15th International Conference
Applications of Computer Algebra

Teaching Principal Components Analysis with Minitab

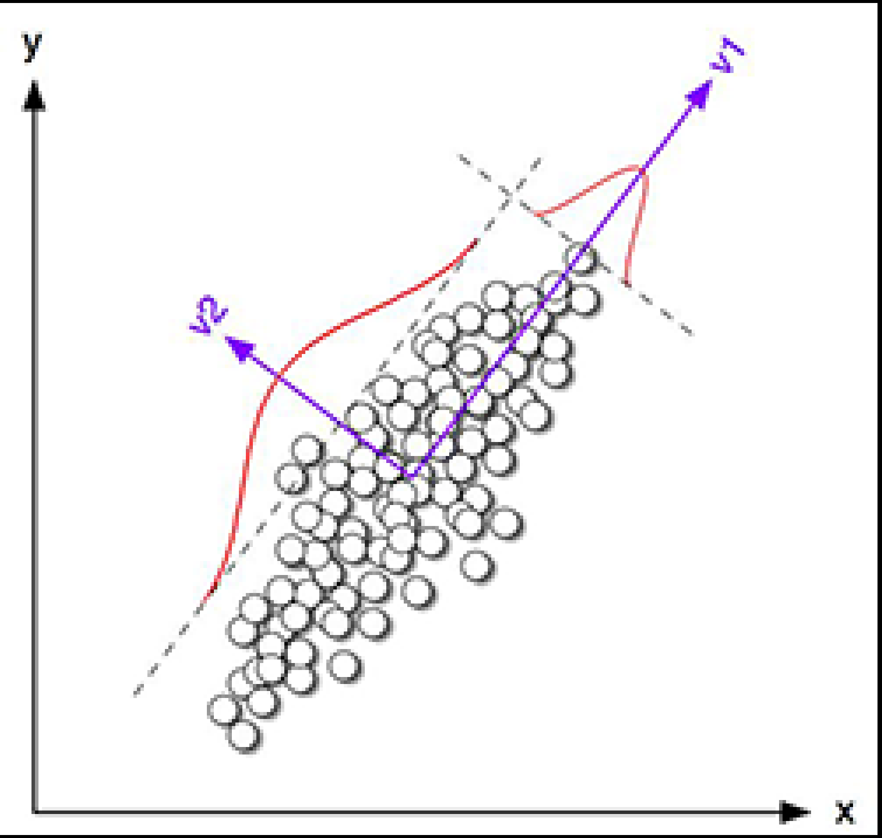
Dr. Jaime Curts

The University of Texas Pan American

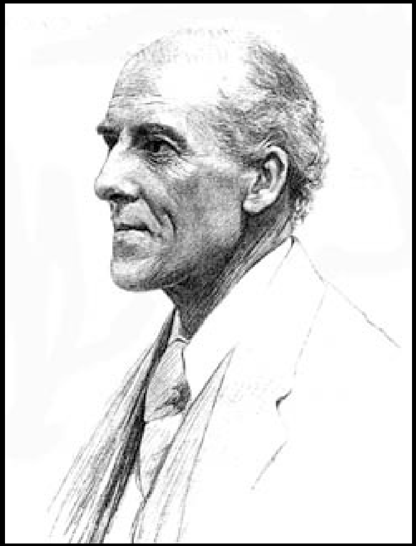
ACA 2009 to be held June 25-28, 2009
at École de Technologie Supérieure (ETS),
Université du Québec, Montréal, Québec, Canada

Introduction

The purpose of this presentation is to introduce the logical and arithmetic operators and simple matrix functions of Minitab® –a well-known software package for teaching statistics- as a computer-aid to teach Principal Components Analysis (PCA) to graduate students in the field of Education.

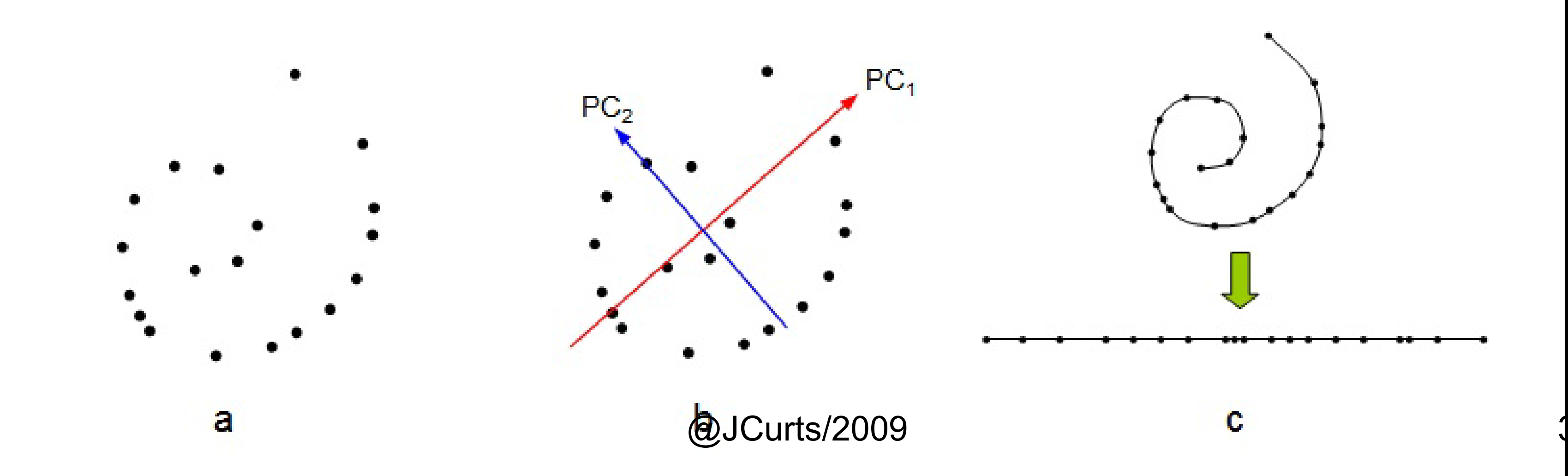


PCA, originally proposed by Pearson (1901) is a mathematical technique –a vector space transform- that has its roots in linear algebra and in statistics.



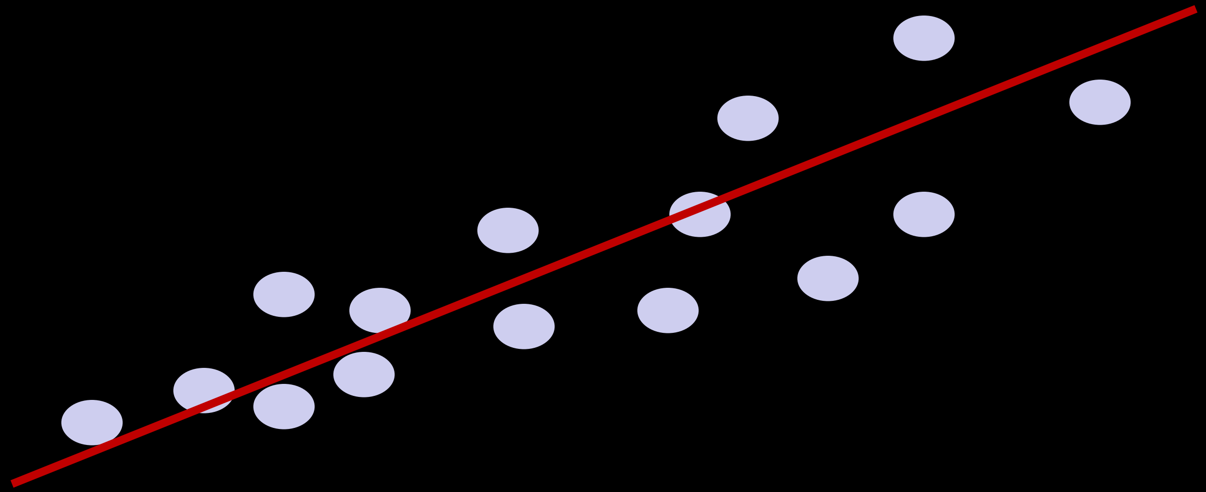
Its main purpose is to reduce a correlated multidimensional data set to an uncorrelated lower dimensional space with maximum variance.

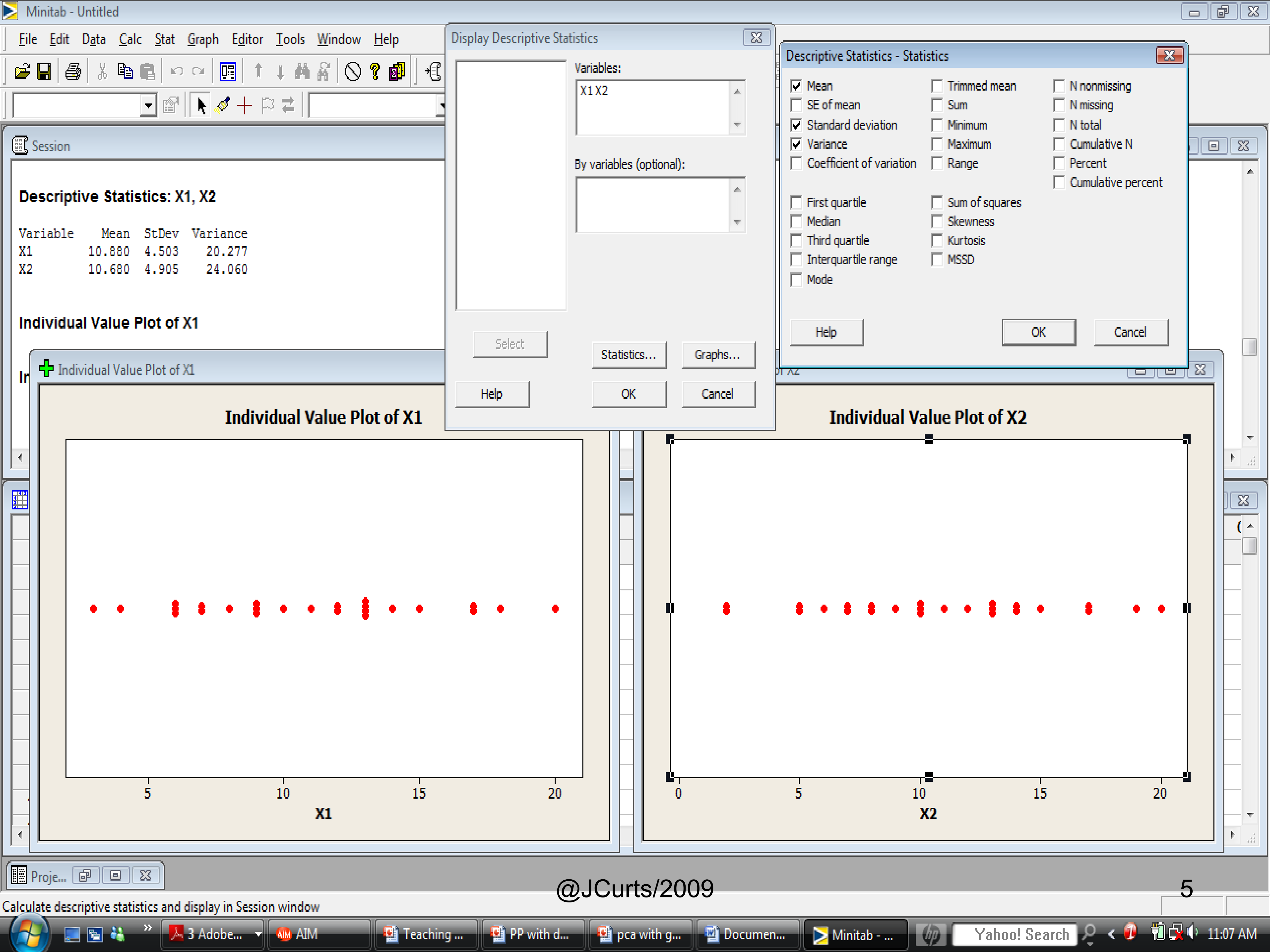
PCA concepts can be a roadblock for non-mathematical oriented students, since statistical definitions (i.e., variance-covariance, correlation) need to be connected to matrix algebra (eigenvectors of a variance-covariance matrix) and to graphical vector representation (including matrix rotation).



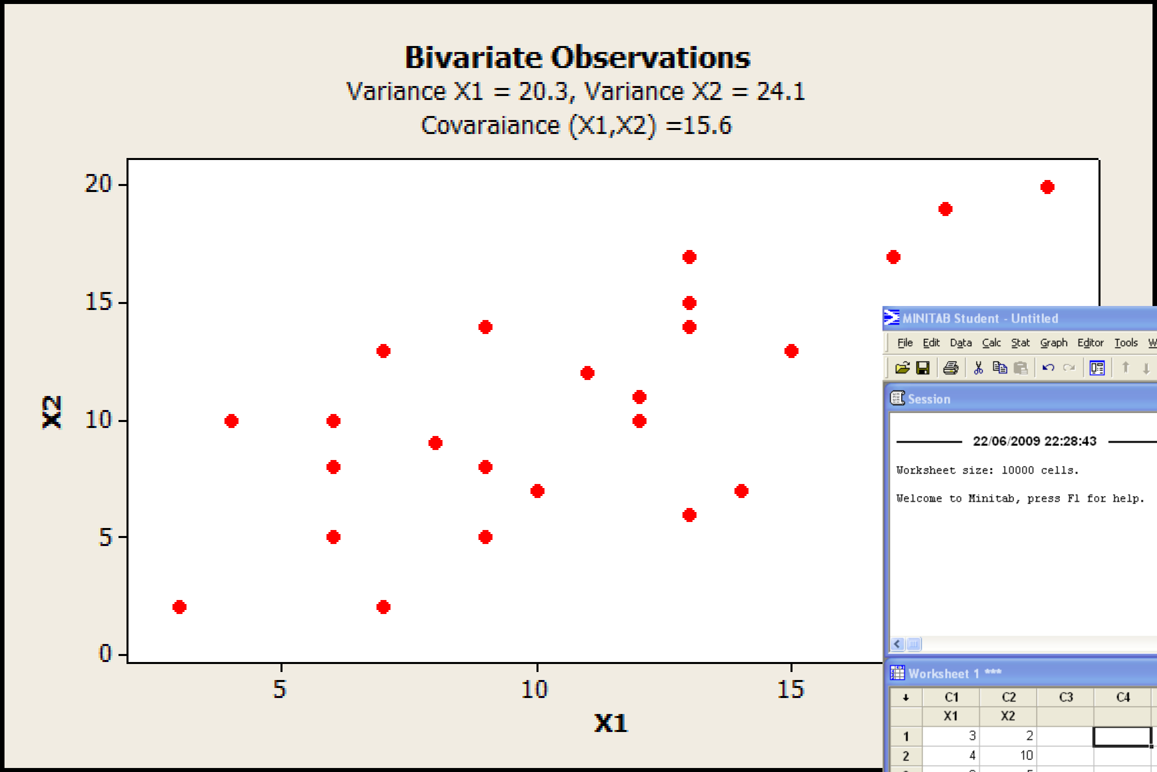
- Given m points in a n dimensional space, for large n, how does one project on to a low dimensional space while preserving broad trends in the data and allowing it to be visualized?

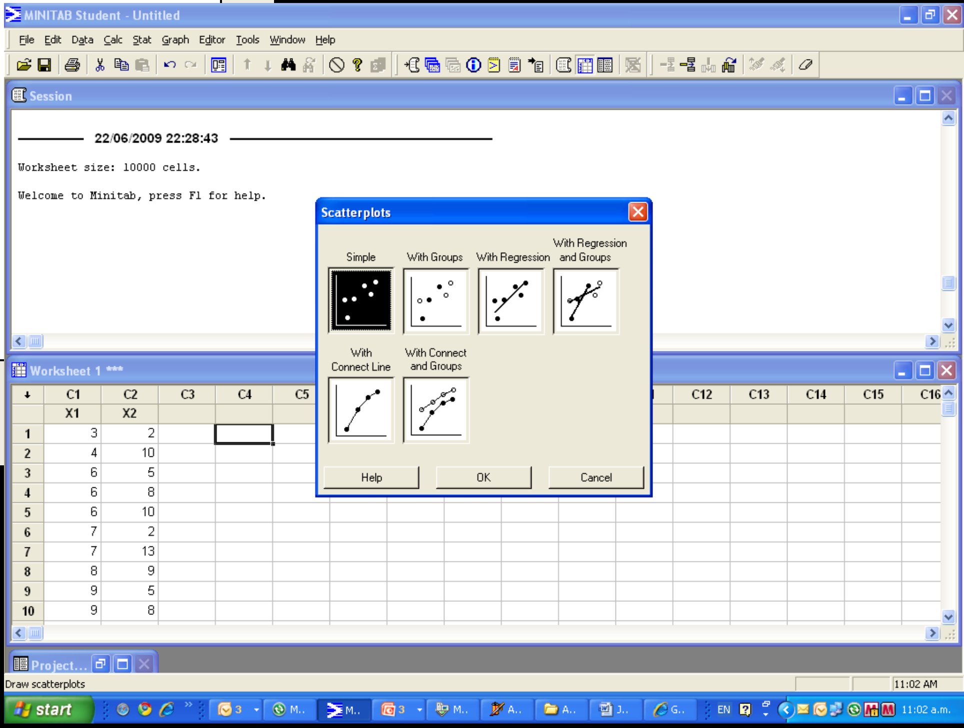
- Choose a line that fits the data so the points are spread




Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Descriptive Statistics: X1, X2
Variable Mean StDev Variance
X1 10.880 4.503 20.277
X2 10.680 4.905 24.060
Individual Value Plot of X1
Display Descriptive Statistics
Variables:
X1 X2
By variables (optional):
Select
Statistics...
Graphs...
Help
OK
Cancel
Descriptive Statistics - Statistics
Mean
SE of mean
Standard deviation
Variance
Coefficient of variation
First quartile
Median
Third quartile
Interquartile range
Mode
Trimmed mean
Sum
Minimum
Maximum
Range
Sum of squares
Skewness
Kurtosis
MSSD
N nonmissing
N missing
N total
Cumulative N
Percent
Cumulative percent
Help
OK
Cancel
Individual Value Plot of X1
X1
Individual Value Plot of X2
X2
Calculate descriptive statistics and display in Session window

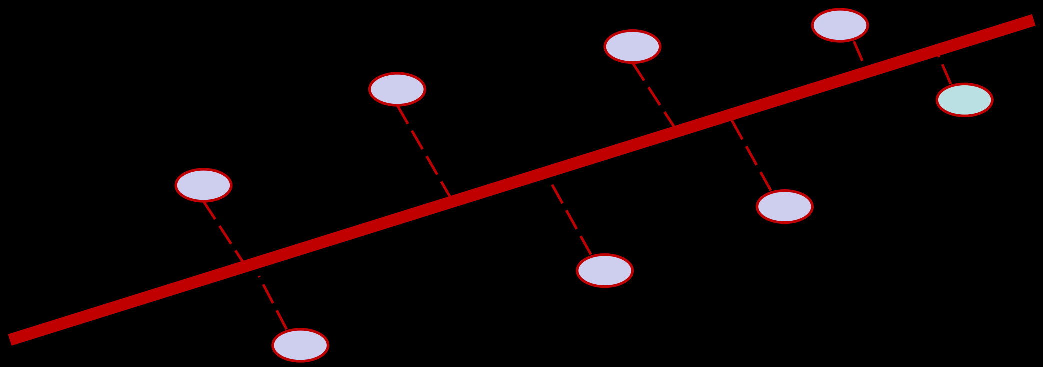
A sample of *n* observations in the 2-D space



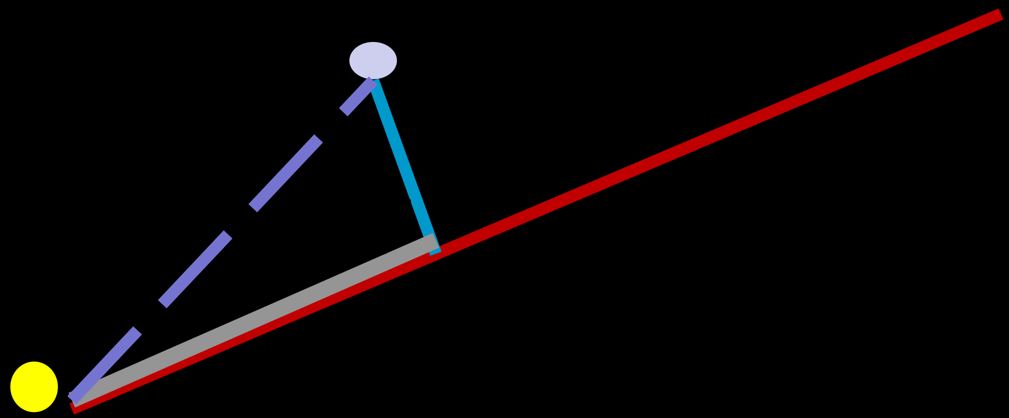


Goal: to account for the variation in a sample in as few variables as possible, to some accuracy

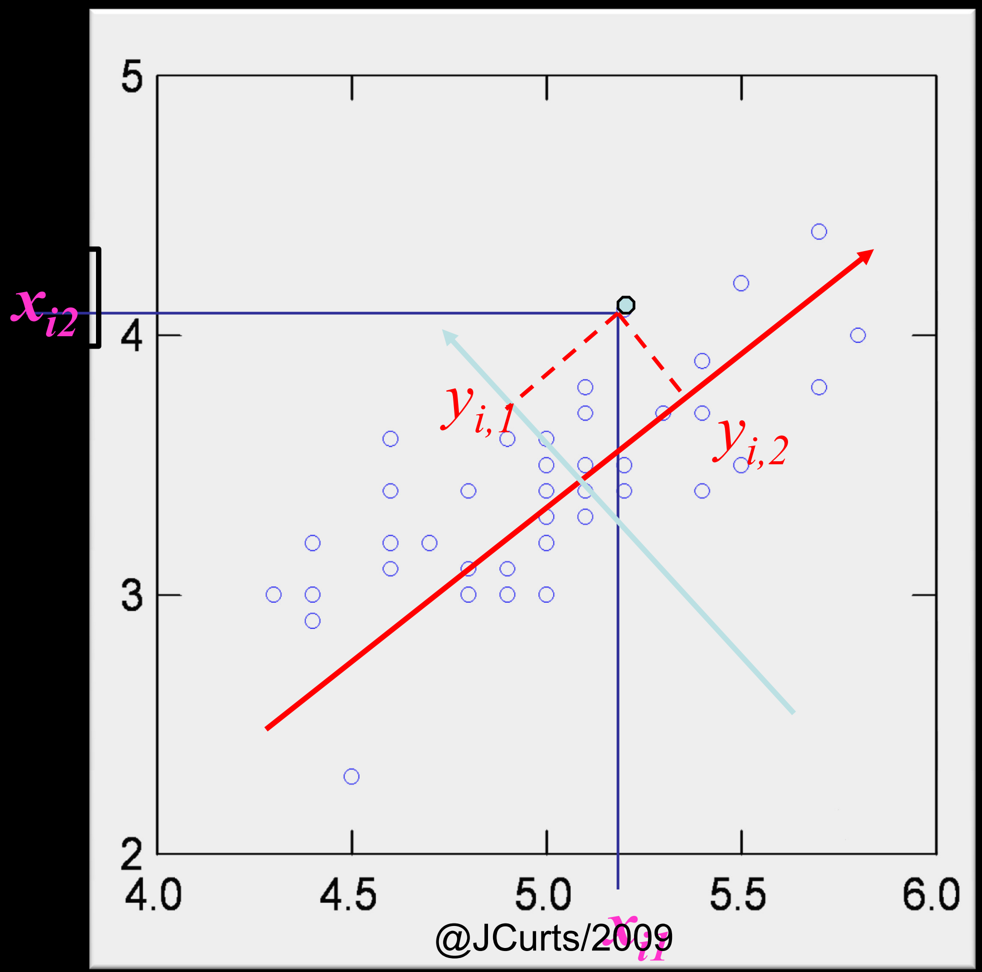
Formally, minimize sum of squares of distances to the line.



Why sum of squares? Because it allows fast minimization, assuming the line passes through 0



Minimizing sum of squares of distances to the line is the same as maximizing the sum of squares of the projections on that line, thanks to Pythagoras.



5
4
3
2
4.0
4.5
5.0
5.5
6.0
x_{i2}
x_{i1}
$y_{i,1}$
$y_{i,2}$
@JCurts/2009

Y1 and Y2 are new coordinates.

- Y1 represents the direction where the data values have the largest uncertainty.
- Y2 is perpendicular to Y1.

To find Y1 and Y2, we need to make transformation from X1 and X2. To simplify the discussion, we move the origin to $(\bar{x}_1, \bar{x}_2)$ and redefine the (X1,X2) coordinate as

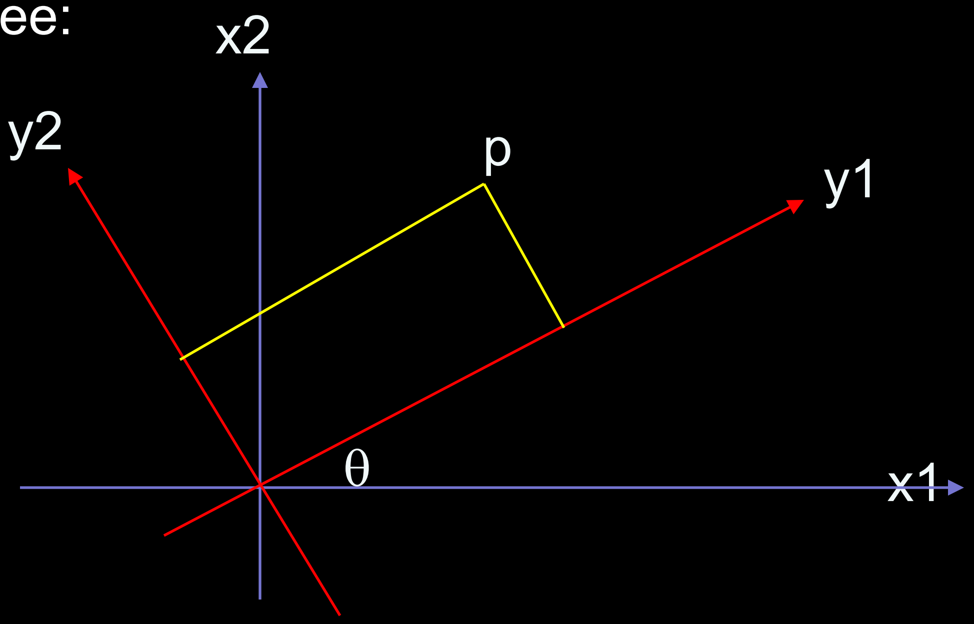
x1 = X1 - $\bar{x}_1$, x2 = X2 - $\bar{x}_2$, so that the origin is (0,0).

The relationship is illustrated in the following graph. We would like to present the data of a given lab, p = (x1,x2) in terms of p = (y1,y2).

From basic geometry relations, we see:

y1 = (cosθ) x1 + (sinθ) x2

y2 = (-sinθ) x1 + (cosθ) x2



The angle θ is determined so that the observations along the Y1 axis has the largest variability.
But HOW?

Minitab - Untitled

File Edit Data Calc Stat Graph Editor Tools Window Help

Session

Descriptive Statistics: X1, X2, Z1, Z2, Y1 with thet, Y1 with thet, ...

Variable	N	N*	Mean	StDev	Variance
X1	25	0	10.880	4.503	20.277
X2	25	0	10.680	4.905	24.060
Z1	25	0	0.000	1.000	1.000
Z2	25	0	0.000	1.000	1.000
Y1 with theta = 5	25	0	-7.155	3.909	15.277
Y1 with theta = 10	25	0	-14.94	5.97	35.62
Y1 with theta = 45	25	0	14.80	6.08	36.95
Y1 = theta = 90	25	0	4.673	3.289	10.814

Calculator

C1 X1
C2 X2
C3 Z1
C4 Z2
C5 Y1 with theta =
C6 Y1 with theta =
C7 Y1 with theta =
C8 Y1 = theta = 9

Store result in variable: c8

Expression:

COS(90) * X1+SIN(90) * X2

Functions: All functions

Absolute value
Antilog
Any
Arcsine
Arccosine
Arctangent

Assign as a formula

Select | Select | Help | OK | Cancel

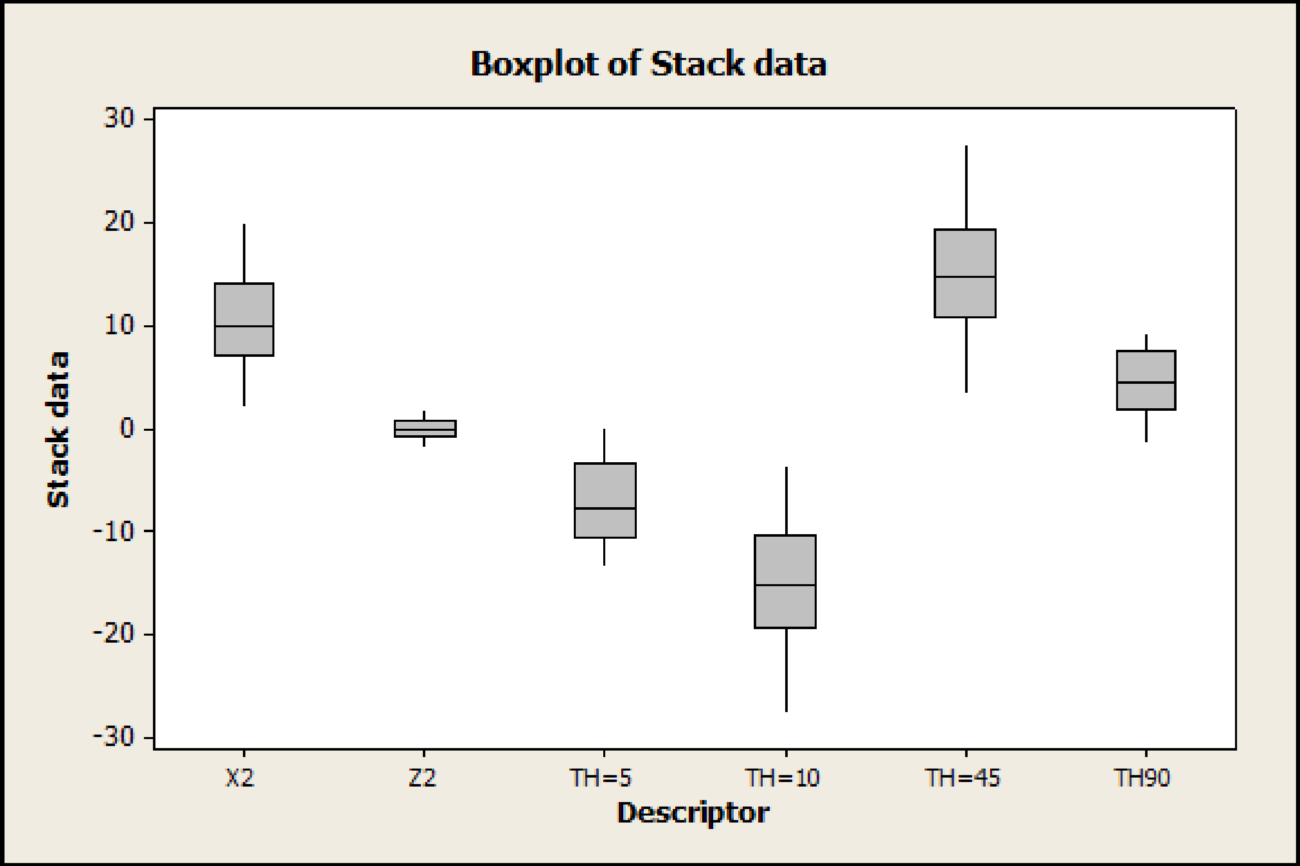
ACA data one.MTW ***

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
	X1	X2	Z1	Z2	Y1 with theta = 5	Y1 with theta = 10	Y1 with theta = 45	Y1 = theta = 90							
16	13	6	0.47080	-0.95411	-2.0659	-14.1721	11.9346	-0.46098							
17	13	14	0.47080	0.67685	-9.7373	-18.5242	18.7418	6.69100							
18	13	15	0.47080	0.88072	-10.6963	-19.0682	19.5927	7.58499							
19	13	17	0.47080	1.28846	-12.6141	-20.1563	21.2945	9.37299							
20	14	7	0.69288	-0.75024	-2.7412	-15.5551	13.3108	-0.01505							
21	15	13	0.91495	0.47298	-8.2111	-19.6583	18.9416	4.90085							
22	17	13	1.35911	0.47298	-7.6438	-21.3365	19.9922	4.00471							
23	17	17	1.35911	1.28846	-11.4795	-23.5126	23.3958	7.58069							
24	18	19	1.58118	1.69619	-13.1136	-25.4397	25.6230	8.92061							
25	20	20	2.02533	1.90006	-13.5052	-27.6619	27.5245	8.91846							

Calculate ranges of expressions using calculator functions

Editable

3 Adobe Reader 8.1 | AIM | 3 Microsoft Offic... | Document1 - Micr... | Minitab - Untitled | Yahoo! Search | 12:44 PM


Boxplot of Stack data
Stack data
30
20
10
0
-10
-20
-30
X2
Z2
TH=5
TH=10
TH=45
TH90
Descriptor

For any given value of theta, then, it is a simple matter to work out the values of Y1 for each of our twenty observations. When θ is 5 degrees, for example, the calculations are:

Z1	Z2	New Variable Y1
1.90	0.47	1.94
0.99	0.85	1.06
1.22	0.09	1.22
0.54	-0.68	0.47
0.31	0.47	0.35
0.08	-1.25	-0.03
-0.15	-0.30	-0.17
-0.60	0.09	-0.59
-1.06	-1.63	-1.20
-1.29	-1.25	-1.39
-1.74	-1.63	-1.88
-1.52	-1.06	-1.60
0.76	0.47	0.80
1.90	2.57	2.12
0.31	0.28	0.33
-0.38	-0.10	-0.38
-0.15	0.66	-0.09
-0.15	0.85	-0.07
-0.38	0.66	-0.32
-0.60	0.47	-0.56

Mean (X1) = 7.65; VAR (X1) = 19.23
Mean (Z1) = Mean (Z2) = 0;
Variance (Z1) = Variance (Z2) = 1

VAR (Y1) = $(\cos\theta)\, x1 + (\sin\theta)\, x2$ = **1.12**

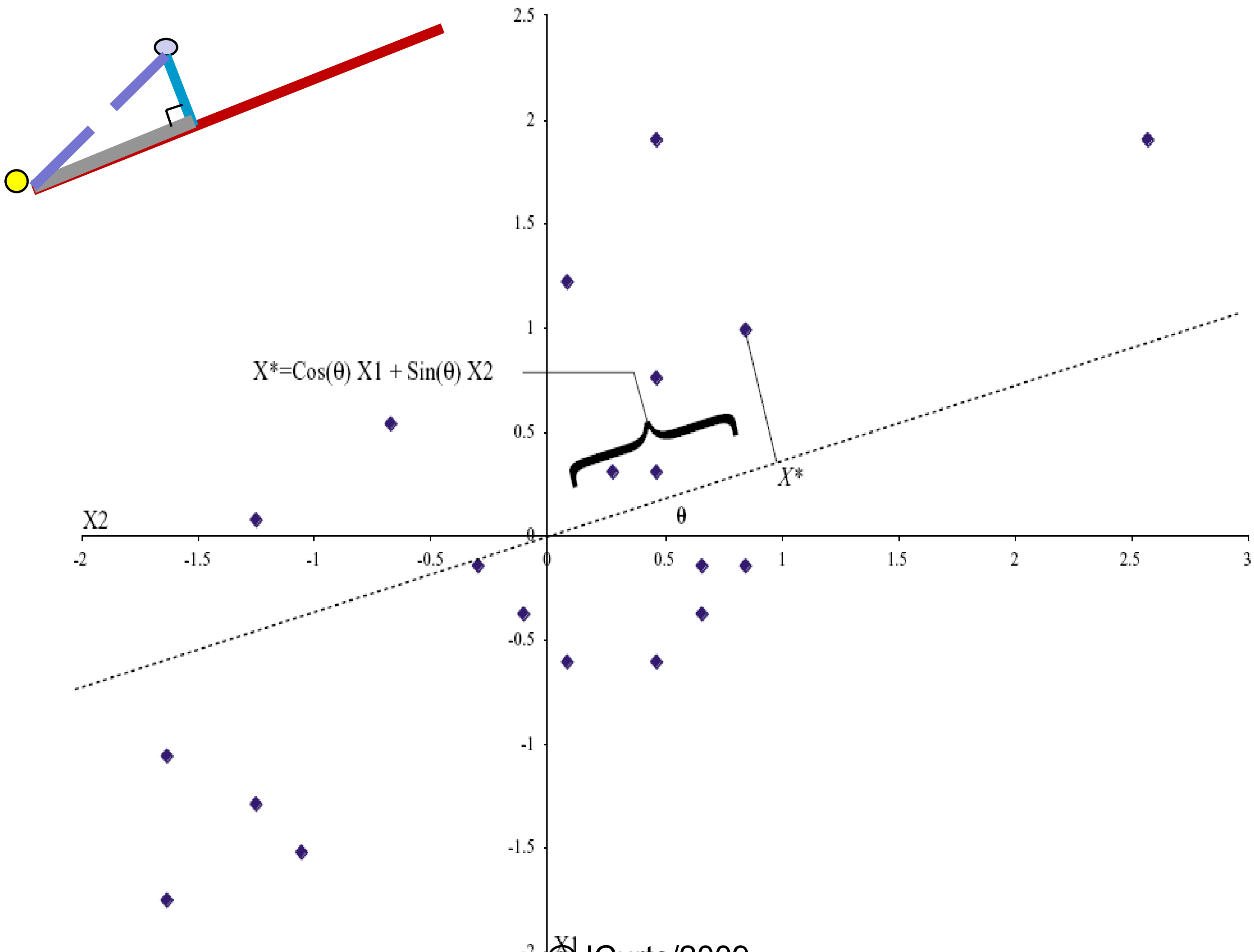
Note that each of the original variables has a variance of 1.0, but the variance of the new axis is 1.12, which constitutes more than half of the total variance for the entire dataset (e.g., 1.12/2.00 or 56%).

Each value of theta will yield a different set of scores on Y1, and will also result in distinct values for the variance term. If we calculate transformed values and variances for different values of theta, we can compare the variance of the new axis to the total for our dataset.

Note that as we increase the angle, the new variable accounts for an increasing fraction of total variance, until 45 degrees, **and then declines**; by the time theta is 90 degrees, the new axis is equivalent to X2, and, not surprisingly, its proportion of variance is back to 1.00 or 50.0%.

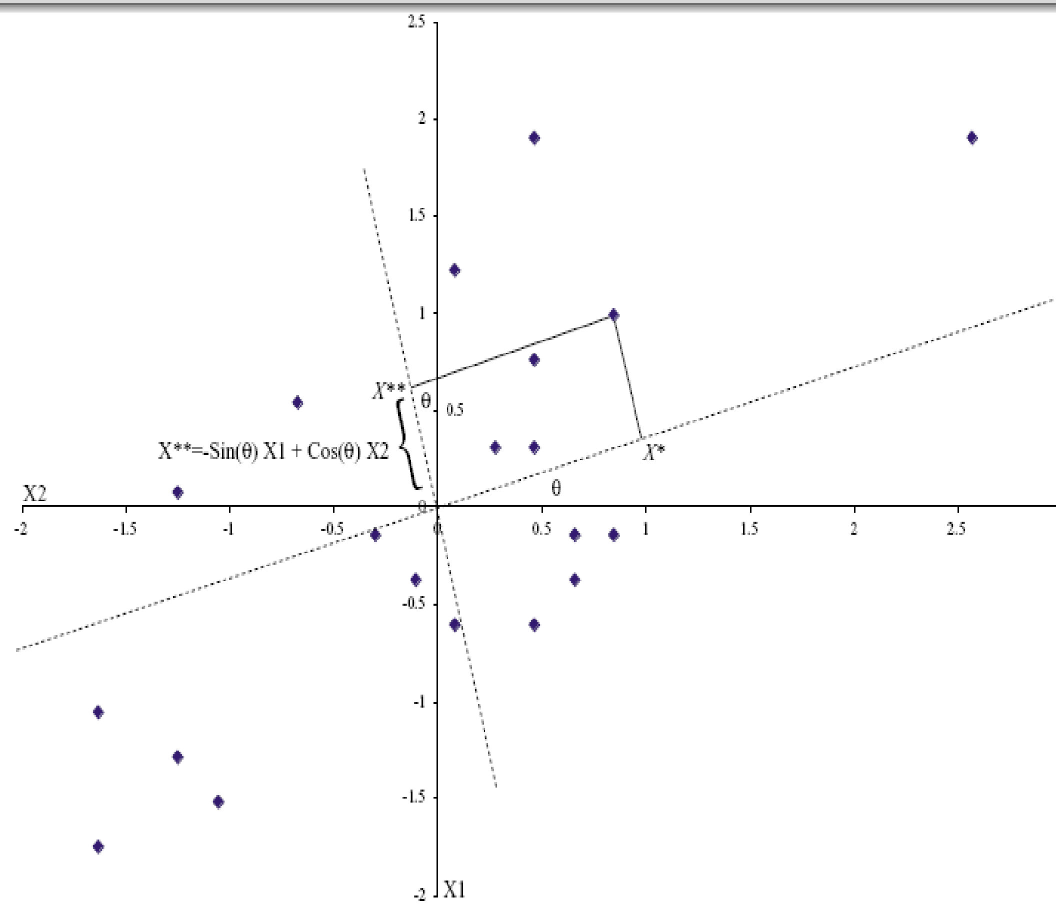
Theta	Var (Y1)	Proportion
5	1.121	56.0%
10	1.238	61.9%
15	1.348	67.4%
20	1.447	72.4%
25	1.533	76.7%
30	1.603	80.1%
35	1.654	82.7%
40	1.685	84.3%
45	**1.696**	**84.8%**
50	1.685	84.3%
55	1.654	82.7%
60	1.603	80.1%
65	1.533	76.7%
70	1.447	72.4%
75	1.348	67.4%
80	1.238	61.9%
85	1.121	56.0%
90	1.000	50.0%

606_pca reference to gould and explanation of theta.pdf - Adobe Reader
File Edit View Document Tools Window Help
7 / 27 130% Find



Principal Componet... Principal COmpone... Microsoft PowerPoi... 606_pca reference t... Yahoo! Search 9:19 PM

606_pca reference to gould and explanation of theta.pdf - Adobe Reader
File Edit View Document Tools Window Help
9 / 27 100% Find



Note that this time, the variance term is much smaller than that of the each of the original variables. But the variances of the two new axes sum to 2.0 -- the total variation in the original dataset. This is the basic approach of principal components analysis: obtaining linear combinations of variables into new axes, such that the first one accounts for the largest share of total variance, the second is orthogonal to the first and accounts for less variance, etc. Several properties hold for these components.

The transformation from (x1,x2) to (y1,y2) results several nice properties

1. The variability along y1 is largest.
2. Y1 and y2 are uncorrelated, that is, orthogonal.
3. The confidence region based on (y1,y2) is easy to construct, and provide useful interpretations of the two sample plots.

Questions remain unanswered are

1. How to determine the angle θ so that the variability of observations along the y1 axis is maximized?
2. How to construct the ellipse for confidence region with different levels of confidences?
3. How to interpret the two-sample plots?

✓How to determine the Y1 and Y2 axis so that the variability of observations along the Y1 axis is maximized and Y2 is orthogonal to Y1?

✓Rewrite the linear relation between (Y1,Y2) and (x1,x2) in matrix notation:

Y1 = (cosθ) x1 + (sinθ) x2

Y2 = (-sinθ) x1 + (cosθ) x2

$$Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} (\cos\theta)x_1 + (\sin\theta)x_2 \\ (-\sin\theta)x_1 + (\cos\theta)x_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} A_1'X \\ A_2'X \end{bmatrix} = AX$$

NOTE: X is bivariate , so is Y, and

V(X) = $\begin{bmatrix} V(x_1) & Cov(x_1,x_2) \\ Cov(x_1,x_2) & V(x_2) \end{bmatrix}$, V(Y) = A'V(X)A = $\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$

λ_1 and λ_2 are called the eigen values. Which are the solutions of $|V(X) - \lambda I| = 0$

And, V(Y1) = λ_1 , V(Y2) = λ_2, Correlation between Y1 and Y2 = 0.

λ_1 and λ_2 are called the eigen values. Which are the solutions of
And, V(Y1) = λ_1 , V(Y2) = λ_2, Correlation between Y1 and Y2 = 0.

The angle θ = $$(.5)\arctan\left(\frac{2\rho\sigma_1\sigma_2}{\sigma_1^2-\sigma_2^2}\right)$$ if $\sigma_1 \neq \sigma_2$

when $\sigma_1 = \sigma_2$, θ = 45^o The angle θ = $$= \arctan\left(\frac{\lambda_1-\sigma_1^2}{\rho\sigma_1\sigma_2}\right)$$

Note the angle depends on the correlation between X1 and X2 , as well as, on the variances of X1 and X2, respectively.

- When ρ is close to zero, the angle is also close to zero. If V(X1) and V(X2) are close, then, the scatter plots are scattered like a circle. That is, there is no clear major principal component.
- When ρ is close to zero and V(X1) is much larger than V(X2), then, the angle will be close to zero, and the data points are likely to be parallel to the X-axis. On the other hand, if V(X1) is much smaller than V(X2), the angle will be close to 90^0, and the data points will be more likely parallel to the Y-axis.

Consider, now, we actually observe the following two sample data:

$$\begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{bmatrix}$$

The sample means are given by $\begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$

The sample variance-covariance matrix is given by

$$\hat{V}(X) = \begin{bmatrix} s_1^2 & rs_1s_2 \\ rs_1s_2 & s_2^2 \end{bmatrix}$$

r is the Pearson's correlation coefficient, and S^2 is the sample variance. S is the sample standard deviation.

V(Y) is the solution of $\left|\hat{V}(X) - \lambda I\right| = 0$

The solutions for λ are given by

$$\frac{(s_1^2 + s_2^2) \pm \sqrt{(s_1^2 + s_2^2)^2 - 4(1 - r^2)(s_1^2 s_2^2)}}{2}$$

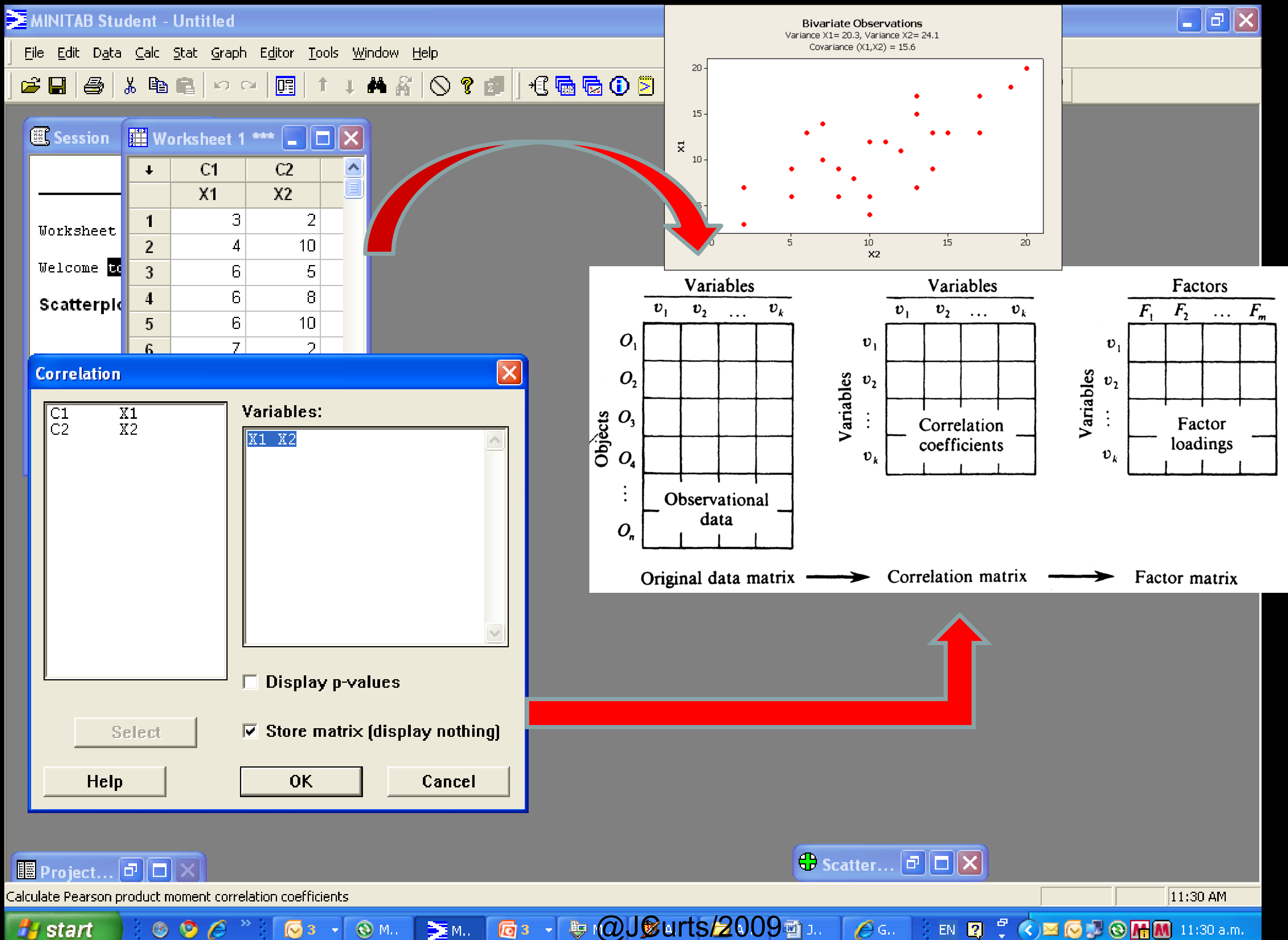
NOTE: V(Y1) + V(Y2) = $\lambda_1 + \lambda_2 = s_1^2 + s_2^2$ = V(X1) + V(X2)

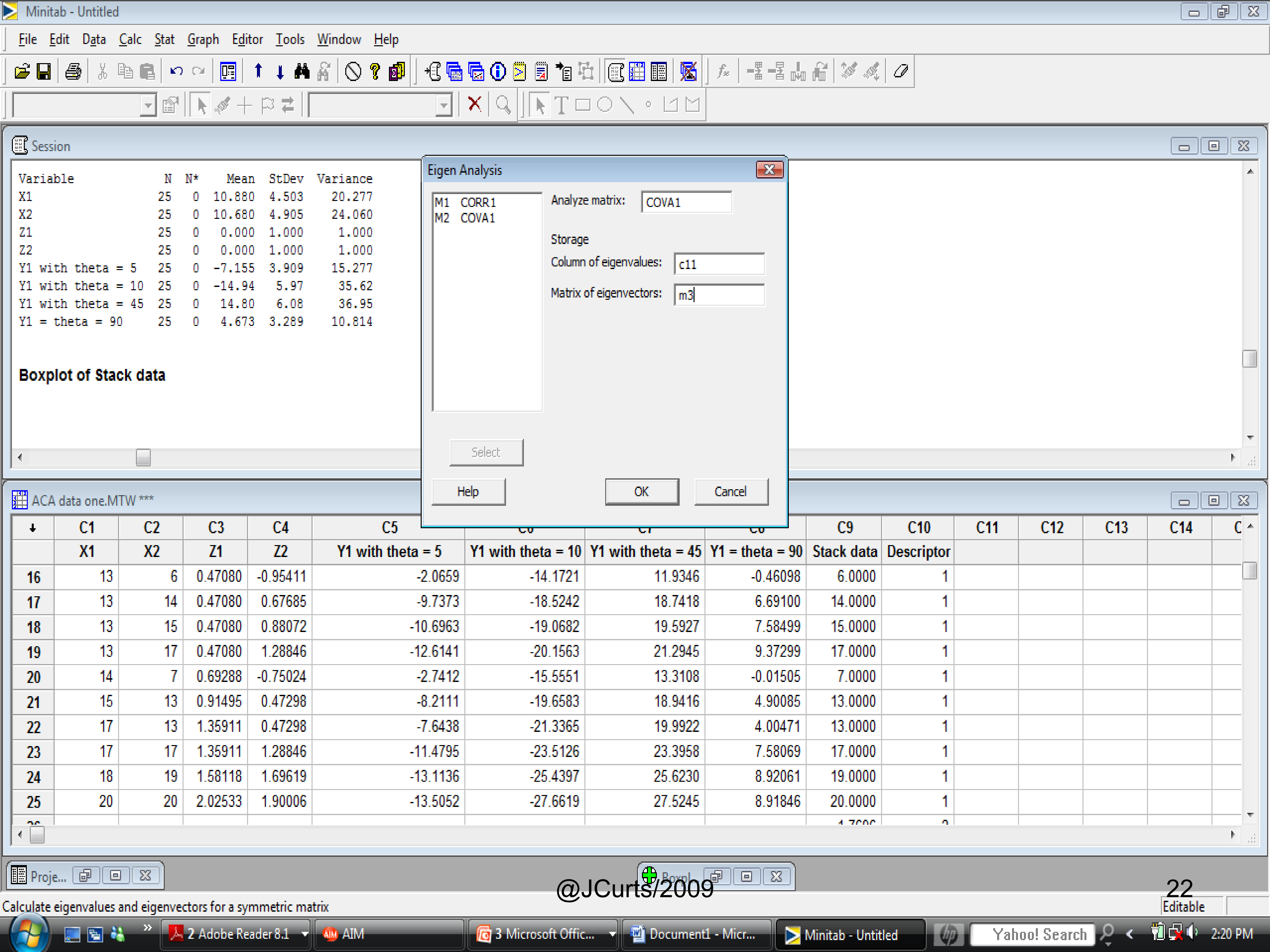
Using the sample data, the angle is estimated by

$$\theta = (.5)\arctan\left(\frac{2rs_1s_2}{s_1^2 - s_2^2}\right) \qquad s_1 \neq s_2$$

$$= \arctan\left(\frac{\lambda_1 - s_1^2}{rs_1s_2}\right)$$

Correlation and Covariance Matrix





Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Variable N N* Mean StDev Variance
X1 25 0 10.880 4.503 20.277
X2 25 0 10.680 4.905 24.060
Z1 25 0 0.000 1.000 1.000
Z2 25 0 0.000 1.000 1.000
Y1 with theta = 5 25 0 -7.155 3.909 15.277
Y1 with theta = 10 25 0 -14.94 5.97 35.62
Y1 with theta = 45 25 0 14.80 6.08 36.95
Y1 = theta = 90 25 0 4.673 3.289 10.814
Boxplot of Stack data
Eigen Analysis
M1 CORR1
M2 COVA1
Analyze matrix: COVA1
Storage
Column of eigenvalues: c11
Matrix of eigenvectors: m3
Select
Help
OK
Cancel
ACA data one.MTW ***
C1 X1 | C2 X2 | C3 Z1 | C4 Z2 | C5 Y1 with theta = 5 | Y1 with theta = 10 | Y1 with theta = 45 | Y1 = theta = 90 | C9 Stack data | C10 Descriptor | C11 | C12 | C13 | C14
16 13 6 0.47080 -0.95411 -2.0659 -14.1721 11.9346 -0.46098 6.0000 1
17 13 14 0.47080 0.67685 -9.7373 -18.5242 18.7418 6.69100 14.0000 1
18 13 15 0.47080 0.88072 -10.6963 -19.0682 19.5927 7.58499 15.0000 1
19 13 17 0.47080 1.28846 -12.6141 -20.1563 21.2945 9.37299 17.0000 1
20 14 7 0.69288 -0.75024 -2.7412 -15.5551 13.3108 -0.01505 7.0000 1
21 15 13 0.91495 0.47298 -8.2111 -19.6583 18.9416 4.90085 13.0000 1
22 17 13 1.35911 0.47298 -7.6438 -21.3365 19.9922 4.00471 13.0000 1
23 17 17 1.35911 1.28846 -11.4795 -23.5126 23.3958 7.58069 17.0000 1
24 18 19 1.58118 1.69619 -13.1136 -25.4397 25.6230 8.92061 19.0000 1
25 20 20 2.02533 1.90006 -13.5052 -27.6619 27.5245 8.91846 20.0000 1
Calculate eigenvalues and eigenvectors for a symmetric matrix
Editable
Adobe Reader 8.1
AIM
3 Microsoft Offic...
Document1 - Micr...
Minitab - Untitled
Yahoo! Search
2:20 PM

@JCurts/2009

Minitab - Untitled - [Session]

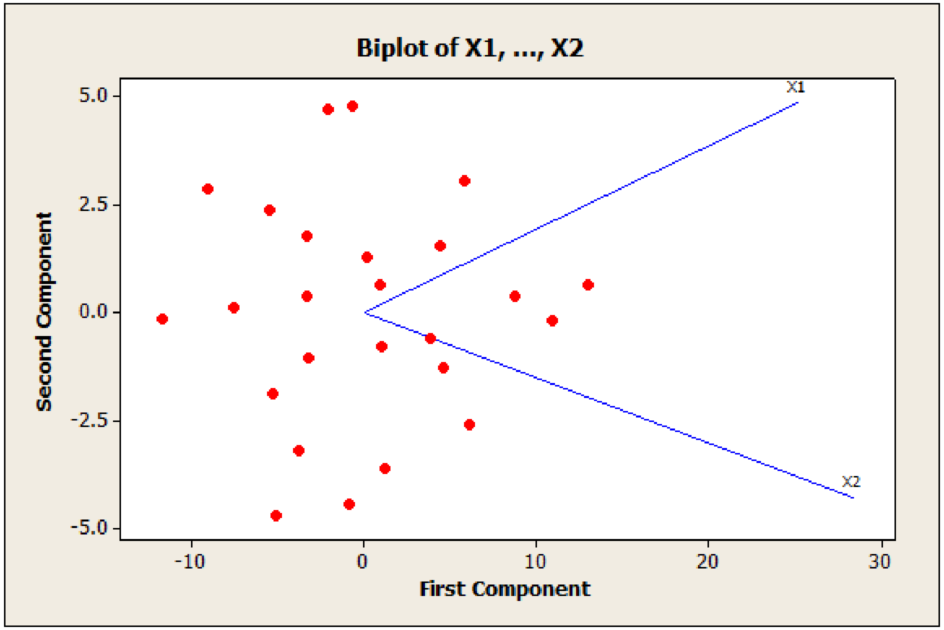
File Edit Data Calc Stat Graph Editor Tools Window Help

Boxplot of Stack data

print

Data Display

```
 Matrix CORR1

1.00000  0.70560
0.70560  1.00000


 Matrix COVA1

20.2767  15.585
15.5850  24.060


 Matrix M3

0.663139   0.748496
0.748496  -0.663139
```



Principal Component Analysis: X1, X2

```
Eigenanalysis of the Covariance Matrix

Eigenvalue  37.868  6.469
Proportion   0.854  0.146
Cumulative   0.854  1.000


Variable    PC1     PC2
X1         0.663   0.748
X2         0.748  -0.663
```

@JCurts/2009

Welcome to Minitab, press F1 for help.   Editable

2 Adobe Reader 8.1 AIM 3 Microsoft Offic... Document1 - Micr... Minitab - Untitled -... Yahoo! Search 2:29 PM

Minitab - Untitled

File Edit Data Calc Stat Graph Editor Tools Window Help

Principal Components Analysis

Variables: X1 X2

Number of components to compute:

Type of Matrix
- ( ) Correlation
- (•) Covariance

Select | Graphs... | Storage... | Help | OK | Cancel

Principal Components Analysis - Storage

C1 X1
C2 X2
C3 Z1
C4 Z2
C5 Y1 with theta =
C6 Y1 with theta =
C7 Y1 with theta =
C8 Y1 = theta = 90
C9 Stack data
C10 Descriptor
C11
C12
C13
C14
C15
C16

Coefficients: c12 c13

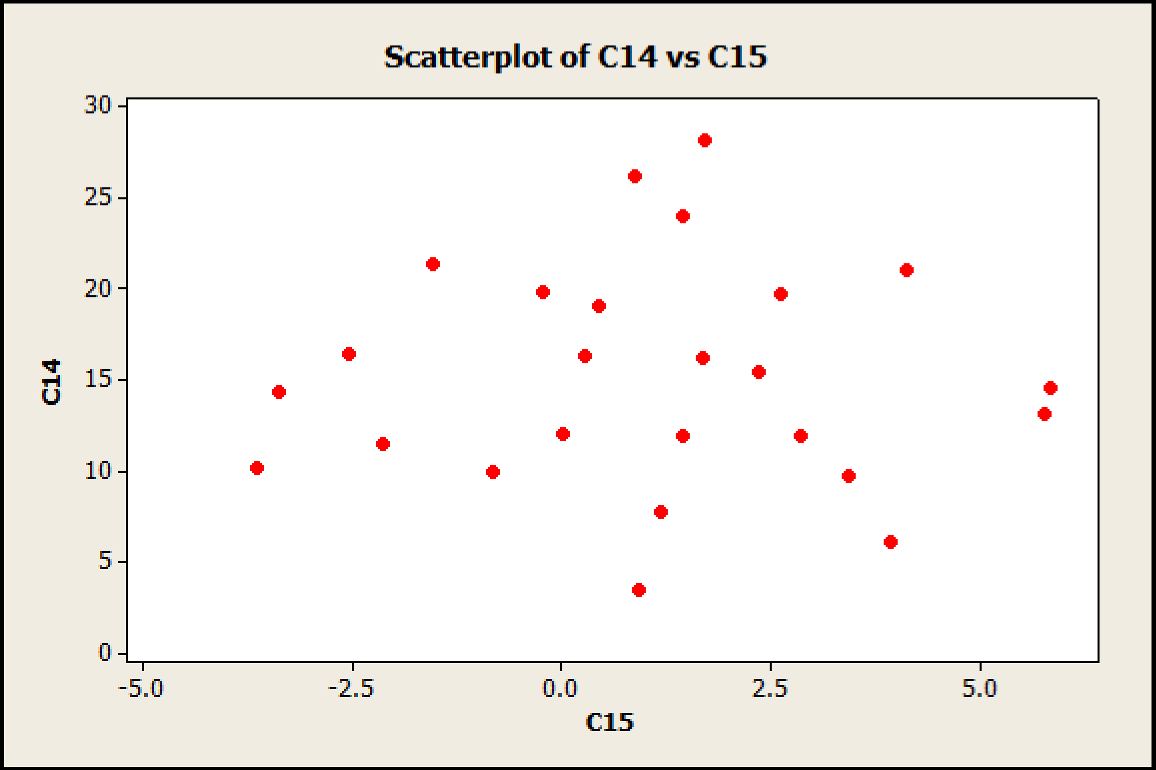
Scores: c14 c15

Eigenvalues: c16

Select | Help | OK | Cancel

					C8	C9	C10	C11	C12	C13	C14	C15	C16	C17
				a = 45	Y1 = theta = 90	Stack data	Descriptor							
1	-1.76959	-1.0669	-3.6053	3.2778	0.44377	2.0000	1	37.8677	0.663139	0.748496	3.4864	0.91921	37.8677	
2	-0.13863	-8.4546	-8.7965	10.6103	7.14767	10.0000	1	6.4690	0.748496	-0.663139	10.1375	-3.63741	6.4690	
3	-1.15798	-3.0926	-7.7545	7.4064	1.78154	5.0000	1				7.7213	1.17528		
4	-0.54637	-5.9694	-9.3866	9.9592	4.46353	8.0000	1				9.9668	-0.81414		
5	-0.13863	-7.8873	-10.4746	11.6610	6.25152	10.0000	1				11.4638	-2.14041		
6	-1.76959	0.0678	-6.9615	5.3791	-1.34852	2.0000	1				6.1390	3.91320		
7	0.47298	-10.4804	-12.9458	14.7390	8.48544	13.0000	1				14.3724	-3.38134		
8	-0.34250	-6.3610	-11.6088	11.8607	4.46138	9.0000	1				12.0416	0.01972		
9	-1.15798	-2.2417	-10.2717	8.9824	0.43732	5.0000	1				9.7107	3.42077		
10	-0.54637	-5.1184	-11.9038	11.5351	3.11931	8.0000	1				11.9562	1.43135		

Perform principal components analysis (PCA)



Pearson correlation of C14 and C15 = -0.000

Thanks……………………..

jbcurts@utpa.edu