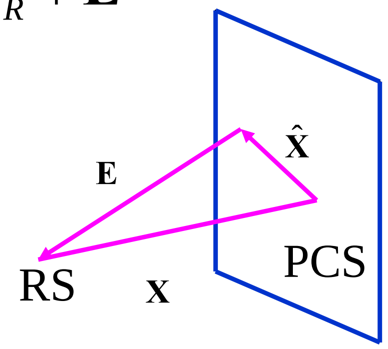
- The *M* observations of **X** in the sample are independent.
- **X** is drawn from an underlying population that follows a *N*-variable normal (Gaussian) distribution with the known covariance matrix **S**.

PCA

- PCs can be computed via SVD: $\mathbf{X}_{M\times N} = \mathbf{U}_{M\times M}\mathbf{\Sigma}_{M\times N}\mathbf{V}^T_{N\times N}$
- Select the columns of loading matrix **P** to correspond to the loading vectors **V** associated with the first *R* singular values.
- The projections of the observations in **X** into lower dimensional space are contained in score matrix $\mathbf{T} = \mathbf{XP}$.
- The projection of **T** back into the *N*-dimensional observation space
- The residual matrix is $\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}}$.

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \cdots + \mathbf{t}_R\mathbf{p}_R^T + \mathbf{E}$$
$$= \mathbf{TP}^T + \mathbf{E}$$





PCA: MATLAB Codes


```
[m,n] = size(data);
cov = (data'*data)/(m-1);
[u,s,v] = svd(cov);
loads  = v(:,1:lv);
scores = data*loads;
```

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_m^T \\ \vdots \\ \mathbf{x}_M^T \end{bmatrix}$$

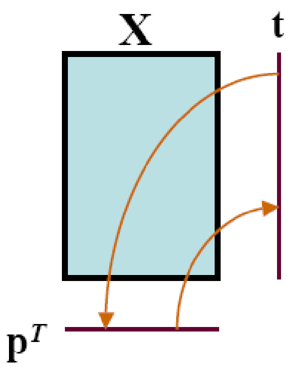
$$\mathbf{x}_m^T = \begin{bmatrix} x_{m1} & x_{m2} & \cdots & x_{mN} \end{bmatrix}$$

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \cdots + \mathbf{t}_R\mathbf{p}_R^T + \mathbf{E}$$
$$= \mathbf{TP}^T + \mathbf{E}$$

Advanced: PCA NIPALS Algorithm

NIPALS (Nonlinear Iterative PArtial Least Squares)

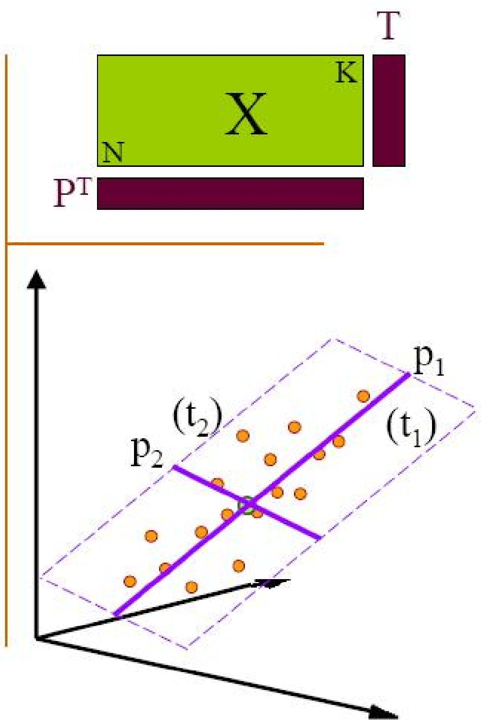
- t (start) = column of X
- Regress all columns of X on t to get loadings: $\mathbf{p}^T = \mathbf{t}^T X/\mathbf{t}^T\mathbf{t}$
- Normalize p to length 1: $\mathbf{p} = \mathbf{p}/(\mathbf{p}^T\mathbf{p})^{1/2}$
- Regress rows of X on p to get scores: $\mathbf{t} = X\mathbf{p}/\mathbf{p}^T\mathbf{p} = X\mathbf{p}$
- Check convergence of **t** (not converged? go to step 2)
- At convergence: compute residual $X = X - \mathbf{t}\mathbf{p}^T$
  Residual matrices used as next X (→ 2)



NIPALS is a variant of power law method for computing eigenvalues of X

PCA Provides an Overview of a Data Table

1. Transformation (optional)
2. Centering: subtract column averages
3. Scaling: usually, divide by column standard deviations
4. PCA = least squares projection of data onto (hyper)-plane
5. scores, t , are coordinates in the (hyper)-plane
6. loadings, p, define the direction of the (hyper)-plane

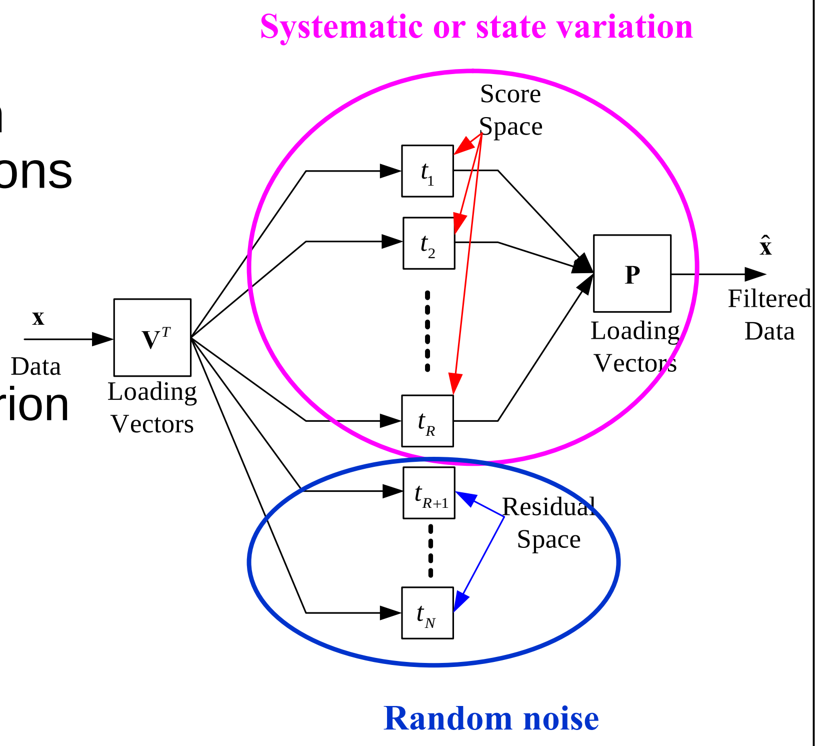


$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \cdots + \mathbf{t}_R\mathbf{p}_R^T + \mathbf{E}$$

$$= \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

Determining Number of Loading Vectors

- To avoid over-fitting in PCA, optimal dimensions should be selected.
  - » Cumulative percent variance
  - » Eigenvalue one criterion
  - » Average eigenvalue
  - » Cross validation



Optimal Dimension –– Cumulative Percent Variance

- Cumulative percent variance (CPV) measures the percent variance captured by the first *R* PCs, which can be expressed

$$CPV(R) = \frac{\sum_{r=1}^{R} \lambda_r}{\sum_{r=1}^{N} \lambda_r} 100\%$$

- *R* PC is chosen if CPV(*R*) can explain a predetermined variance, say 95%.

Optimal Dimension –– Eigenvalue One Criterion

- Only those PCs whose variances (equals to the corresponding eigenvalues of X^TX) are greater than one are retained in the model.

Optimal Dimension –– Average Eigenvalue

- Select the eigenvalues which are greater than the mean of all eigenvalues and discard eignevalues smaller than the mean.

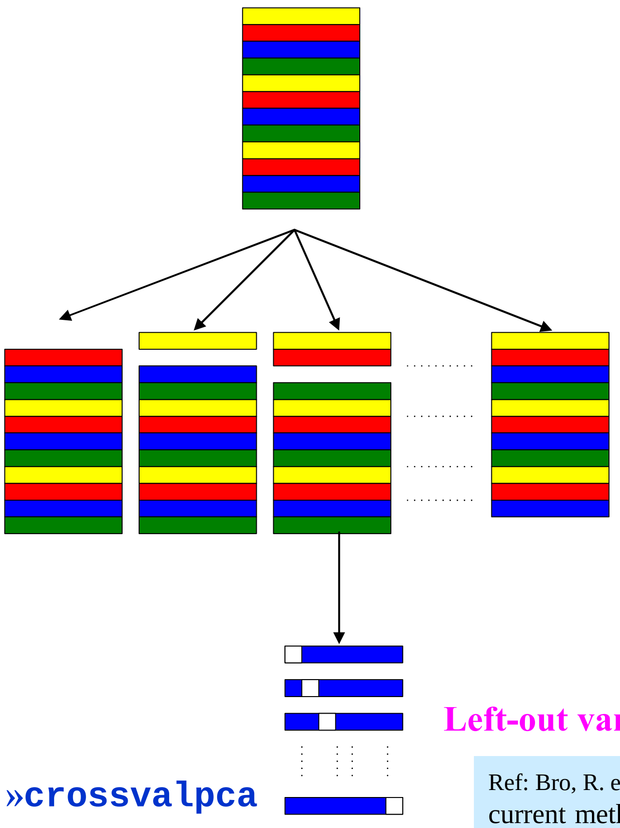
Optimal Dimension –– Cross Validation

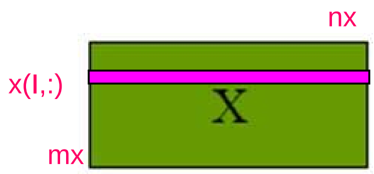
- The training set is divided into several blocks. Each time, one block ($\mathbf{X}$) is left out, and PCA is performed on the remaining blocks. The PREdiction Sum of Squares (PRESS) statistics is calculated based on the block which is left out

$$PRESS(r) = \frac{1}{(BlockSize)N} \left\| \mathbf{X} - \hat{\mathbf{X}} \right\|^2$$

- PRESS for one block is computed based on various dimensions of the score space using all the other blocks. It is repeated for each block. Adding all the resulting PRESS together gives a cumulative PRESS. The minimum cumulative PRESS determines the dimension of the score space.

Optimal Dimension –– Cross Validation (LOO)





$$e_{mn}(r) = x_{mn} - \hat{x}_{mn}(r)$$

$$PRESS(r) = \sum_{m}^{M} \sum_{n}^{N} \left[e_{mn}(r) \right]^2$$

Left-out sample *m*

$$\hat{x}_{mn}(r) = \mathbf{t}\mathbf{p}_j^T$$

$$\mathbf{t}^T = \mathbf{x}_i \mathbf{P} \left(\mathbf{P}^T \mathbf{P} \right)^{-1}$$

Left-out variable *n*

»crossvalpca

Ref: Bro, R. et al. Cross-validation of component models: A critical look at current methods, Anal Bioanal Chem (2008) 390:1241–1251

Illustrated Example: #PCs (SFCM)

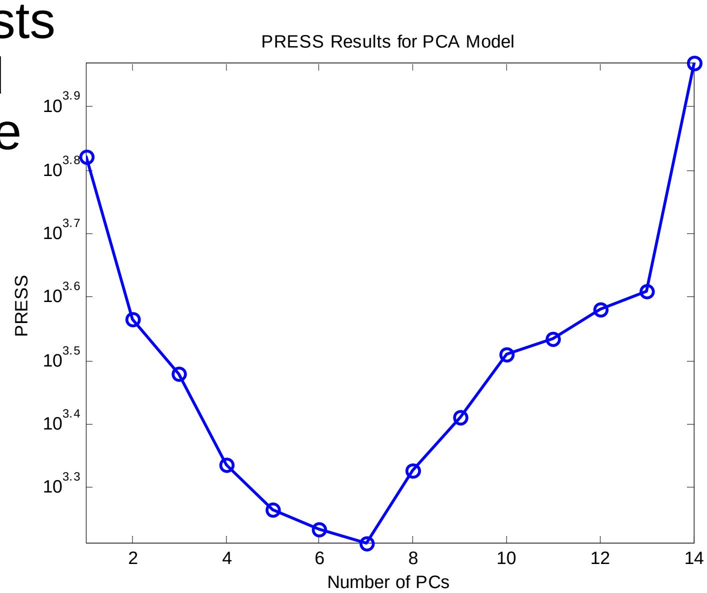
- Chose the number of principal components to keep in the model.

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	7.64e+000	36.37	36.37
2	6.35e+000	30.25	66.62
3	2.13e+000	10.13	76.75
4	1.83e+000	8.72	85.47
5	8.20e-001	3.90	89.37
6	6.15e-001	2.93	92.30
7	4.21e-001	2.00	94.30
8	3.07e-001	1.46	95.77
9	2.30e-001	1.10	96.86
10	1.85e-001	0.88	97.74
11	1.30e-001	0.62	98.36
12	9.54e-002	0.45	98.82
13	7.71e-002	0.37	99.18
14	6.25e-002	0.30	99.48
15	4.41e-002	0.21	99.69
16	2.44e-002	0.12	99.81

»**pcademo**

Illustrated Example: #PCs (SFCM)

- Chose the number of principal components to keep in the model.
- Cross validation suggests that up to 7 PCs should be the best. The relative size of the eigenvalues suggests that 4 PCs should be retained.

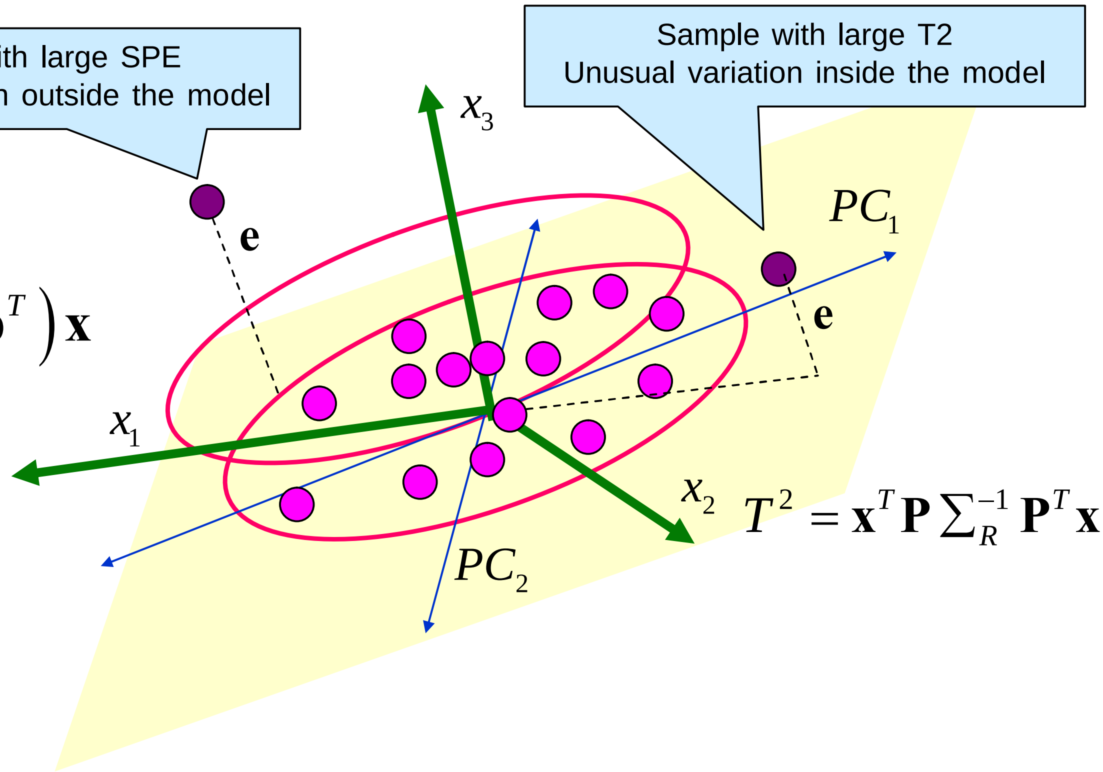


»**pcademo**

Fault Monitoring & Detection

$$Q = \mathbf{e}^T \mathbf{e}$$

$$\mathbf{e} = \left(\mathbf{I} - \mathbf{p}\mathbf{p}^T\right)\mathbf{x}$$



$$T^2 = \mathbf{x}^T \mathbf{P} \textstyle\sum_R^{-1} \mathbf{P}^T \mathbf{x}$$

Control Limits: PCA

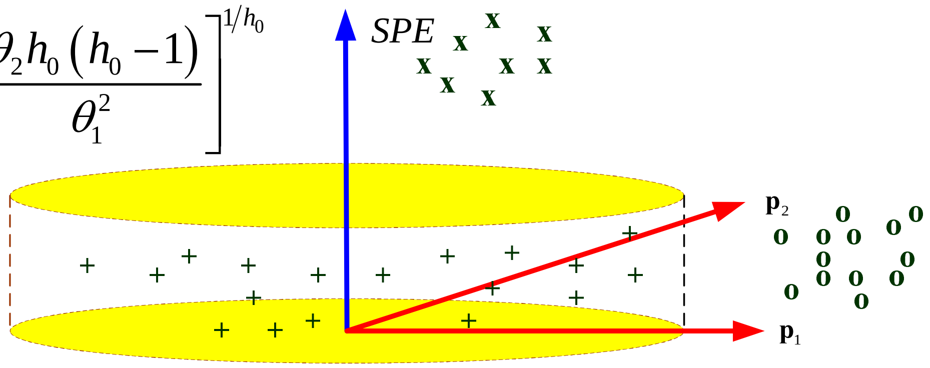
$$Q = \mathbf{e}^T \mathbf{e}$$

$$\mathbf{e} = \left(\mathbf{I} - \mathbf{p}\mathbf{p}^T\right)\mathbf{x}$$

$$Q_\alpha = \theta_1 \left[\frac{h_0 c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0 \left(h_0 - 1\right)}{\theta_1^2} \right]^{1/h_0}$$

$$\theta_i = \sum_{j=R+1}^{N} \lambda_j^i$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$$

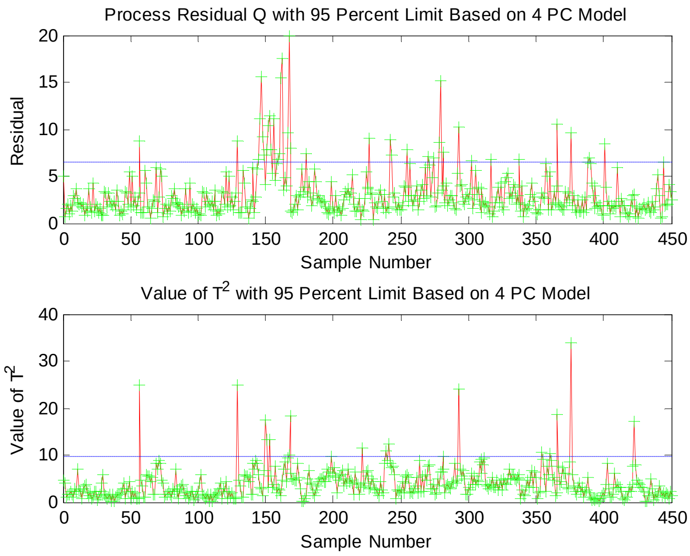


$$T^2 = \mathbf{x}^T \mathbf{P} \textstyle\sum_R^{-1} \mathbf{P}^T \mathbf{x}$$

$$T_\alpha^2 = \begin{cases} \chi_\alpha^2(R) & \text{mean and coivariance are known} \\ \dfrac{R(M-1)(M+1)}{M(M-R)} F_\alpha(R, M-R) & \text{covariance is estimated} \end{cases}$$

Illustrated Example: Control Charts (SFCM)

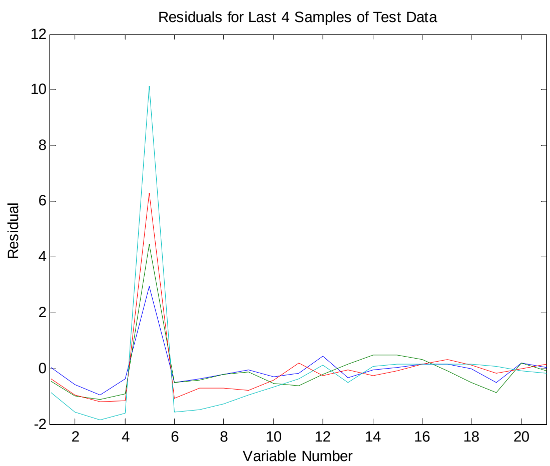
- Control limits can be placed on the process scores T^2 and residual Q.

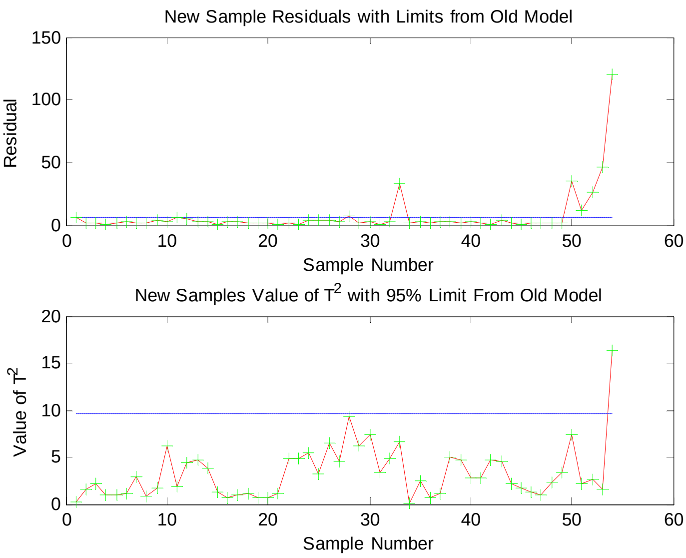


»**pcademo**

Illustrated Example: Test Set #1 (SFCM)

- At right near the end of the period, the Q residual goes over the 95% limit and stays there.
- The residual on the fifth variable is very large. It is an indication that the sensor has failed.

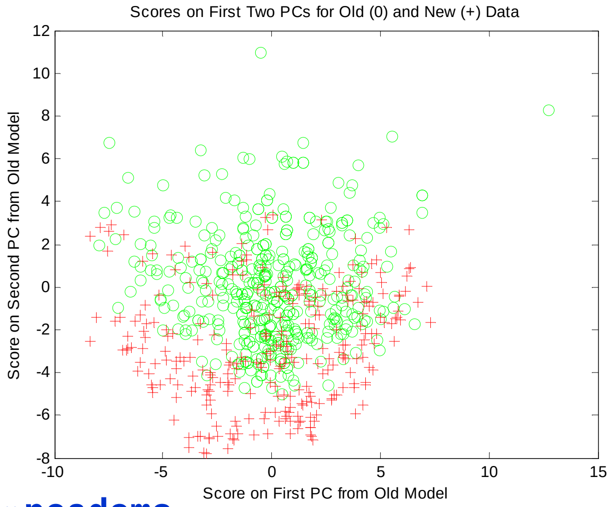




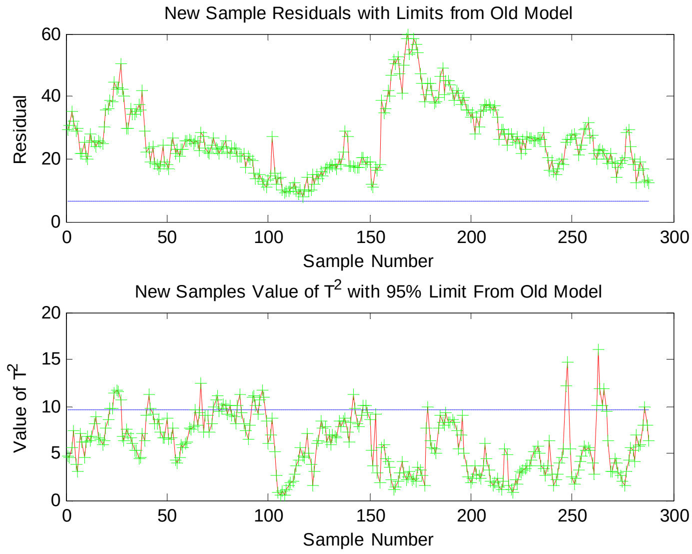
»**pcademo**

Illustrated Example: Test Set #2 (SFCM)

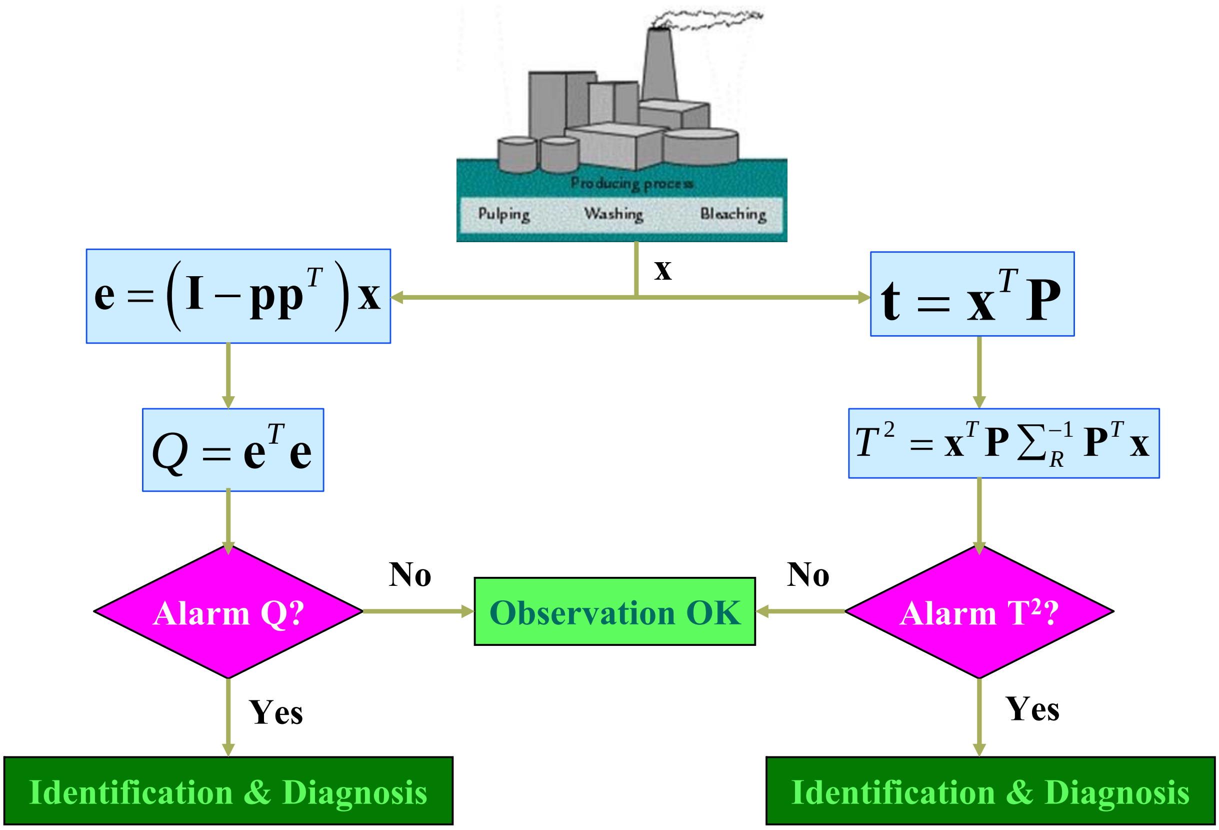
- The scores from set #2 don't fall between the limits calculated for Train. The residual is also relatively large.
- From the score plot, it is even more evident that set #2 is very different.
- This indicates that a major change has taken place in the process.



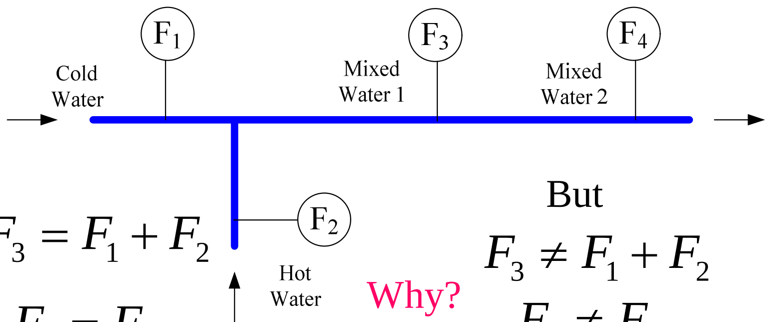
»pcademo



Monitoring with PCA



Illustrated Example: Flow Rate System



$$F_3 = F_1 + F_2$$

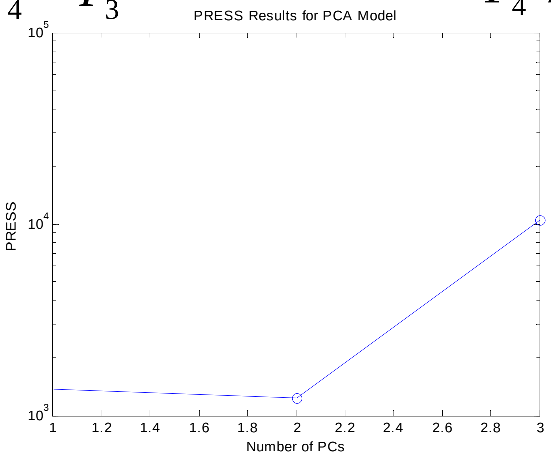
$$F_4 = F_3$$

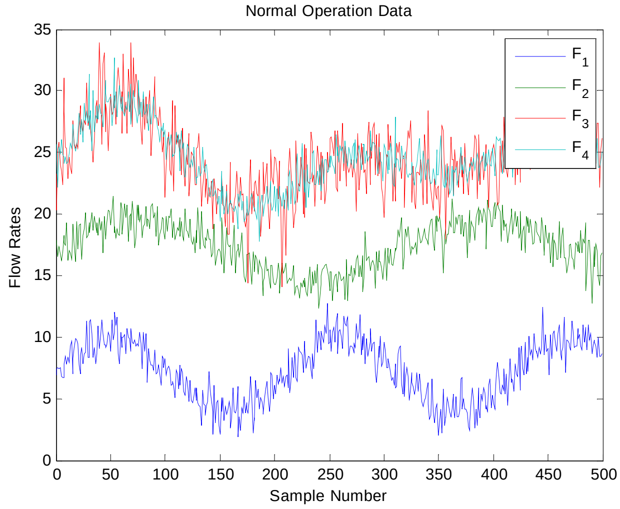
Why?

But

$$F_3 \neq F_1 + F_2$$

$$F_4 \neq F_3$$

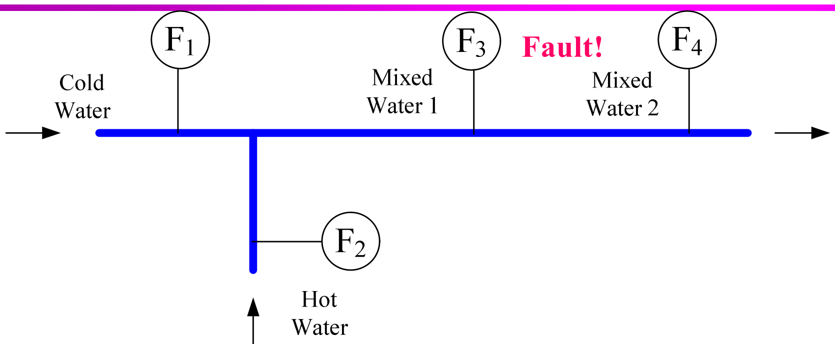


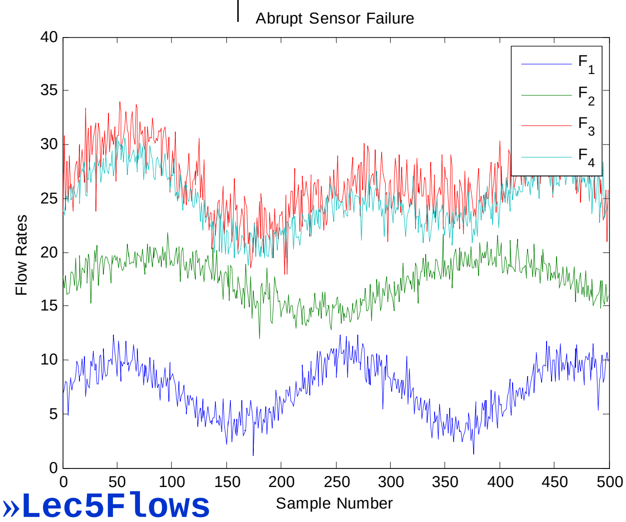


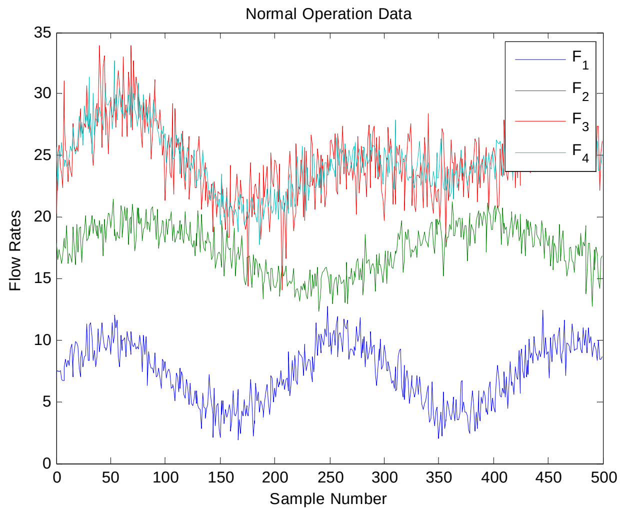
Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	2.27e+000	56.75	56.75
2	**1.15e+000**	**28.74**	**85.49**
3	3.87e-001	9.67	95.16
4	1.94e-001	4.84	100.00

»Lec5Flows

Flow Rate System: Abrupt Sensor Fault Detection

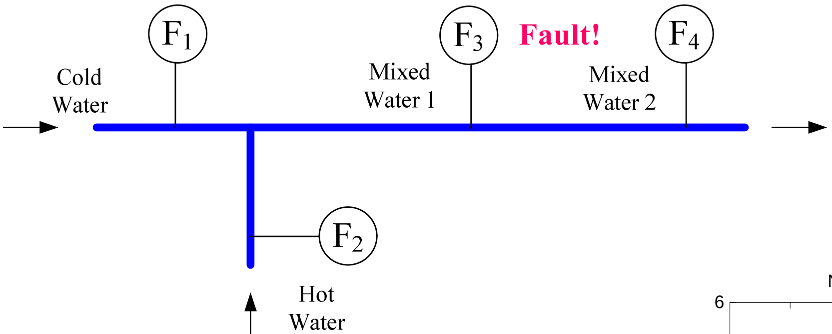


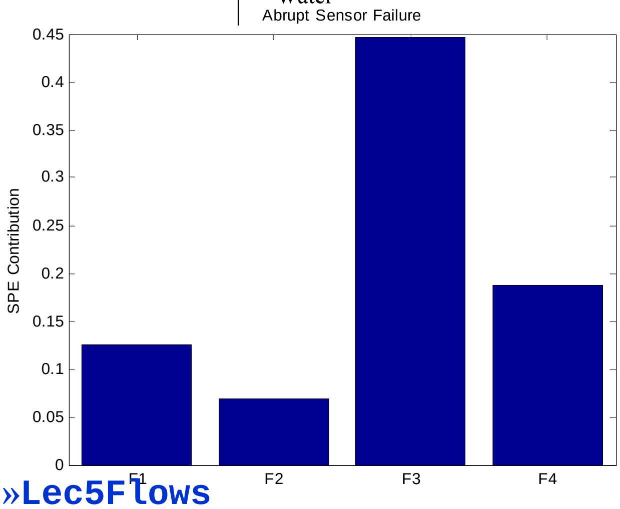


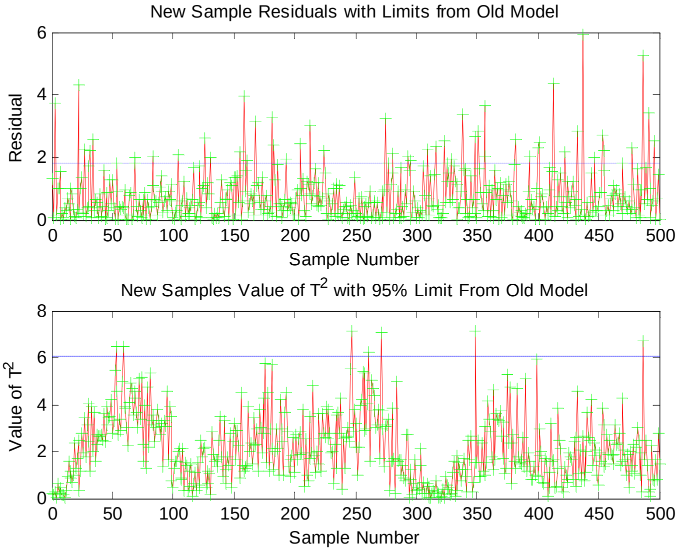


»Lec5Flows

Flow Rate System: Abrupt Sensor Fault Detection

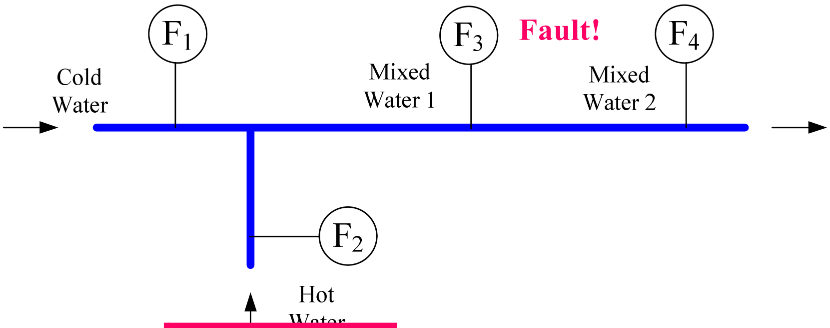


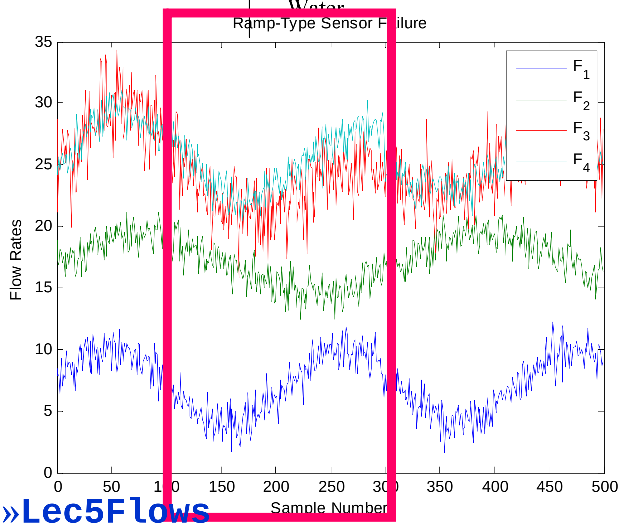


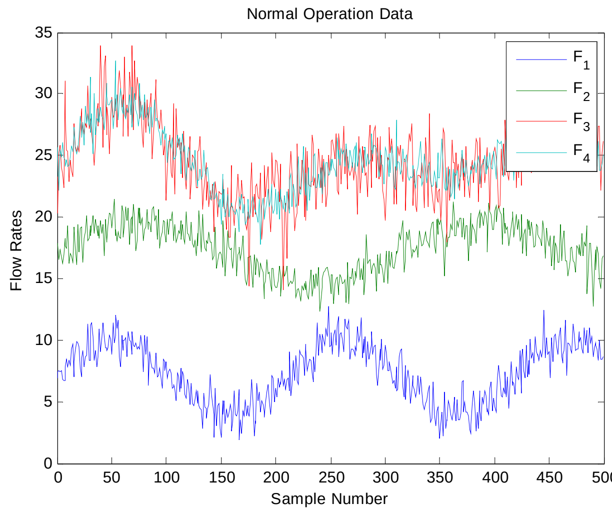


»Lec5Flows

Flow Rate System: Ramp-type Sensor Fault Detection

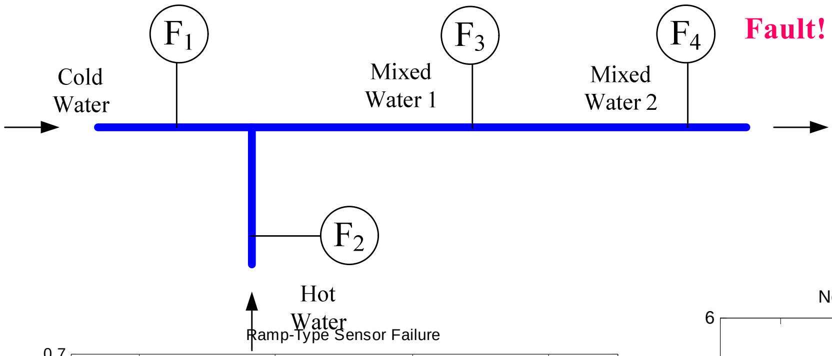


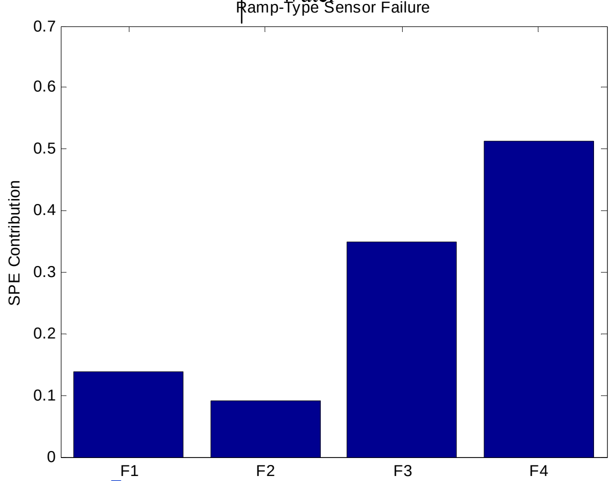


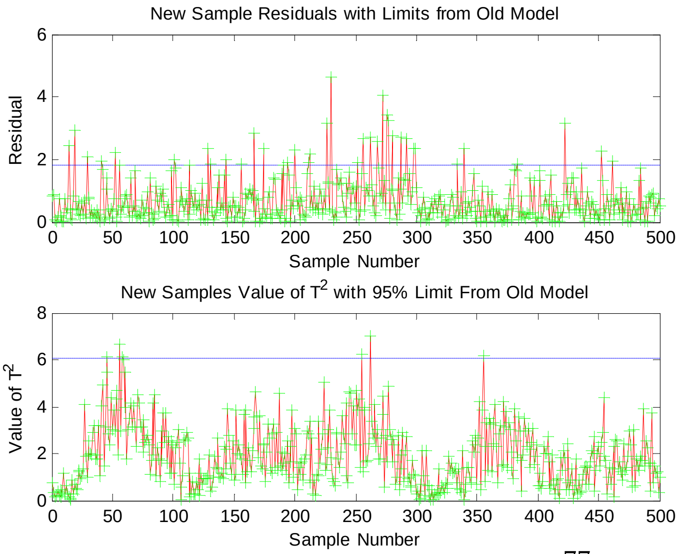


»Lec5Flows

Flow Rate System: Ramp-type Sensor Fault Detection

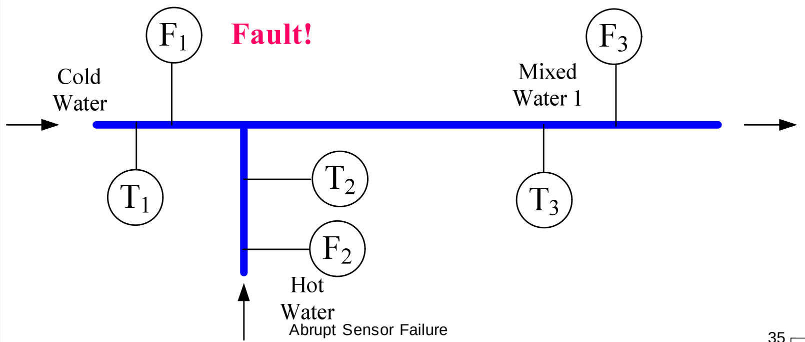






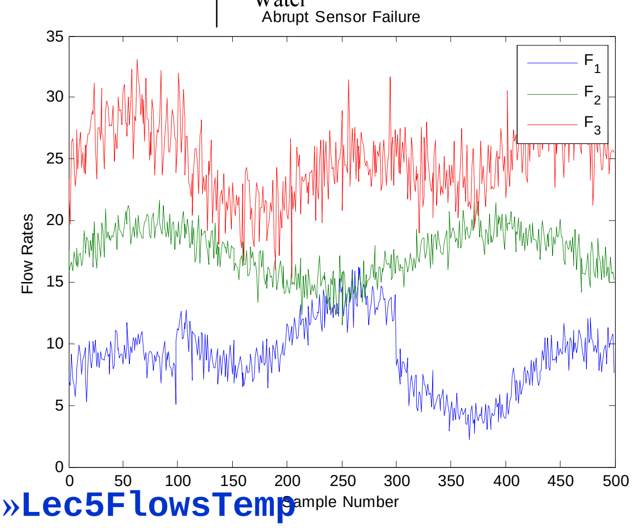
»Lec5Flows

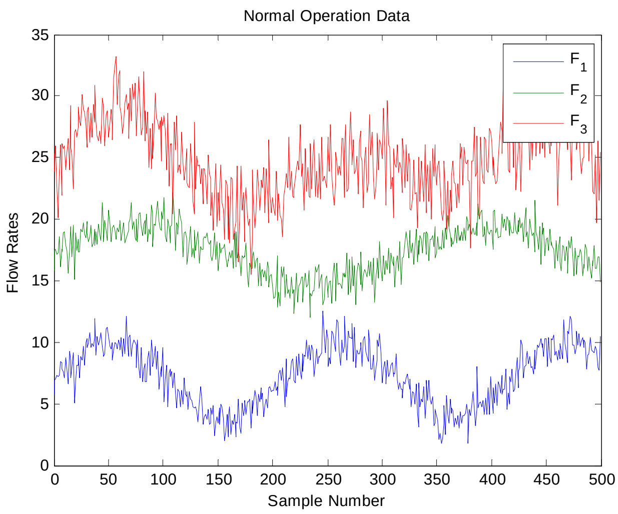
Limitation of PCA Model for MSPC



Mass Balance

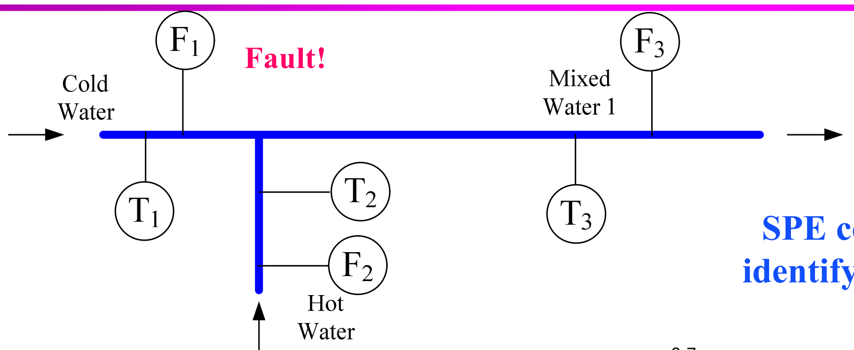
$$F_3 = F_1 + F_2$$





»Lec5FlowsTemp

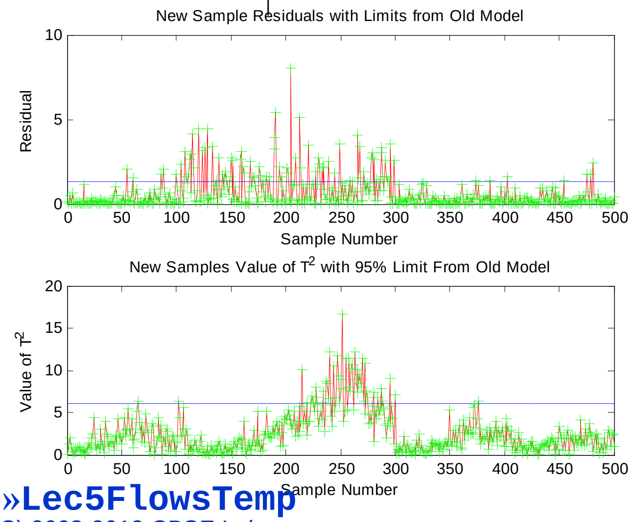
Limitation of PCA Model for MSPC

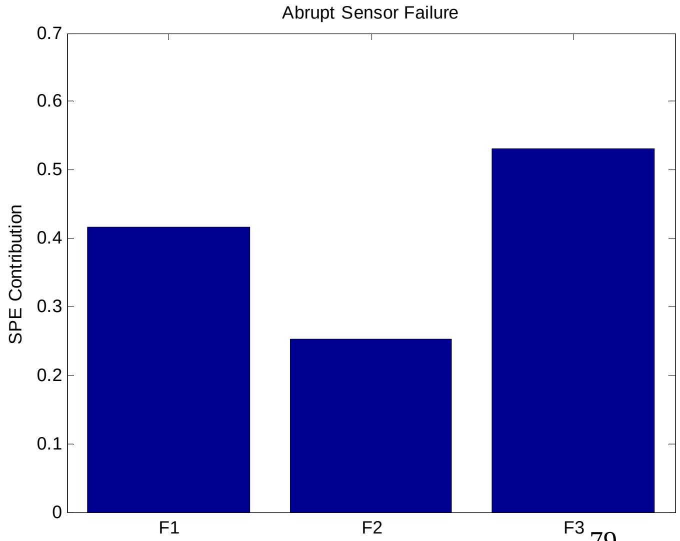


Mass Balance

$$F_3 = F_1 + F_2$$

SPE contribution is unable to identify the correct fault sensor. Why?

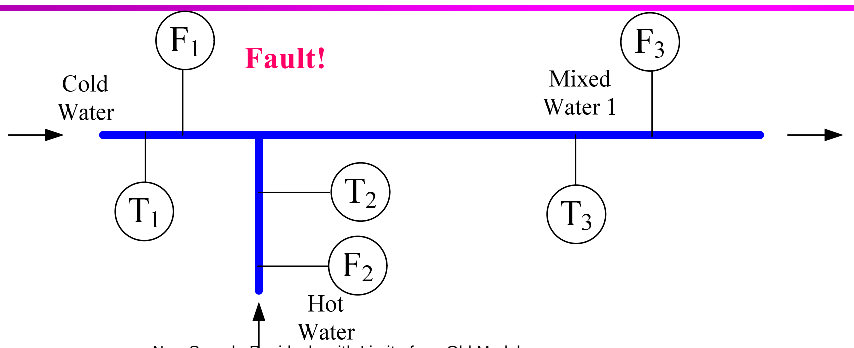




»Lec5FlowsTemp

J. Chen

Limitation of PCA Model for MSPC

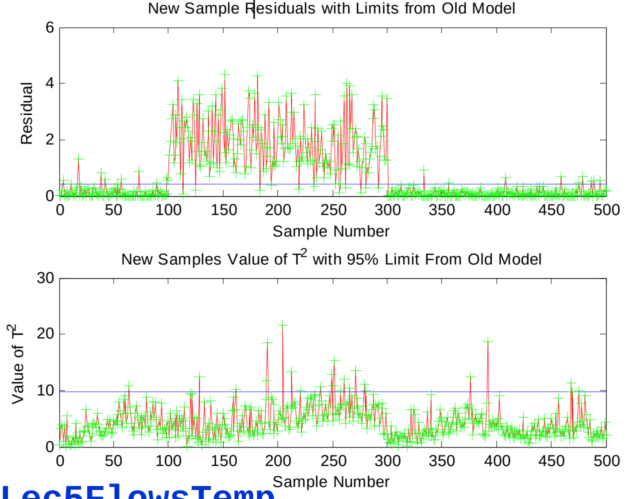


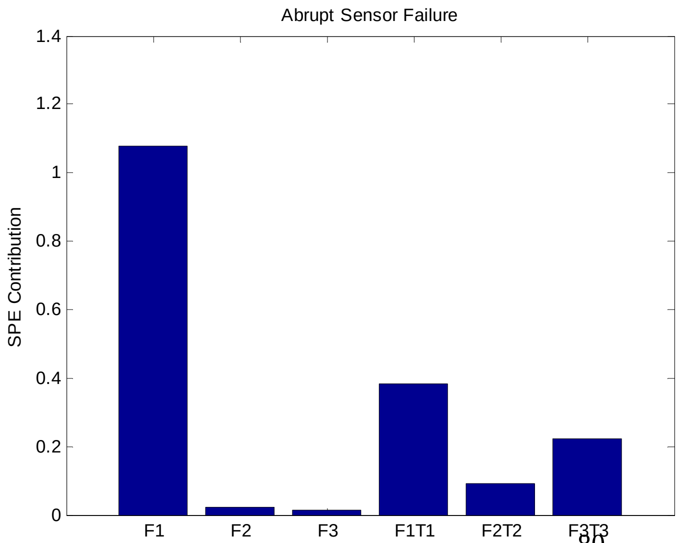
Mass Balance

$$F_3 = F_1 + F_2$$

Energy Balance

$$F_3T_3 = F_1T_1 + F_2T_2$$

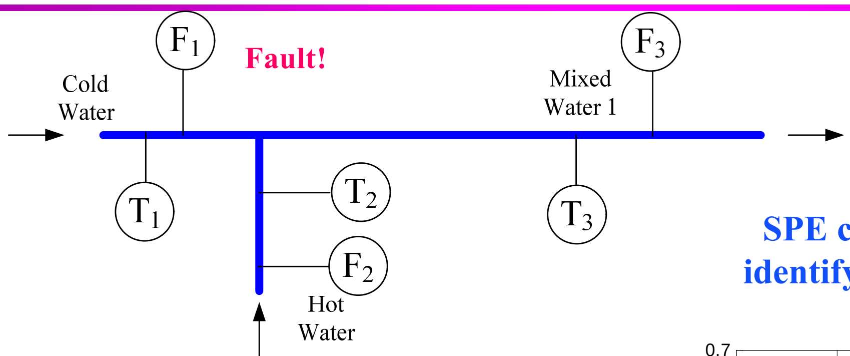




»Lec5FlowsTemp

J. Chen

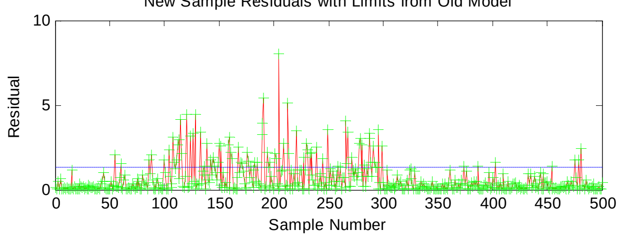
Limitation of PCA Model for MSPC

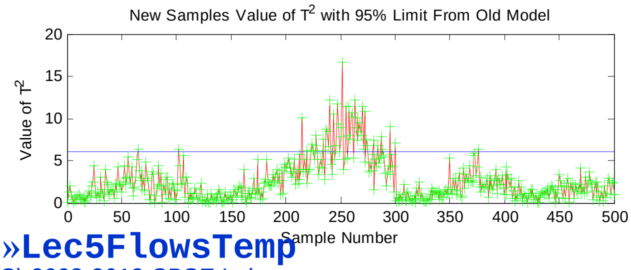


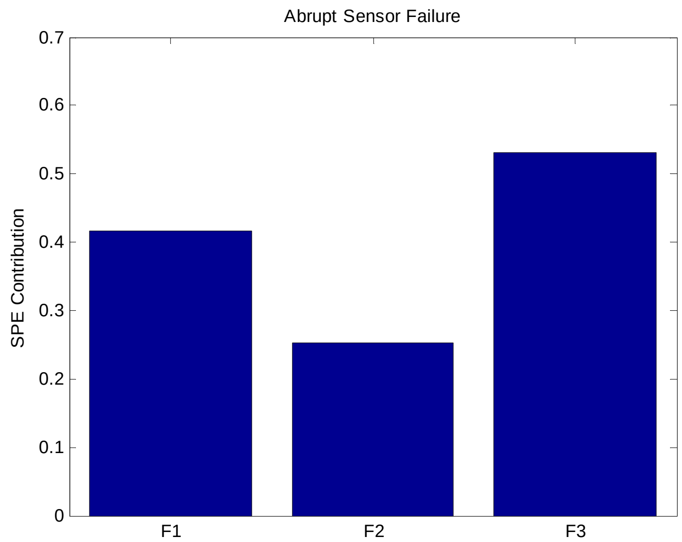
Mass Balance

$$F_3 = F_1 + F_2$$

SPE contribution is unable to identify the correct fault sensor. Why?

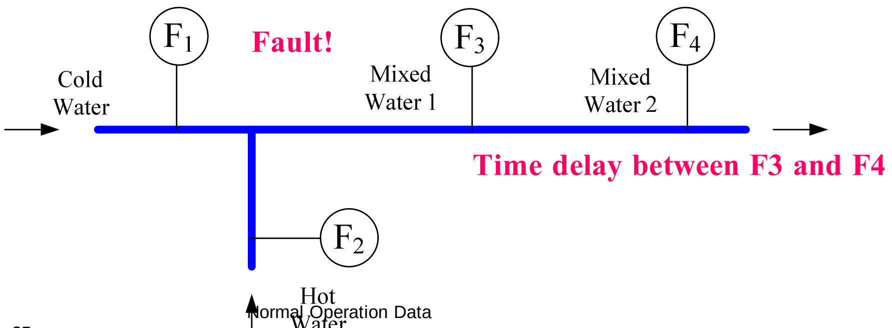






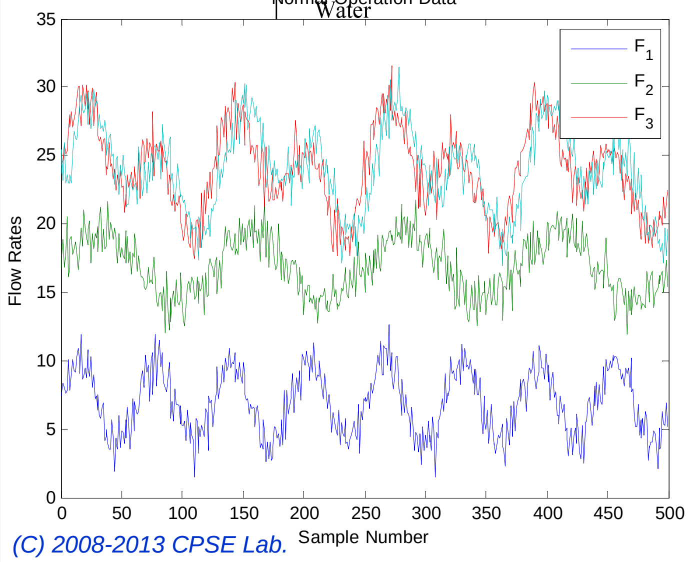
»**Lec5FlowsTemp**

Dynamic Process of PCA Model for MSPC



$$F_3 = F_1 + F_2$$

$$F_4(t+7) = F_3$$

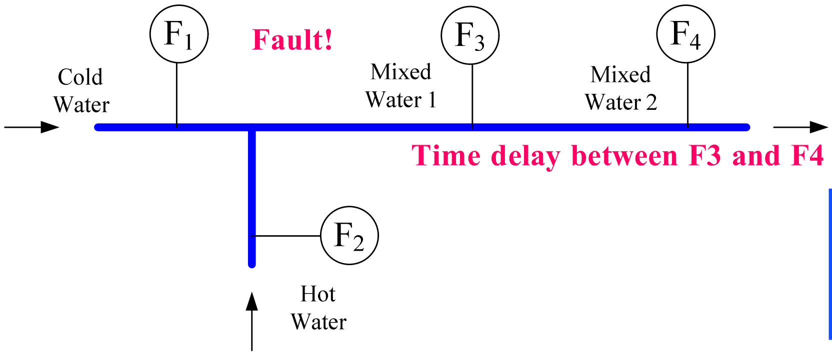


$$X1 = [F_1(t) \quad F_2(t) \quad F_3(t) \quad F_4(t)]$$

$$X2 = [F_1(t) \quad F_2(t) \quad F_3(t) \quad F_4(t+7)]$$

»**Lec5FlowsDealy**

Dynamic Process of PCA Model for MSPC



$$F_3 = F_1 + F_2$$

$$F_4(t+7) = F_3$$

To build a correct PCA model, it is important to include lagged variable when time delay exists.

$$X1 = [F_1(t) \quad F_2(t) \quad F_3(t) \quad F_4(t)] \quad X2 = [F_1(t) \quad F_2(t) \quad F_3(t) \quad F_4(t+7)]$$

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
---------	----------	----------	----------
1	2.47e+000	61.68	61.68
2	1.05e+000	26.31	87.98
3	3.22e-001	8.05	96.03
4	1.59e-001	3.97	100.00

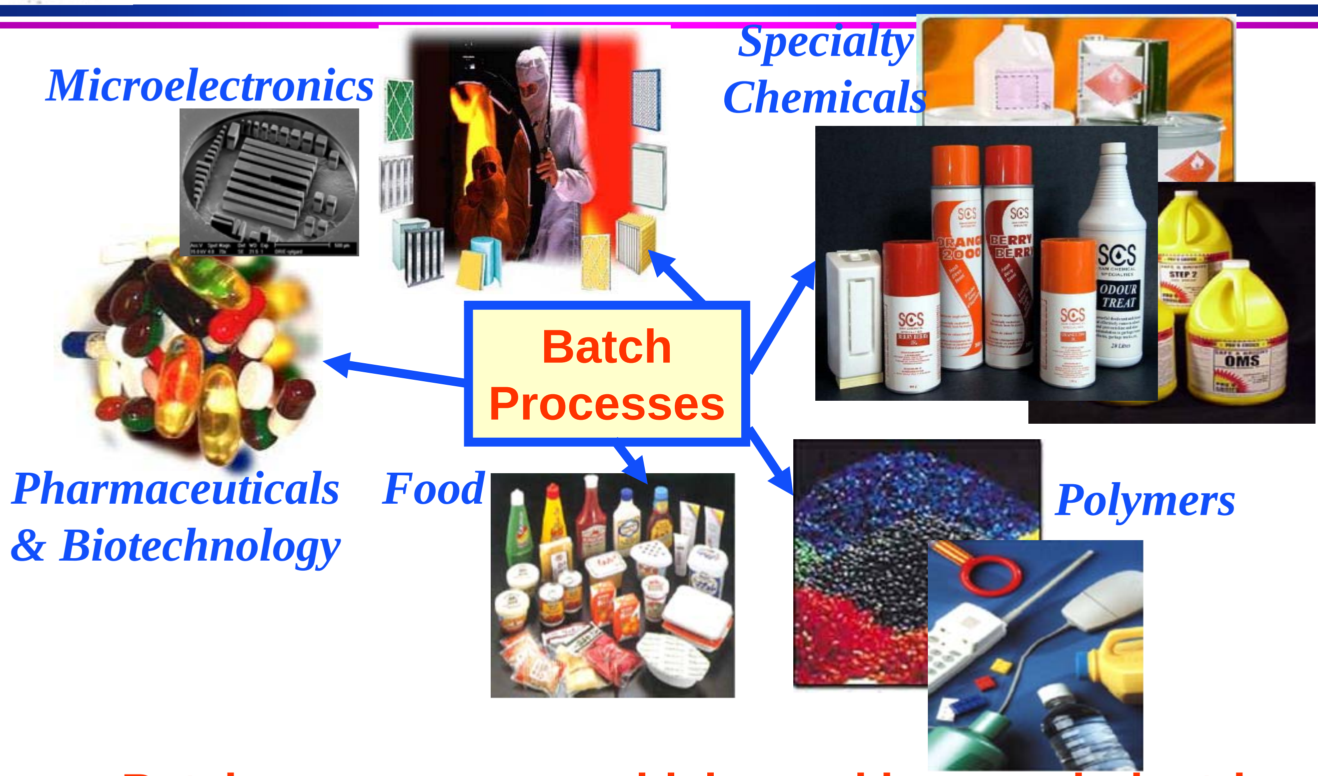
Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
---------	----------	----------	----------
1	2.66e+000	66.45	66.45
2	1.05e+000	26.36	92.81
3	1.71e-001	4.28	97.10
4	1.16e-001	2.90	100.00

»Lec5FlowsDealy

Monitoring for Batch Processes

- Batch processes, unlike continuous reactors which are most often used for high-throughput plants, are also frequently used in situations where production rates are low.
- Increasing quality and performance demands require to drive processes near limits
  - » Batch polymerization reactors permit the production of polymer with a more narrow molecular weight distribution.
  - » Batch fermentors use the lifecycle of the “bugs” to grow the organisms by feeding them substrate and letting them produce the desired chemical
- The batch reactor is quite flexible and can be used to produce a number of different products under a variety of conditions in the same vessel

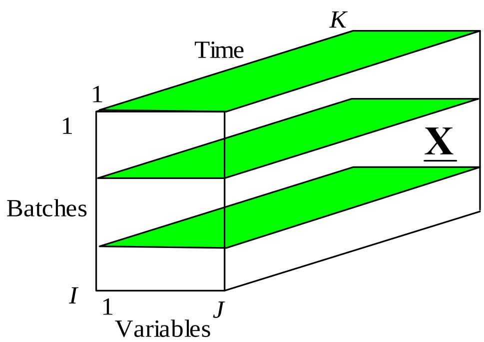
Batch Processes

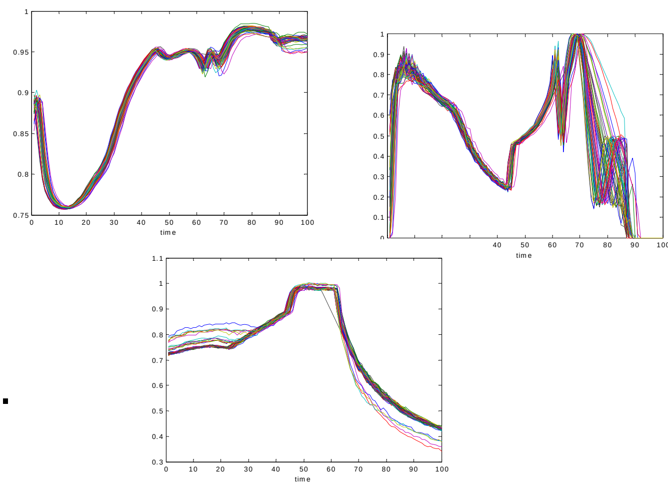


- **Batch processes are widely used in many industries**
- **Reduced time-to-market, flexible operation**

Batch Data Structure

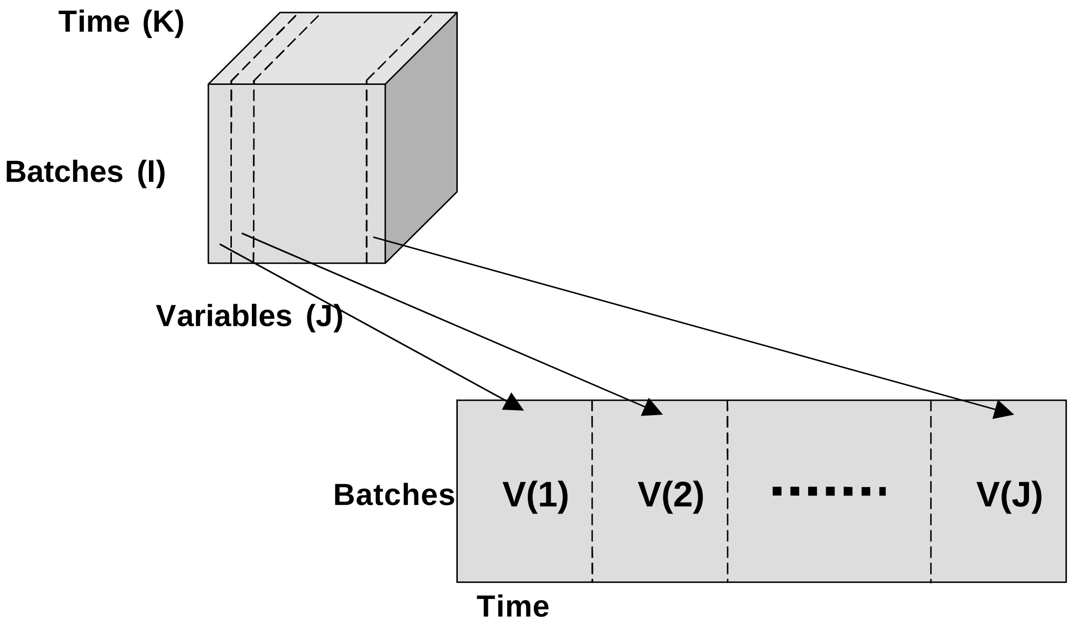
- Batch data presents a 3-dimensional problem.
- With continuous processes it is just the relationships between the variables that are important.
- Batch data includes the add dimension of time since the entire past history of the trajectory contributes to the overall performance of the process.
- To analyze the data, the 3-D matrix must first be unfolded.





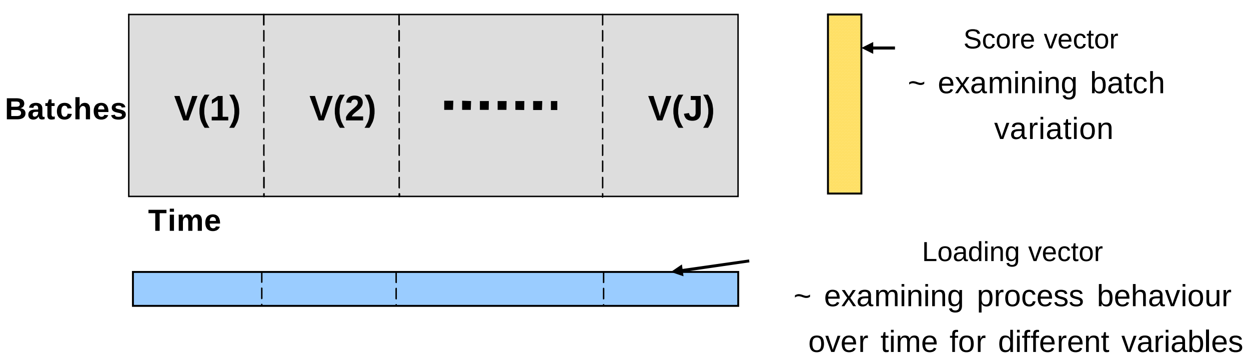
Unfolding Batch Data

- Multi-way unfolding (Nomikos and MacGregor, 1994)



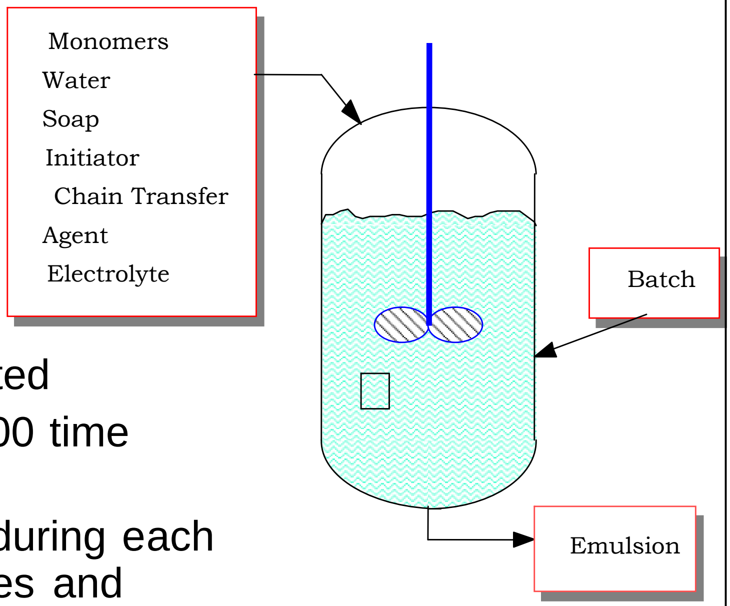
MPCA for Unfolding Batch Data

- Apply PCA to the unfolded equalised batch data
- Extract the principal component score vectors
- Batch performance can be investigated



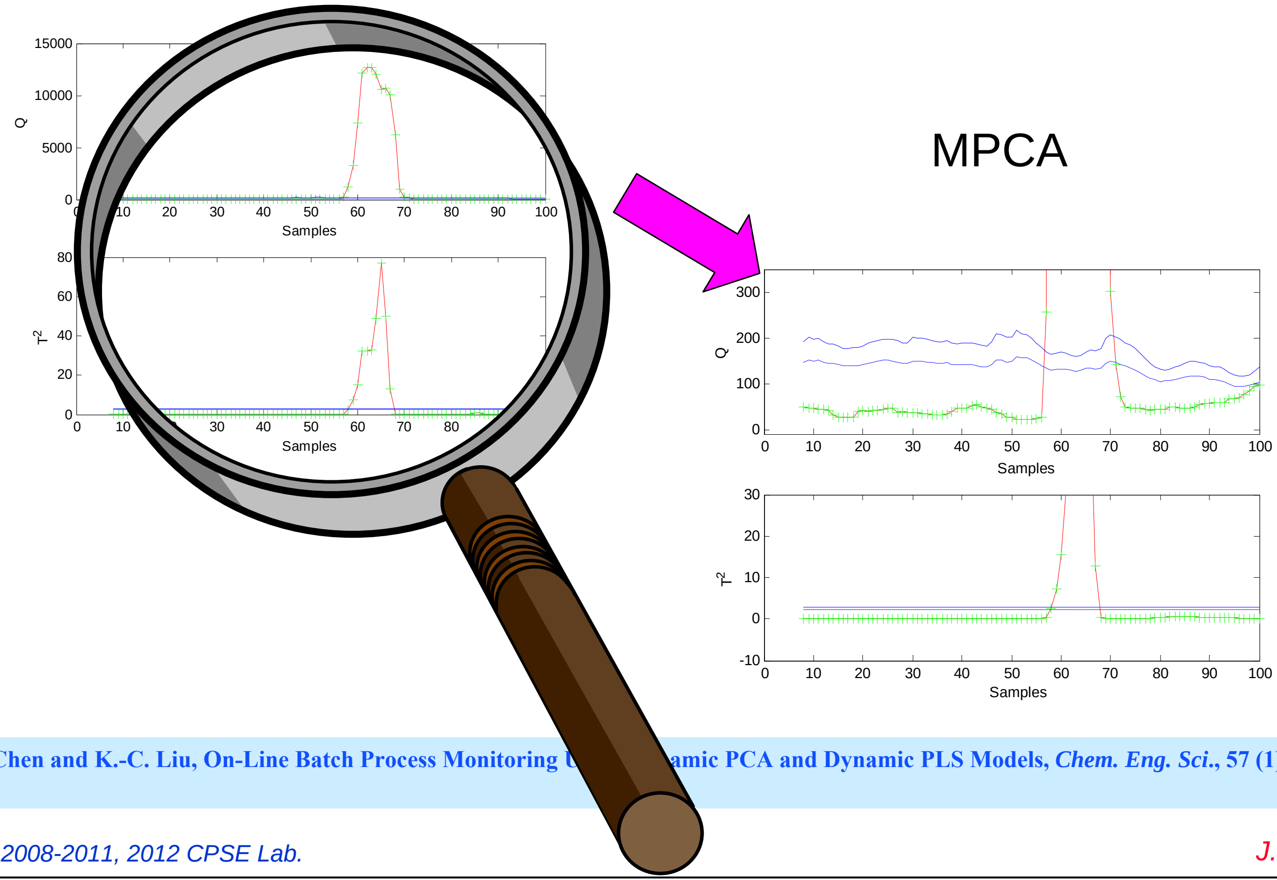
DuPont Batch Polymerization Data

- 2hr are needed to finish one batch run
- 12hr are needed to analyze the properties



- 36 normal batches are collected
- Each batch has duration of 100 time intervals
- Ten variables are measured during each batch (temperatures, pressures and flowrate)

MPCA for DuPont Batch Polymerization Data

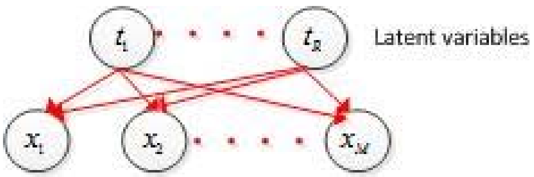


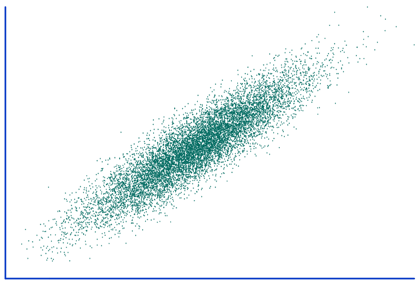
J. Chen and K.-C. Liu, On-Line Batch Process Monitoring U[illegible]amic PCA and Dynamic PLS Models, *Chem. Eng. Sci.*, 57 (1) 63-75,

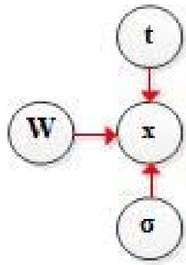
Probabilistic Modeling

- Input: A set of training data
- Procedure
  1. Define the generative model
  2. Derive the likelihood of the data
  3. Specify model parameters
  4. Bayesian: Assign priors (with some hyperparameters)
  5. Model learning: find best parameters/hyperparameters
  6. Inference: make prediction for test data via Bayes' rule
- Advantages
  - » Deep foundation in probability theory and statistics
  - » Many learning and inference algorithms available
    - – Expectation Maximization, Variational Bayes, …

Probabilistic PCA







- Latent variable model

$$\mathbf{x} = \mathbf{W}\mathbf{t} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

Latent variable $\mathbf{t} \sim N(0, \mathrm{I})$

Mean vector

Noise process $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\sigma}^2 \mathrm{I})$

- **The PPCA model indicates that given the latent variable t, x is Gaussian distributed:** $\mathbf{x} | \mathbf{t} \sim N(\mathbf{W}\mathbf{t} + \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mathbf{I})$

- **If $\boldsymbol{\sigma}^2 \to 0$, PPCA leads to PCA solution (up to a rotation and scaling factor)**

Probabilistic PCA

Determination of parameters:

- Given observed variables **x** , the log likelihood function is :

$$\ln p(\mathbf{x} \mid \mu, \mathbf{W}, \sigma^2) = \sum_{i=1}^{n} \ln p(\mathbf{x}_n \mid \mu, \mathbf{W}, \sigma^2)$$

- The parameters can be obtained by EM algorithm :

$$E(\mathbf{t} \mid \mathbf{x}) = \mathbf{M}^{-1}\mathbf{W}_{ML}{}^{T}(\mathbf{x}-\mu)$$

$$E(\mathbf{t}\mathbf{t}^T \mid \mathbf{x}) = \sigma^2\mathbf{M}^{-1} + E(\mathbf{t} \mid \mathbf{x})E^T(\mathbf{t} \mid \mathbf{x})$$

E step : Using the observed variables and the current (fixed) parameters to compute latent variables

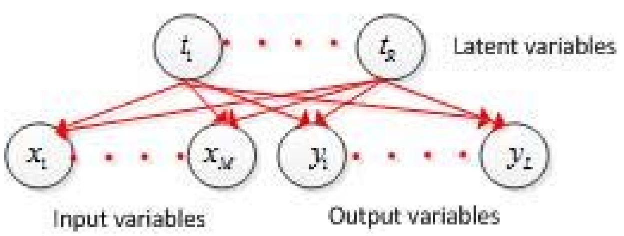
$$\mathbf{W} = [\sum_{i=1}^{n} E(\mathbf{t}_i \mid \mathbf{x}_i)\mathbf{x}_i^T][\sum_{i=1}^{n} E(\mathbf{t}_i\mathbf{t}_i^T \mid \mathbf{x}_i)]^{-1}$$

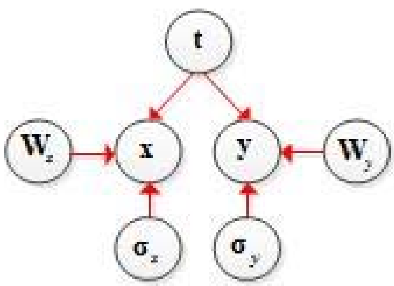
$$\sigma^2 = \frac{\sum_{i=1}^{n}\{\mathbf{x}_i^T\mathbf{x}_i - 2E^T(\mathbf{t}_i \mid \mathbf{x}_i)\mathbf{W}^T\mathbf{x}_i + Tr[E(\mathbf{t}_i\mathbf{t}_i^T \mid \mathbf{x}_i)\mathbf{W}^T\mathbf{W}]\}}{nm}$$

M step: Maximize the likelihood function with respect to parameters by the latent variables from E step

Tipping, M. E.; Bishop, C. M. Probabilistic principal component analysis. *J. R. Stat. Soc.* **1999**, *61*, 611–622.

Supervised Probabilistic PLVR





- Supervised latent variable model

$$\mathbf{x} = \mathbf{W}_x\mathbf{t} + \boldsymbol{\mu}_x + \boldsymbol{\varepsilon}_x$$

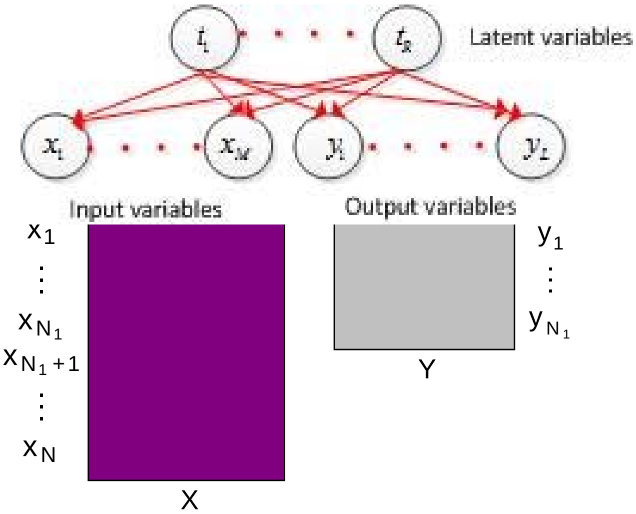
$$\mathbf{y} = \mathbf{W}_y\mathbf{t} + \boldsymbol{\mu}_y + \boldsymbol{\varepsilon}_y \qquad \boldsymbol{\varepsilon}_y \sim N(0, \boldsymbol{\sigma}_y^2 \mathrm{I})$$

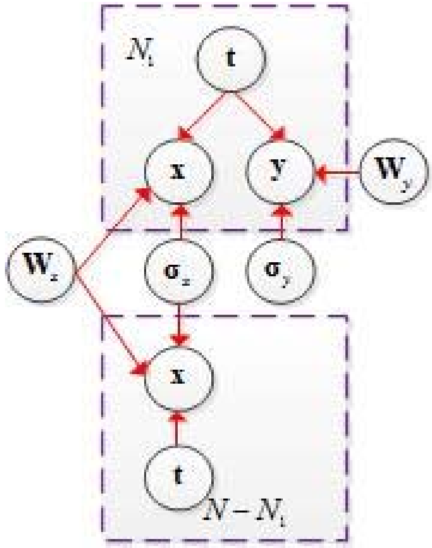
Noise process

  » Inputs and outputs are conditionally independent
  » All input/output dimensions are conditionally independent

$$\mathbf{x}\,|\,\mathbf{t} \sim N(\mathbf{W}_x\mathbf{t} + \boldsymbol{\mu}_x, \boldsymbol{\sigma}_x^2\mathbf{I}) \qquad \mathbf{y}\,|\,\mathbf{t} \sim N(\mathbf{W}_y\mathbf{t} + \boldsymbol{\mu}_y, \boldsymbol{\sigma}_y^2\mathbf{I})$$

Semi-Supervised Probabilistic PLVR





- Idea: A PLVR model with missing data!

Non-Gaussian Process Monitoring Methods

method	advantages	disadvantages
ICA	(1) simple model structure, easy to understand	(1) difficult to determine the control limit
	(2) able to extract high-order data information	(2) the monitoring result may be unstable
	(3) provide latent variables that are independent to each other	(3) difficult to select the number of independent components
GMM	(1) simple model structure, easy to understand	(1) difficult to determine the number of local models
	(2) be able to monitoring processes with multiple operating conditions	(2) model training is complicated
	(3) can also handle the nonlinearity of the process	(3) may not be able to model all types of non-gaussian data
SVDD	(1) the developed model can be directly used for process monitoring	(1) the kernel parameter of the model should be tuned
	(2) can handle both of the linear and nonlinear process data	(2) the tighter control limit of SVDD may cause more false alarms
	(3) has no assumption of the data distribution	(3) process analyses and interpretations become more difficult

Nature of Process Data

- Very high dimensional data matrices
  - » Many variables and many observations
- Non-causal in nature
  - » Can't generally use data to imply cause and effect relationships
  - » But can get informative correlation relationships
- Variables are not independent
  - » High correlation among variables – not independent
- Missing data
  - » 10-20 % missing is common
- Low signal to noise ratio
  - » Each variable contains little information – need multivariate methods

Need efficient multivariate methods to treat these problems!

Conclusion

- Clustering and projection are important tasks in DM
- Probabilistic modeling would be a good way to apply to both tasks
- Joint clustering-projection models
  - » Principled way to iterate clustering and projection
  - » Convergence is guaranteed, with better performance