Application of Computer in Chemistry SSC 3533

MULTIVARIATE CALIBRATION

Prof. Mohamed Noor Hasan
Dr. Hasmerya Maarof
Department of Chemistry

Multiple Linear Regression

Y is dependent on more than one *x*'s

Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots.\ \beta_m x_m$$

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots.\ b_m x_m$$

Example

Absorbance of a mixture of more than one components is determined by the concentration of each component in the sample:

$$A = b_1C_1 + b_2C_2 + b_3C_3$$

A – absorbance

b – molar absorbtivity

c – concentration

Calculation Method

- The same method used to get simple linear equation is also used here i.e. least squares method
- Find an equation that minimize the sum squared deviation.

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- To simplify the calculation, matrix operation is used to obtain b_0, b_1, b_2 ...b_m

Matrix Solution

The Y and X variables are expressed in matrix form

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1,p-1} \\ 1 & x & x_{22} & . & . & x_{2,p-1} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n,1} & x_{n,2} & . & . & x_{n,p-1} \end{bmatrix} \qquad b = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ b_{p-1} \end{bmatrix}$$

The regression equation

$$\hat{Y} = Xb$$

$$b = (X'X)^{-1}X'Y$$

Determination of a mixture

Example:

The absorbance of a mixture of Cl_2 and Br_2 in chloroform was measured at various wavelengths using a UV-Visible Spectrophotometer.

Based on the following data, determine the concentration of chlorine and bromine in the mixture.

Data

Wave Number	Molar absorbtivity (k)		Absorbance (A)
	Cl_2	Br_2	
22	4.5	168	34.10
24	8.4	211	42.95
26	20	158	33.55
28	56	30.0	11.70
30	100	4.70	11.00
32	71	5.30	7.98

Multivariate Calibration

- In Chemistry, we often encounter determination of more that one parameters at a time.
- Not only in *X* but sometimes in *Y* too!!

Determination of protein and water in wheat using near-IR

- Objective: To determine protein and water content in wheat using near-IR.

Calibration

- Get IR spectra for 3 samples of wheat
- Measure 'near-IR reflectance' at 8 wavelengths

➢ Data matrix: X (3 x 8)

- Determine, using normal method, percentage of protein (Kjeldahl method for N) and percentage of water in the three samples of wheat.

➢ Concentration matrix: C (3 x 2)

Data Matrix *

$$X_{3 \times 8} = \begin{bmatrix} 2 & 1 & 2 & 1 & 1 & 5 & 3 & 1 \\ 1 & 2 & 2 & 3 & 7 & 4 & 7 & 1 \\ 1 & 2 & 1 & 3 & 6 & 2 & 6 & 2 \end{bmatrix}$$

$$C_{3 \times 2} = \begin{bmatrix} 2 & 4 \\ 3 & 5 \\ 8 & 9 \end{bmatrix}$$

* Note, these are not real values

Multiple Linear Equation

$$\hat{\mathbf{C}} = \mathbf{X.B}$$

$$B = (X'.X)^{-1}.X'.C$$

Notice that C is now the Y variable and Absorbance is now X the variable – Inverse Calibration – why?

Prediction

$$\mathbf{C}_{\mathbf{unk}} = \mathbf{X}_{\mathbf{unk}}\mathbf{B}$$

- Make measurement on unknown smples
- Use the equation and stored value of **B** to Calculate **C** for the unknown samples

Summary

Calibration

- Using standards, measure concentration – **C**
- Obtain spectrum of standards – **X**
- Develop model: **C** = **XB** - store **B**

Prediction

- Obtain spectrum of unknown – $\mathbf{X}_{unk}$
- $\mathbf{C}_{unk} = \mathbf{X}_{unk}\,\mathbf{B}$

Problems with MLR

- Only suitable in ideal conditions:
- Strong correlation between X variables can cause problems.
- Number of samples must be greater than number of variables (Five to one rule)

Need a more robust method!!

Principal Component Regression

- Original variable, X, transformed into a new set of variables called principal components, PC
- Principal components, by definition, are not correlated with each other (orthogonal)
- Number of variables can be reduced – take only the important factors

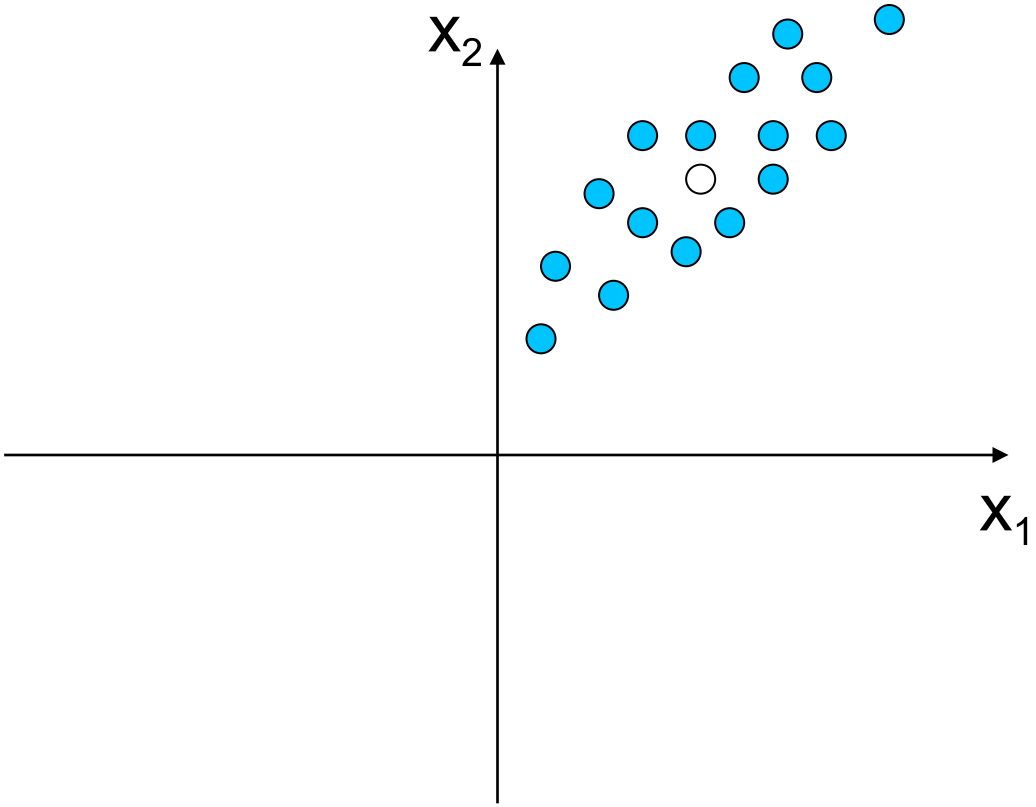
Principal Components Analysis

- Theory and original idea in mathematics and physics (1800s) – eigen analysis
- First applied in psychometrics (1930s, 1940s)
- Also called factor analysis
- Became popular in Chemistry 1970s

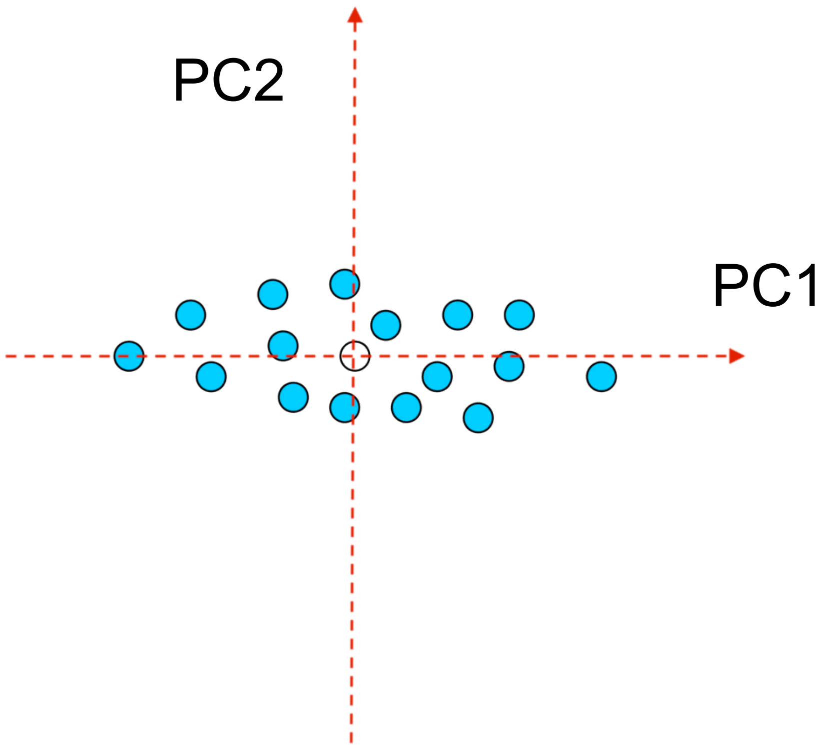
Concept of PCA

- Transformation from original coordinates into new coordinates
- The new coordinates are selected in such a way so that the highest variance is along PC1
- PC2 is at right angle to PC1

Original Data



Transformed data



PCA

$$X = T \cdot P + E$$

X – Original data (I x J)

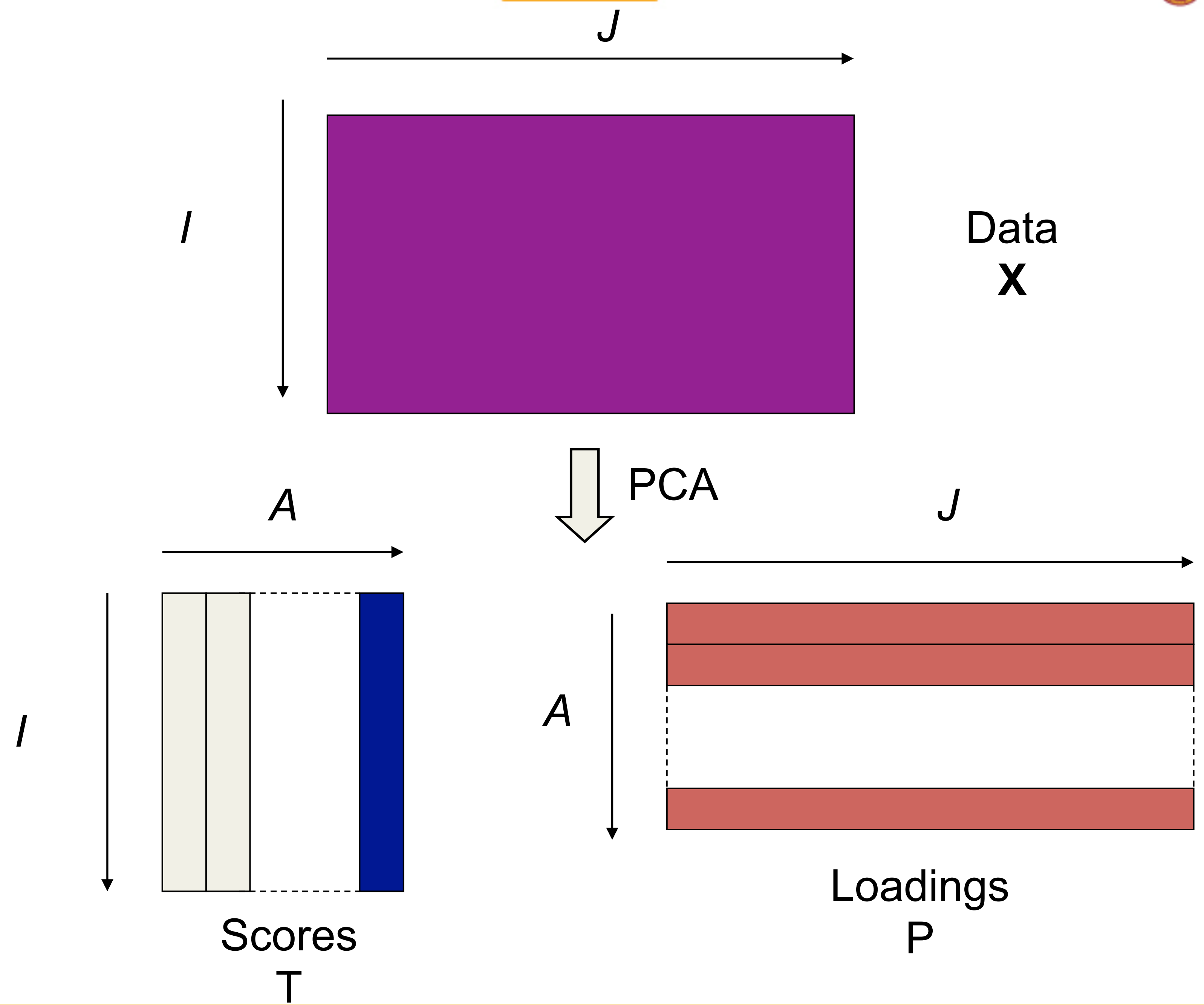
T – Scores (I x A)

P – Loading (A x J)

E – Error

Example

- Chromatogram obtained from HPLC using diode array detector
- Experimental Data:
  - Sampling 30 points (every 1 s)
  - At 28 wavelengths (interval 5 nm)
  - Absorbance Data (AU)
- Data matrix, X (30 x 28)


J
I
Data
X
PCA
A
I
Scores
T
J
A
Loadings
P

Summary of PCR

1. Get principal components using PCA

$$X = T \, . \, P + E$$

2. Regress C on T

$$C = T \, . \, R + E$$

3. Prediction:

$$C_{unk} = T_{unk} . \, R$$

Example of PCR

- Mixture of 10 polyaromatic hydrocarbon compounds (PAH)
  - Pyrene, Acenaphthene, Anthracene, Acenaphtylene…..
- 25 EAS Spectra at 27 wavelengths (from 220 – 350 nm at 5 nm intervals)

Partial Least Squares Regression, PLS

PCA is performed not only on X block but also on C block

$$X = T \,.\, P + E$$

$$C = U \,.\, Q + F$$

T – Score for X

U – Score for C

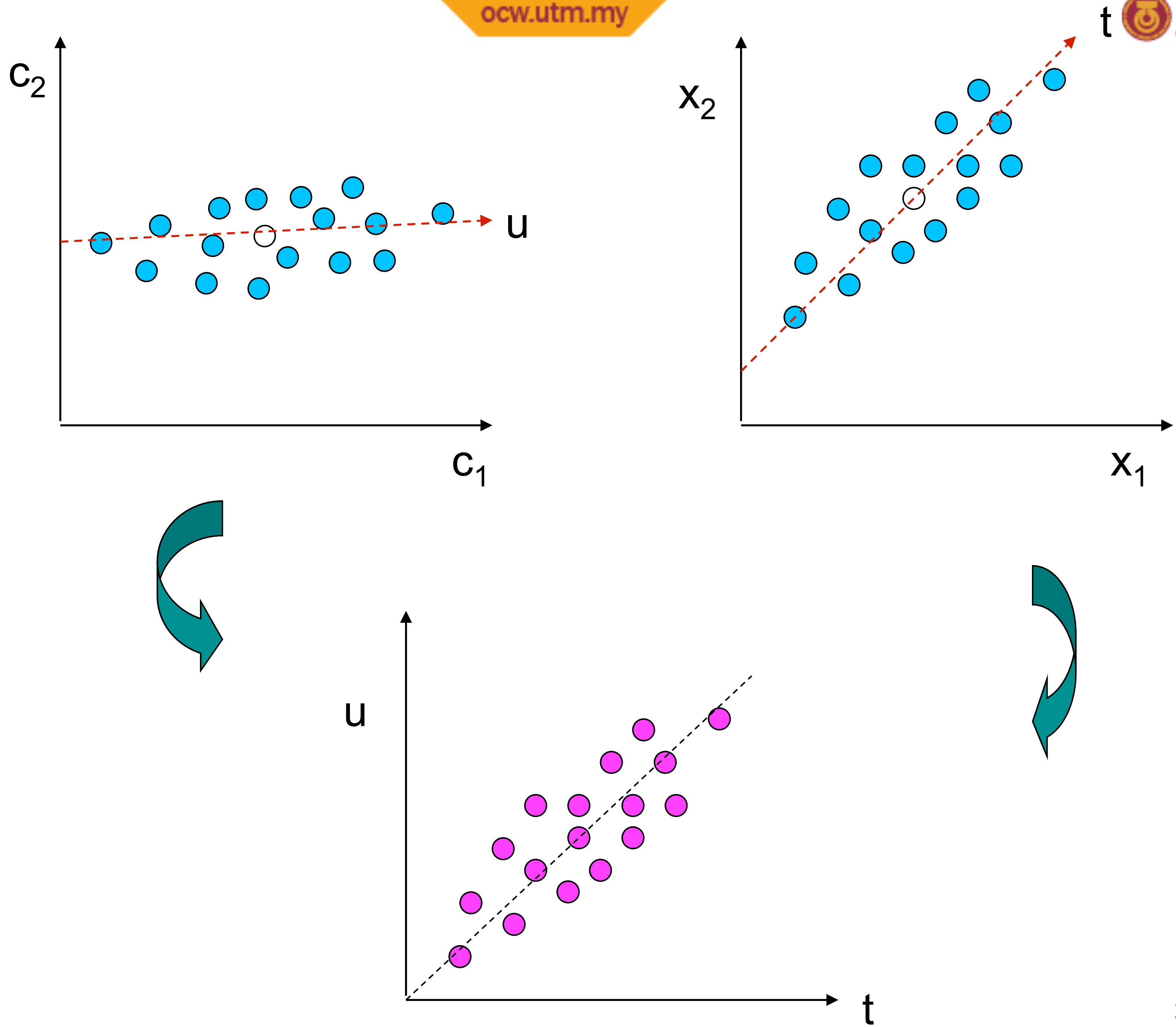
P – Loading for X

Q – Loading for C

PLS

- In ideal conditions, factors in X and factors in C should be the same.
- In reality, they are different. Get the relationship between U and T

  $u = bt + e$


c_2
u
c_1
x_2
t
x_1
u
t