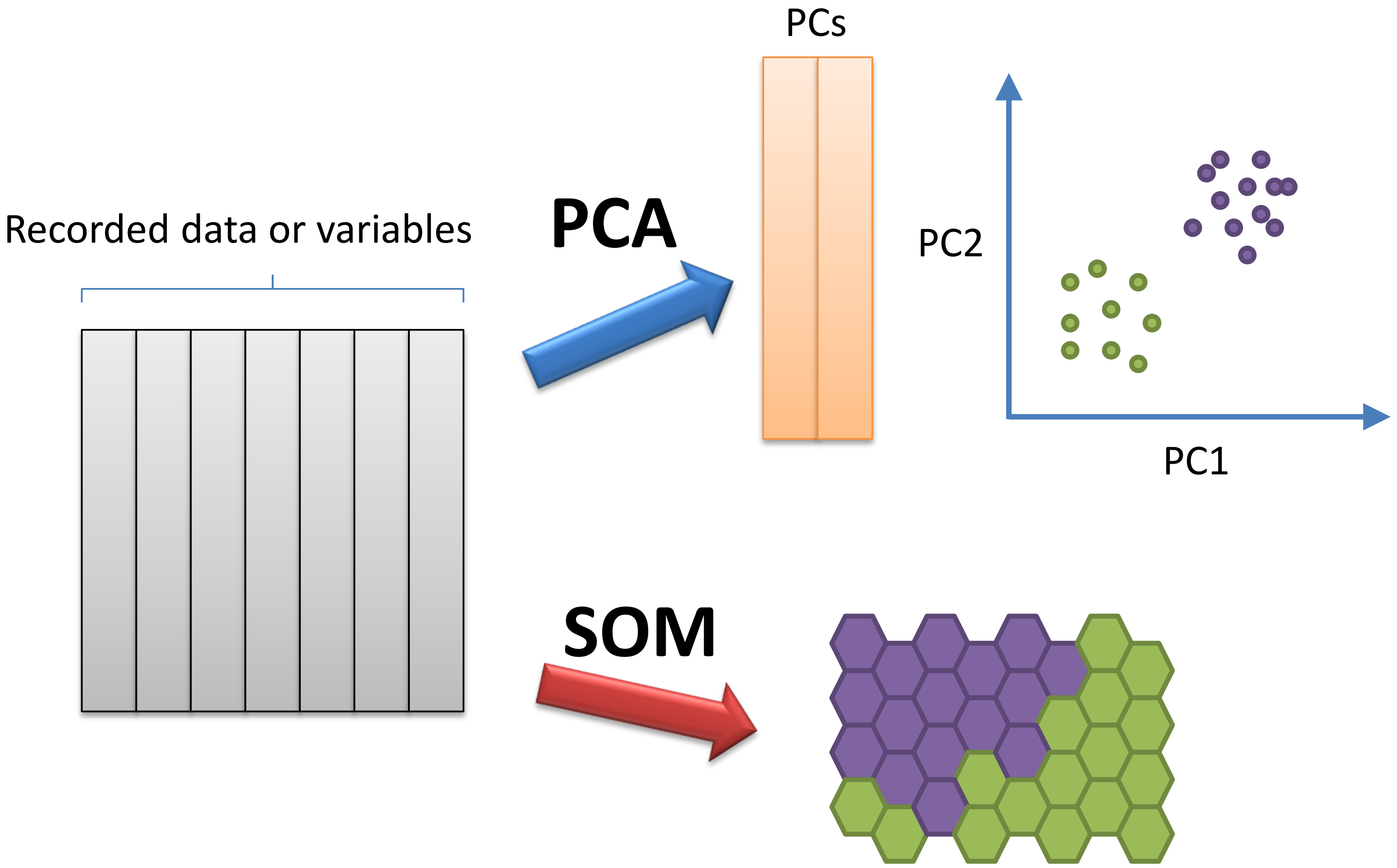
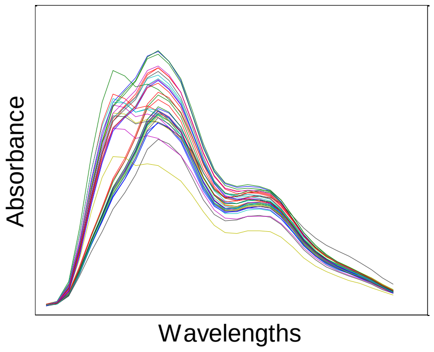
# Exploratory data analysis (EDA)

- The aim of EDA is to detect the similarity or dissimilarity in data.
- To answer:
  - – What is the relationship between samples and between variables?
  - – Are there any grouping in the data?
  - – What are the trends in the data?
  - – Are there any outliers?
- Principal component analysis (PCA) is the most common EDA method.

**Principal component analysis (PCA)** and **self organizing map (SOM)** are among the most used EDA techniques.

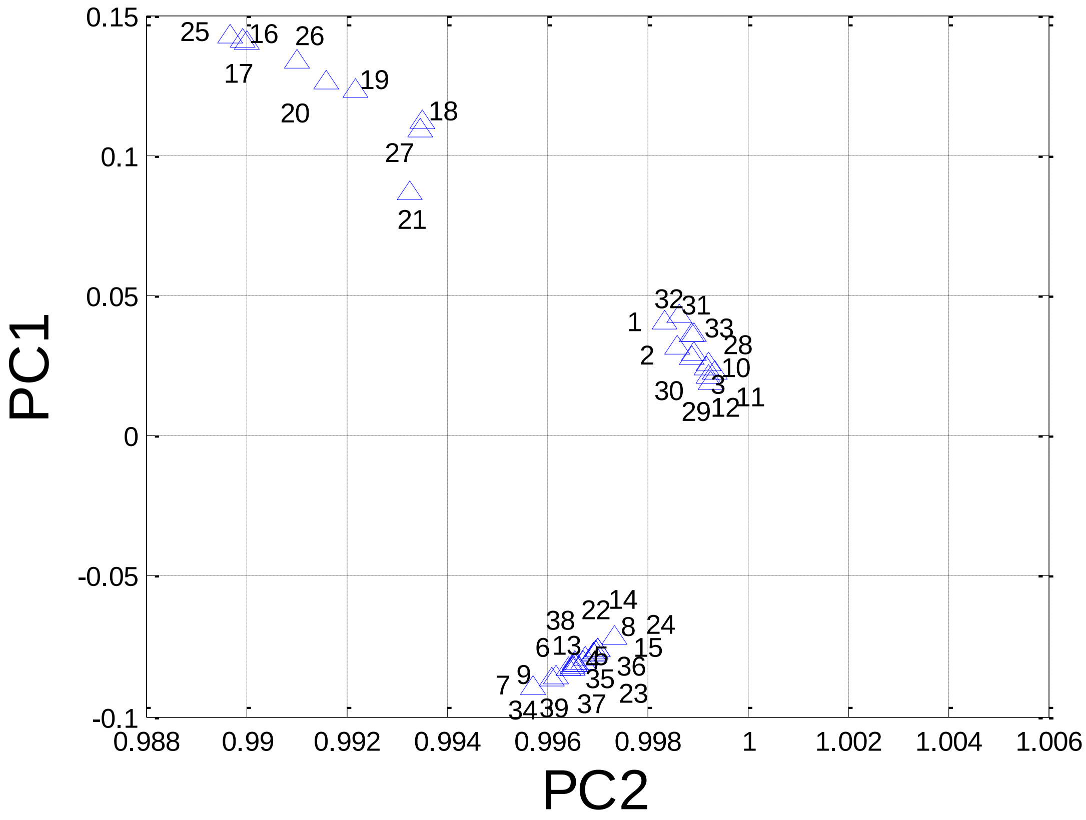


# Principal Component Analysis (PCA)



Spectrum data having 39 samples with 24 variables

**PCA**



**Score plot using PC1 and PC2 of the 39 spectrum data**

- PCA is an abstract mathematical transformation of the original data into some new factors.
- These factors can be more effectively used to represent the variation in the data.
- PCA can be represent by the equation:

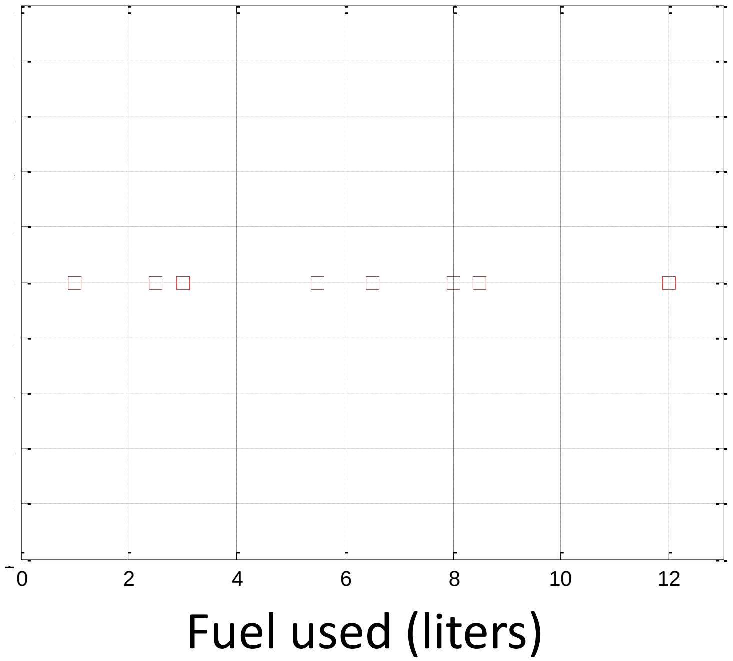
$$\boldsymbol{X = T.P + E}$$

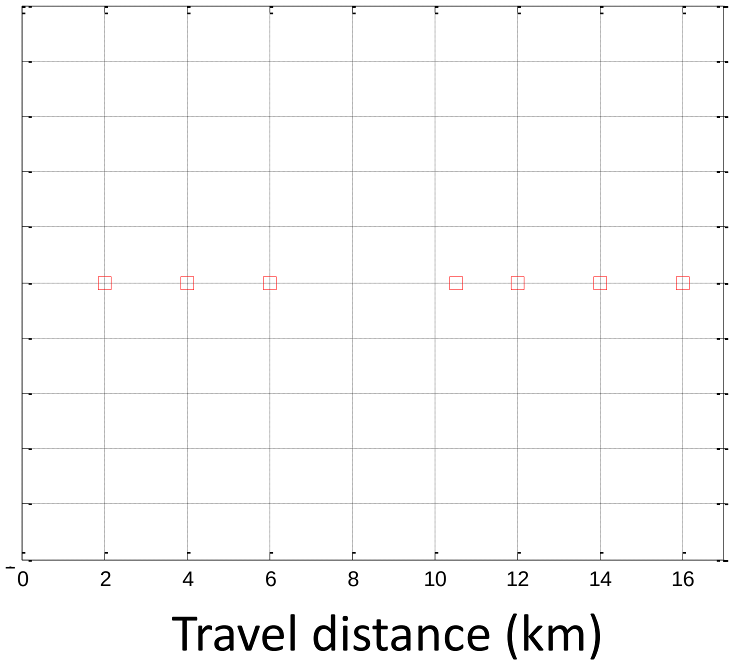
- It is expected to see less complicate data after the PCA transformation.

- A study case

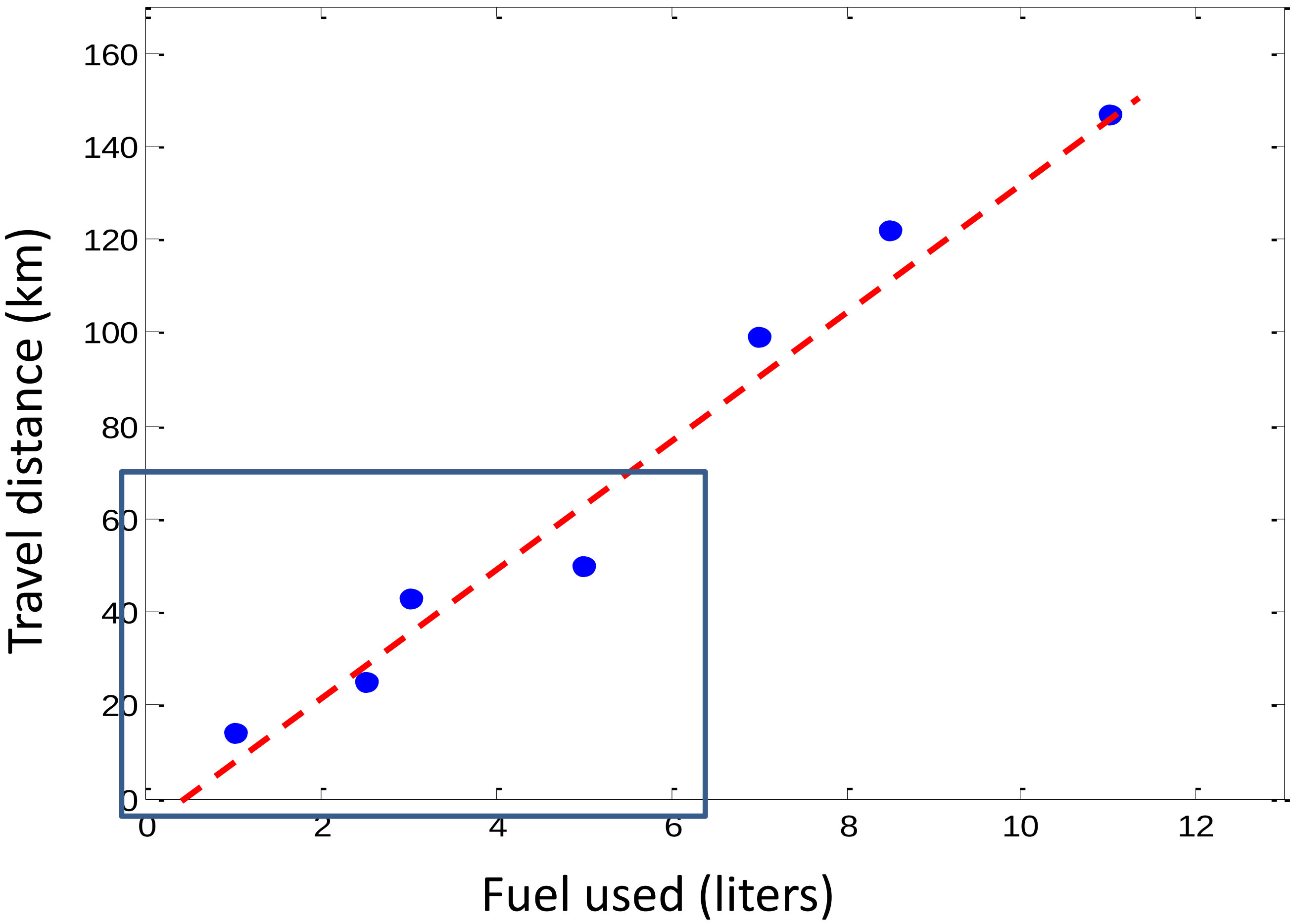
| Exp no. | variable 1<br>(fuel used; liters) | variable 2<br>(distance; km) |
|---|---|---|
| 1 | 1.0 | 2.0 |
| 2 | 2.5 | 4.0 |
| 3 | 3.0 | 6.0 |
| 4 | 5.5 | 6.0 |
| 5 | 6.5 | 10.5 |
| 6 | 8.0 | 12.0 |
| 7 | 8.5 | 14.0 |
| 8 | 12.0 | 16.0 |

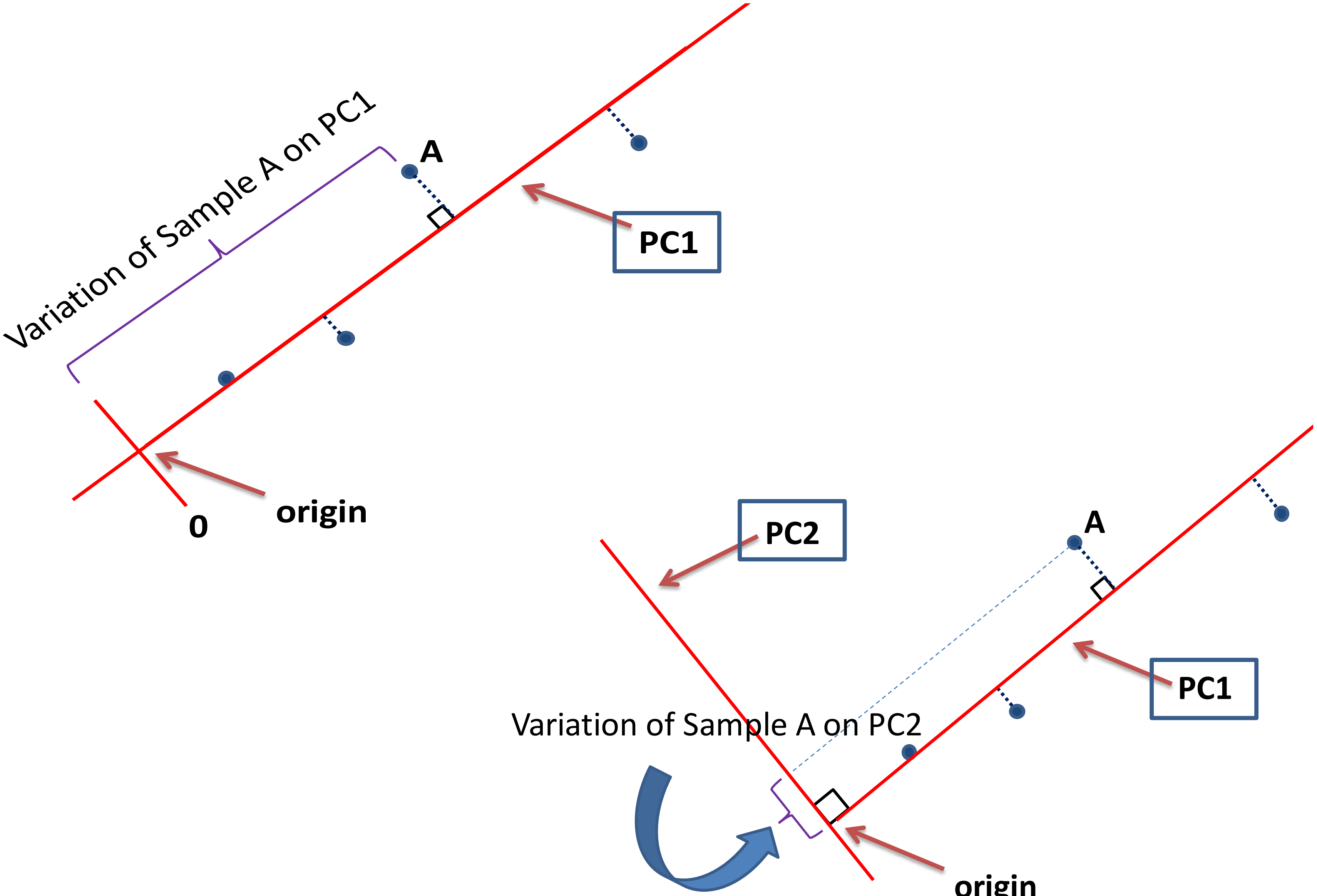
# Data visualization using 1-dimensional graphs





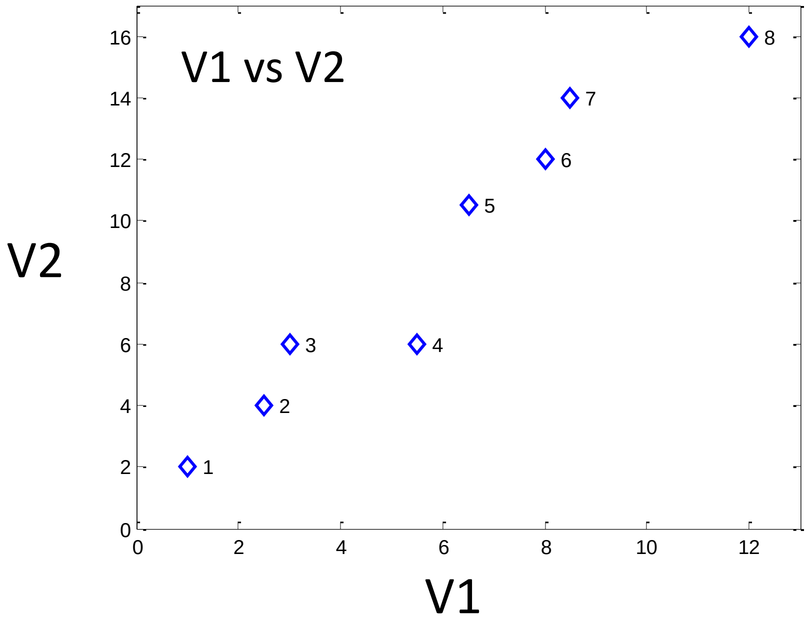
# Data visualization using a 2-dimensional plot

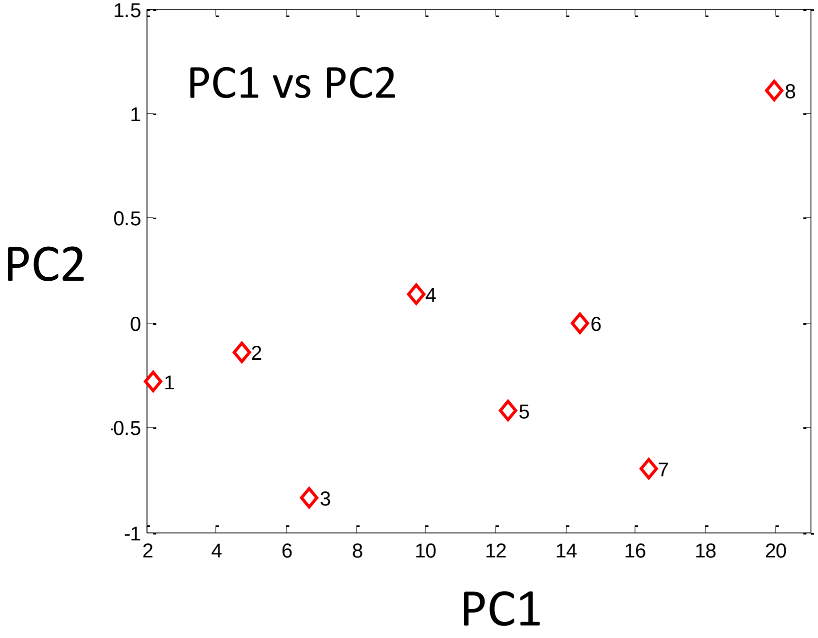




PC → principal component

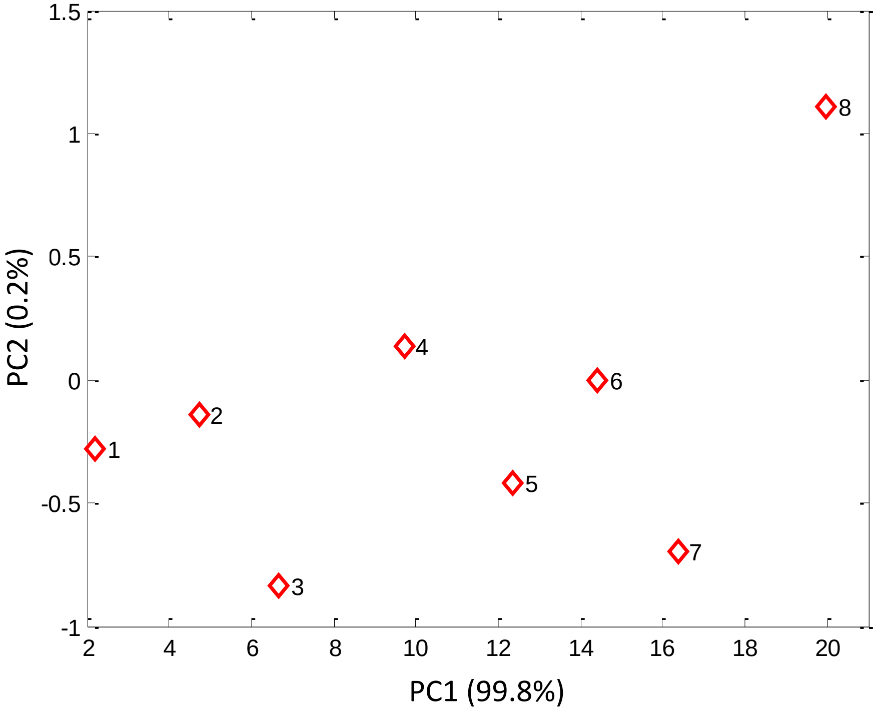
| Sample no. | V1 | V2 | PC1 | PC2 |
|---|---|---|---|---|
| 1 | 1.0 | 2.0 | 2.2 | -0.3 |
| 2 | 2.5 | 4.0 | 4.7 | -0.1 |
| 3 | 3.0 | 6.0 | 6.7 | -0.8 |
| 4 | 5.5 | 6.0 | 9.7 | 0.1 |
| 5 | 6.5 | 10.5 | 12.3 | -0.4 |
| 6 | 8.0 | 12.0 | 14.4 | 0.0 |
| 7 | 8.5 | 14.0 | 16.4 | -0.7 |
| 8 | 12.0 | 16.0 | 20.0 | 1.1 |

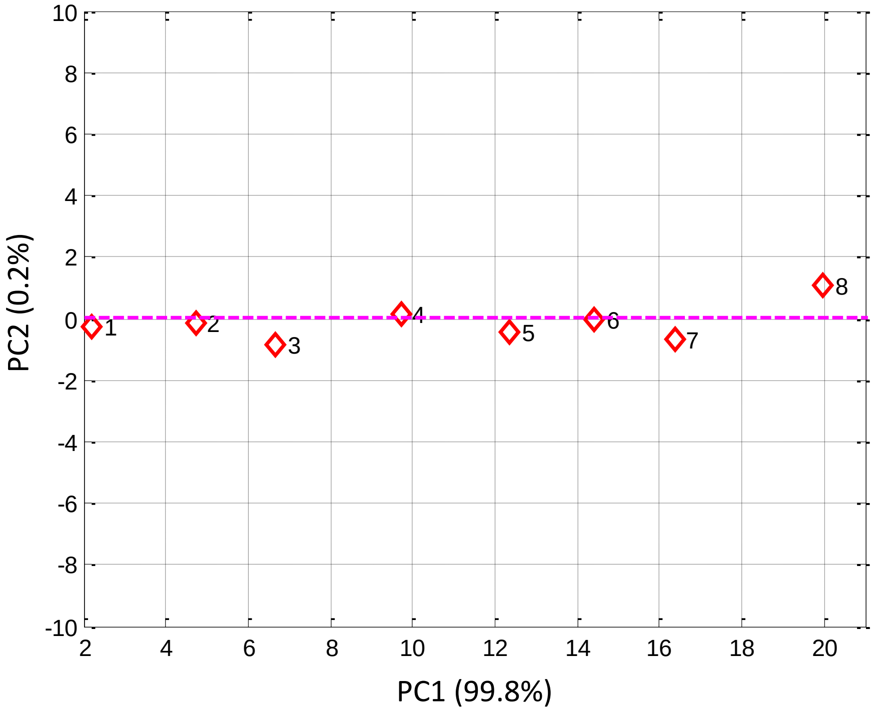




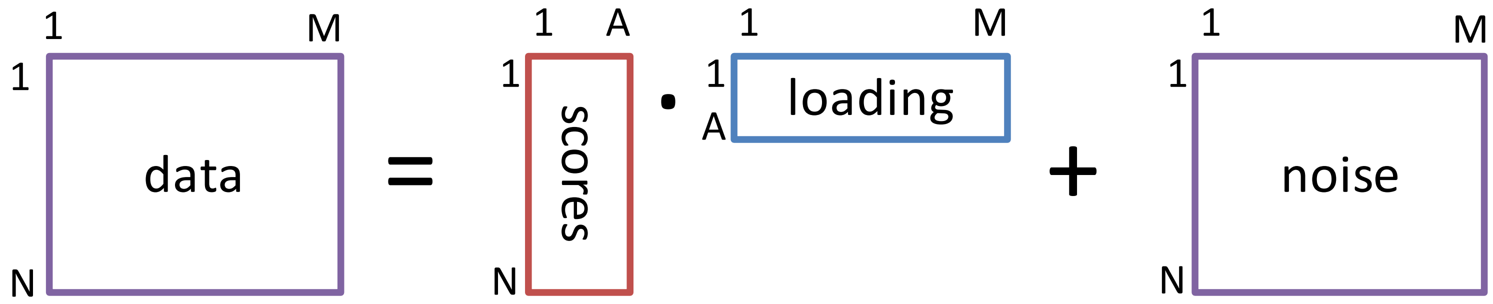
| Sample no. | Variable 1 | | Variable 2 | | PC1 | | PC2 | |
|---|---|---|---|---|---|---|---|---|
| | Value | ^2 | Value | ^2 | Value | ^2 | Value | ^2 |
| 1 | 1.0 | 1.0 | 2.0 | 4.0 | 2.2 | 4.9 | -0.3 | 0.1 |
| 2 | 2.5 | 6.3 | 4.0 | 16.0 | 4.7 | 22.2 | -0.1 | 0.0 |
| 3 | 3.0 | 9.0 | 6.0 | 36.0 | 6.7 | 44.3 | -0.8 | 0.7 |
| 4 | 5.5 | 30.3 | 6.0 | 36.0 | 9.7 | 94.2 | 0.1 | 0.0 |
| 5 | 6.5 | 42.3 | 10.5 | 110.3 | 12.3 | 152.3 | -0.4 | 0.2 |
| 6 | 8.0 | 64.0 | 12.0 | 144.0 | 14.4 | 208.0 | 0.0 | 0.0 |
| 7 | 8.5 | 72.3 | 14.0 | 196.0 | 16.4 | 267.8 | -0.7 | 0.5 |
| 8 | 12.0 | 144.0 | 16.0 | 256.0 | 20.0 | 398.8 | 1.1 | 1.2 |
| Sum of squared | | 369.0 | | 798.3 | | 1192.6 | | 2.7 |
| | 1167.3 | | | | 1195.2 | | | |
| %contribution | 31.6 | | 68.4 | | 99.8 | | 0.2 | |

- PC1 contributes 99.8% of the overall variation whereas PC2 accounts only %0.2
- Only the first PC could be enough to visualize this data.
- PC2 may contain only noise.


1.5
1
0.5
0
-0.5
-1
2
4
6
8
10
12
14
16
18
20
PC1 (99.8%)
PC2 (0.2%)
1
2
3
4
5
6
7
8


10
8
6
4
2
0
-2
-4
-6
-8
-10
2
4
6
8
10
12
14
16
18
20
PC1 (99.8%)
PC2 (0.2%)
1
2
3
4
5
6
7
8

# Calculation of PCA

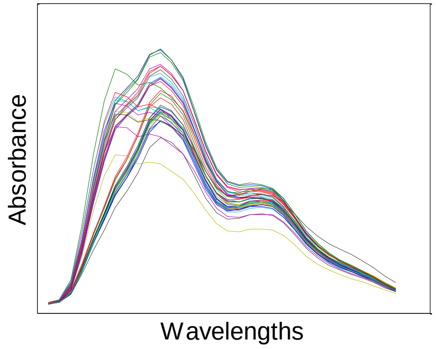


$$X = T.P + E$$

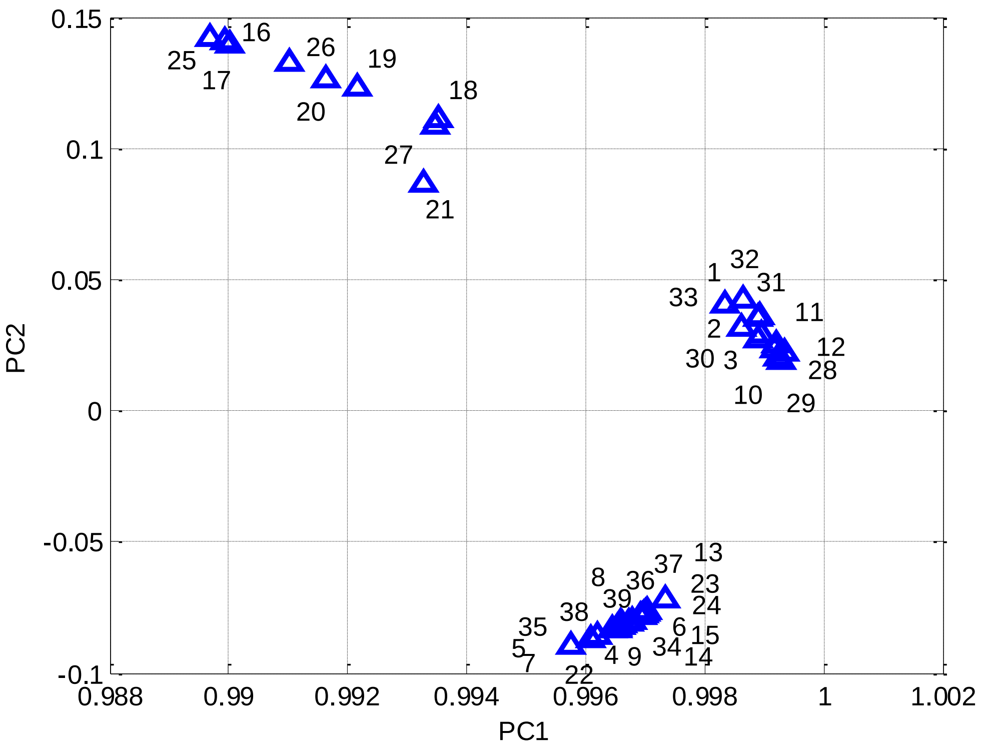
$N$ = Number of samples

$M$ = Number of parameters

$A$ = Number of PCs used in the PCA modelling

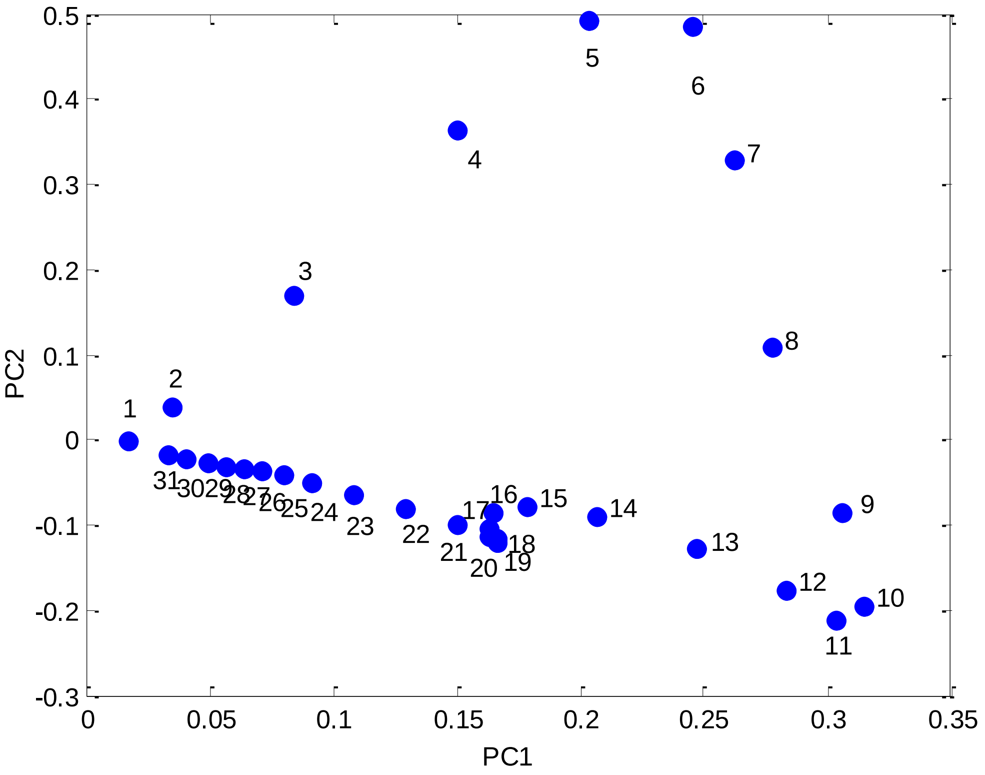


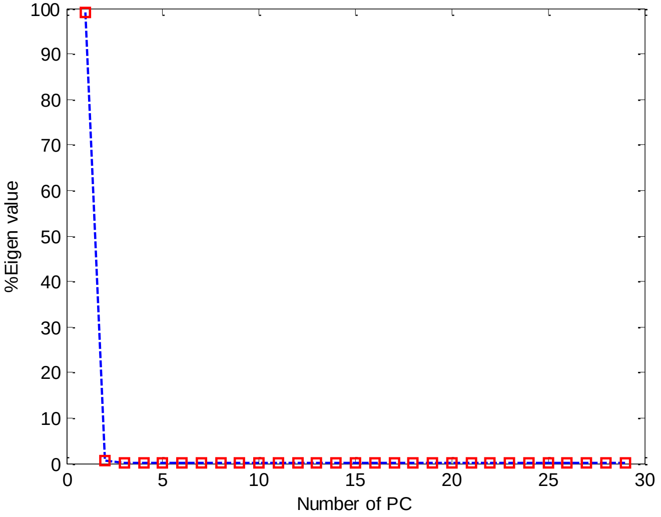
PCA

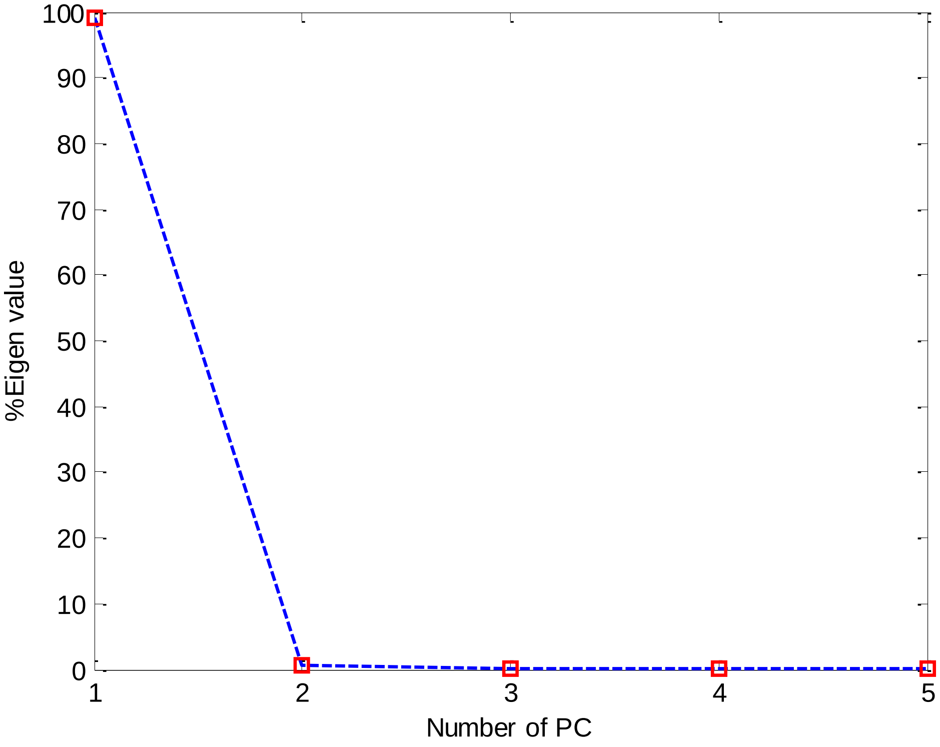


$$[N \times M] = [N \times A].[A \times M] + [N \times M]$$

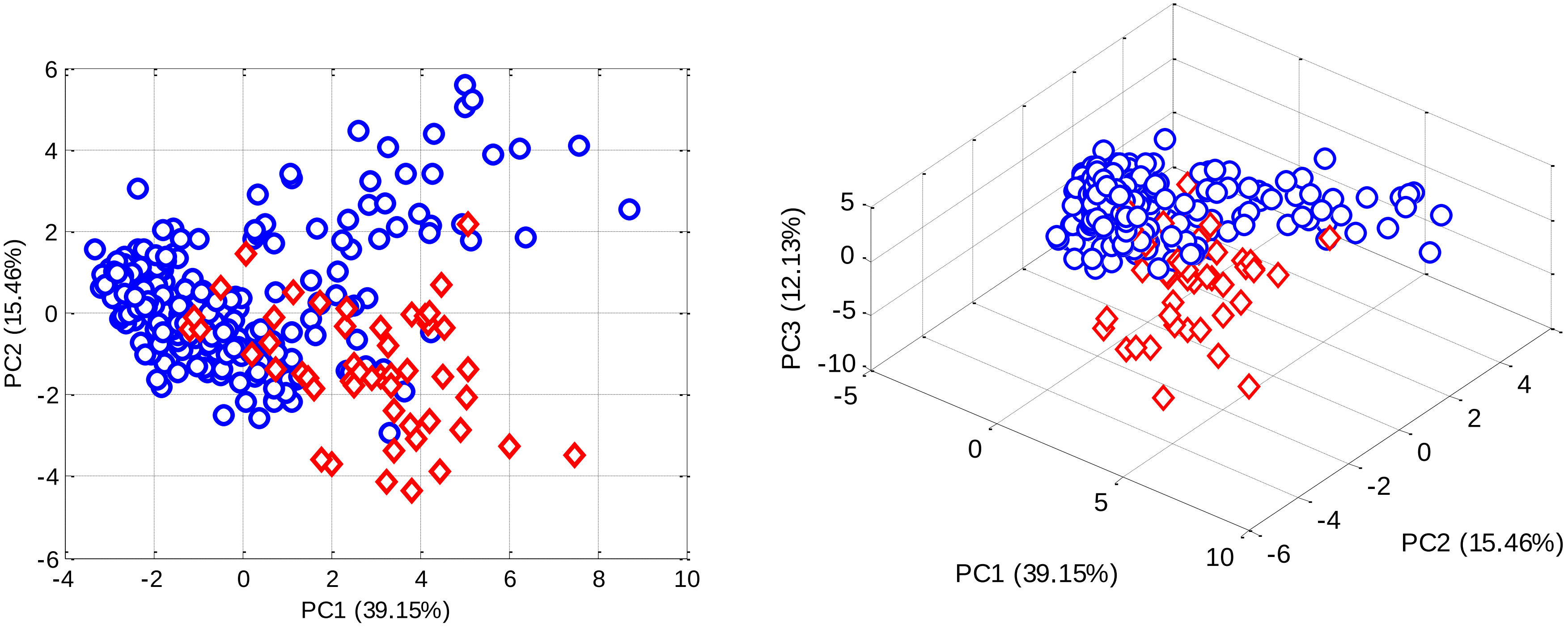
$$[24 \times 39] = [24 \times 2].[2 \times 39] + [24 \times 39]$$


0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
PC2
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
PC1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31


100
90
80
70
60
50
40
30
20
10
0
%Eigen value
0
5
10
15
20
25
30
Number of PC


100
90
80
70
60
50
40
30
20
10
0
%Eigen value
1
2
3
4
5
Number of PC

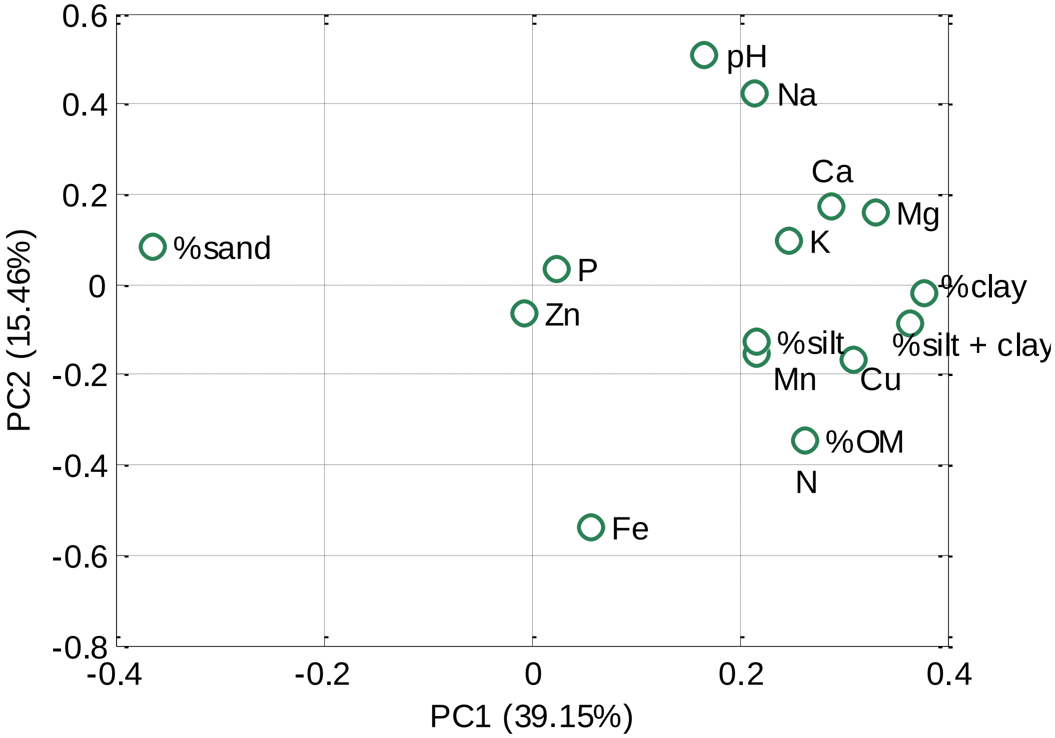
# PCA of physico-chemical parameters data of 704 soil samples from some provinces in the north and northeast of Thailand



## Score (*T*) plot

○ Northeast

◇ North

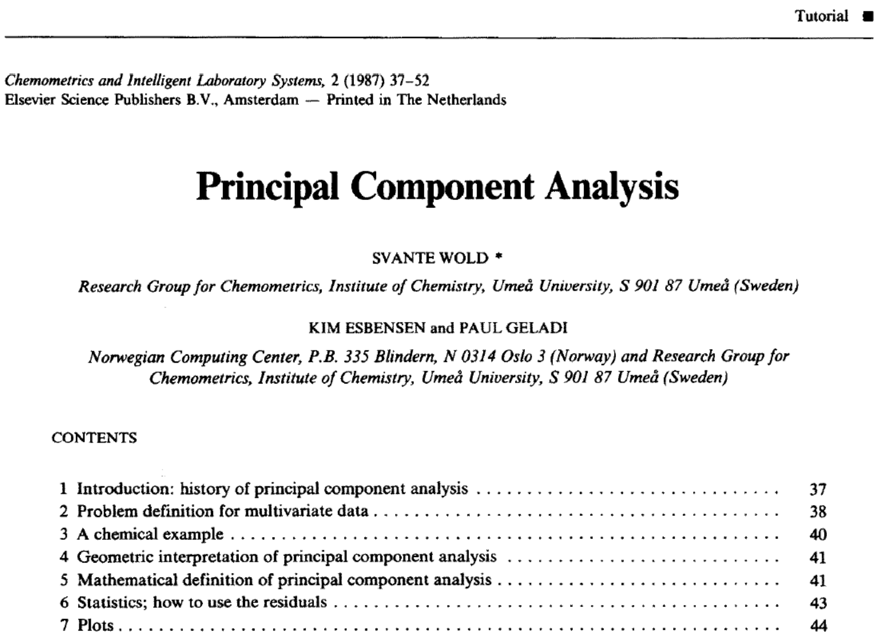


# Loading (*P*) plot

# In conclusion,

$$\boldsymbol{X = T.P + E}$$

- Scores (***T***) visualize the relationship between samples.
- Loading (***P***) can be used to investigate the behaviors of the studied parameters.
- In most cases, the first few of PCs can be used to contain most of the systematic variation.
- The variation that is not modeled is in residual (noise or non-systematic variation, ***E***).



Tutorial ■

*Chemometrics and Intelligent Laboratory Systems*, 2 (1987) 37–52
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

# Principal Component Analysis

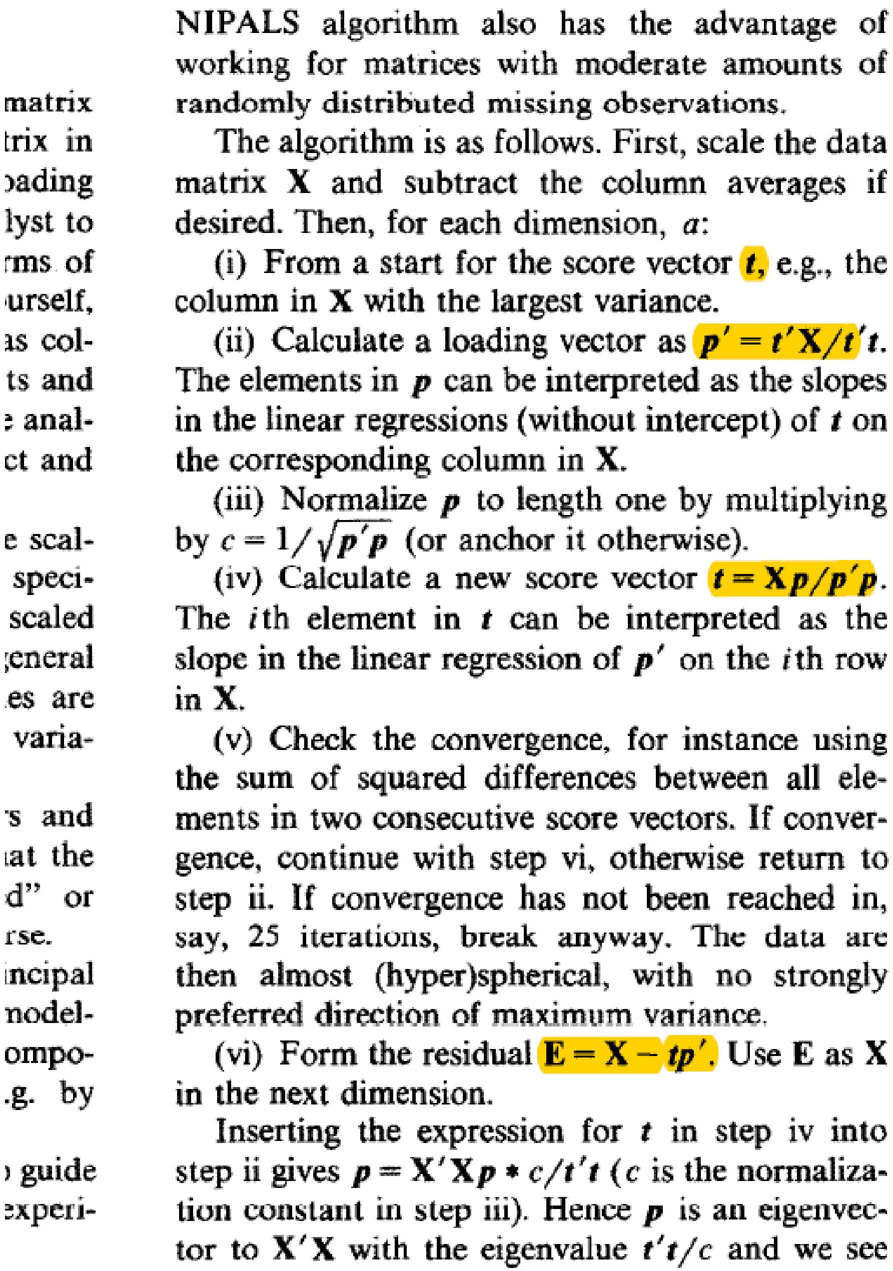
SVANTE WOLD *

*Research Group for Chemometrics, Institute of Chemistry, Umeå University, S 901 87 Umeå (Sweden)*

KIM ESBENSEN and PAUL GELADI

*Norwegian Computing Center, P.B. 335 Blindern, N 0314 Oslo 3 (Norway) and Research Group for Chemometrics, Institute of Chemistry, Umeå University, S 901 87 Umeå (Sweden)*

CONTENTS



matrix
trix in
ɔading
lyst to
rms of
urself,
ıs col-
ts and
: anal-
ct and

e scal-
speci-
scaled
;eneral
es are
varia-

's and
ıat the
d" or
rse.
ncipal
nodel-
ompo-
.g. by

) guide
:xperi-

NIPALS algorithm also has the advantage of working for matrices with moderate amounts of randomly distributed missing observations.

The algorithm is as follows. First, scale the data matrix $\mathbf{X}$ and subtract the column averages if desired. Then, for each dimension, $a$:

(i) From a start for the score vector $\boldsymbol{t}$, e.g., the column in $\mathbf{X}$ with the largest variance.

(ii) Calculate a loading vector as $\boldsymbol{p}' = \boldsymbol{t}'\mathbf{X}/\boldsymbol{t}'\boldsymbol{t}$. The elements in $\boldsymbol{p}$ can be interpreted as the slopes in the linear regressions (without intercept) of $\boldsymbol{t}$ on the corresponding column in $\mathbf{X}$.

(iii) Normalize $\boldsymbol{p}$ to length one by multiplying by $c = 1/\sqrt{\boldsymbol{p}'\boldsymbol{p}}$ (or anchor it otherwise).

(iv) Calculate a new score vector $\boldsymbol{t} = \mathbf{X}\boldsymbol{p}/\boldsymbol{p}'\boldsymbol{p}$. The $i$th element in $\boldsymbol{t}$ can be interpreted as the slope in the linear regression of $\boldsymbol{p}'$ on the $i$th row in $\mathbf{X}$.

(v) Check the convergence, for instance using the sum of squared differences between all elements in two consecutive score vectors. If convergence, continue with step vi, otherwise return to step ii. If convergence has not been reached in, say, 25 iterations, break anyway. The data are then almost (hyper)spherical, with no strongly preferred direction of maximum variance.

(vi) Form the residual $\mathbf{E} = \mathbf{X} - \boldsymbol{t}\boldsymbol{p}'$. Use $\mathbf{E}$ as $\mathbf{X}$ in the next dimension.

Inserting the expression for $\boldsymbol{t}$ in step iv into step ii gives $\boldsymbol{p} = \mathbf{X}'\mathbf{X}\boldsymbol{p} * c/\boldsymbol{t}'\boldsymbol{t}$ ($c$ is the normalization constant in step iii). Hence $\boldsymbol{p}$ is an eigenvector to $\mathbf{X}'\mathbf{X}$ with the eigenvalue $\boldsymbol{t}'\boldsymbol{t}/c$ and we see