

A Guide to the Practical Use of Chemometrics - with applications for Static SIMS

Joanna Lee, Ian Gilmore

National Physical Laboratory, Teddington, UK

Email: joanna.lee@npl.co.uk

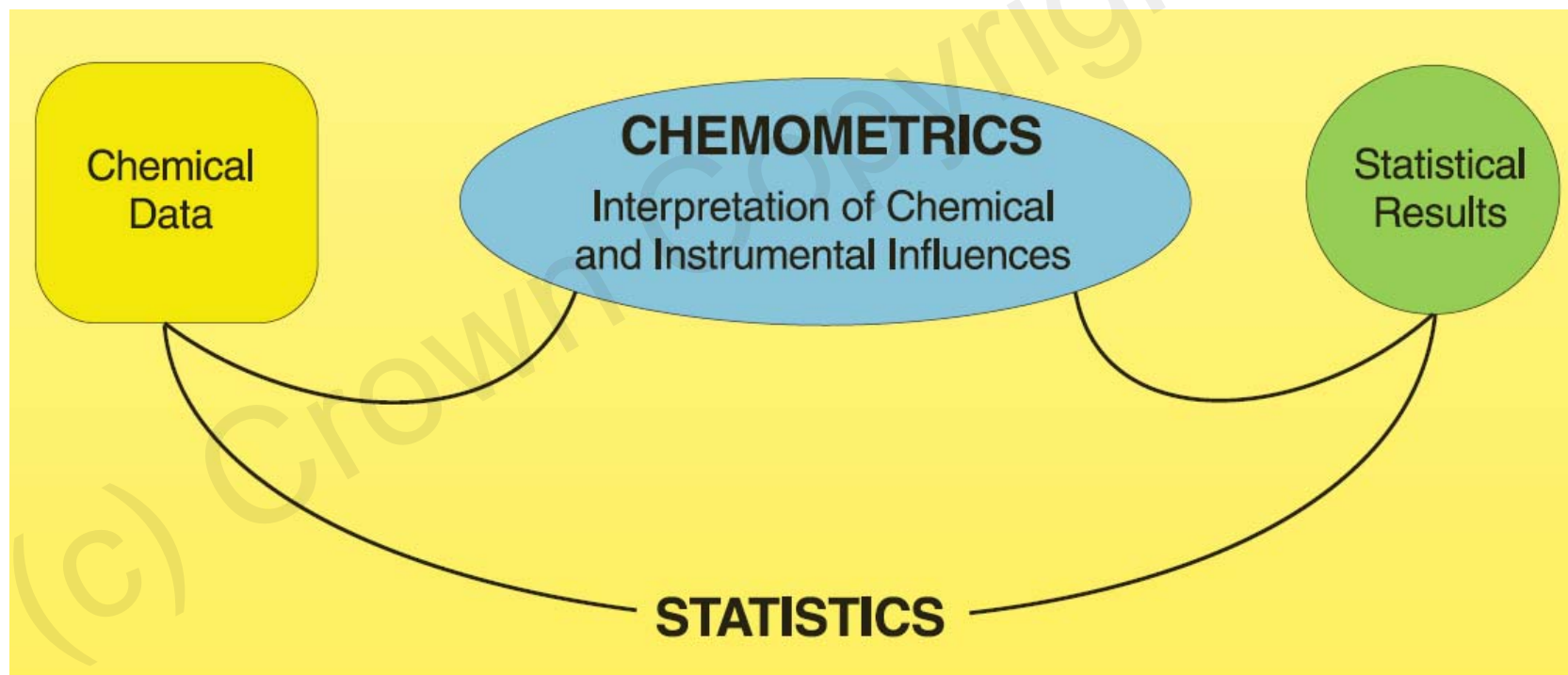
Web: <http://www.npl.co.uk/nanoanalysis>

Department for
**Innovation,
Universities &
Skills**

1. Introduction
2. Linear algebra
3. Factor analysis
 - Principal component analysis
 - Multivariate curve resolution
4. Multivariate regression
 - Multiple linear regression
 - Principal component regression
 - Partial least squares regression
5. Classification
 - Principal component discriminant function analysis
 - Partial least squares discriminant analysis
6. Conclusion

Chemometrics

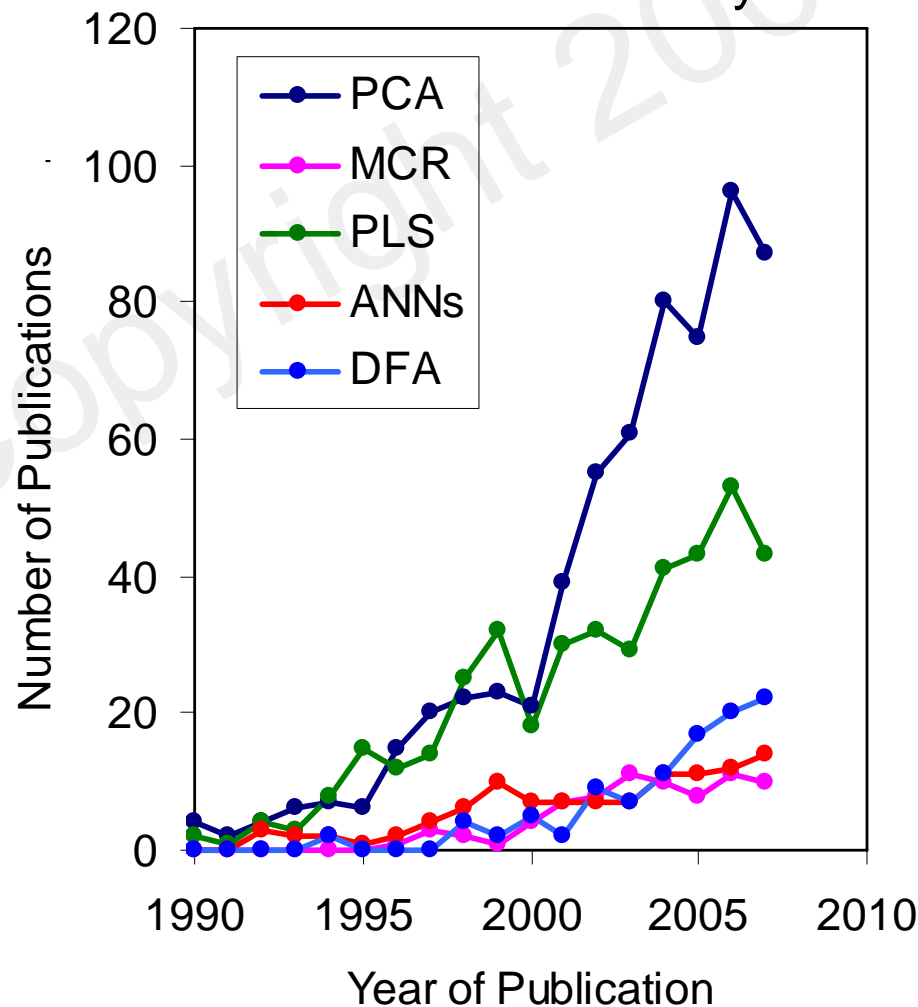
Chemometrics is the science of relating measurements made on a chemical system to the state of the system via application of **mathematical** or **statistical** methods



Multivariate analysis

- Analysis involving a simultaneous **statistical procedure** for two or more **dependent** variables, e.g. mass (SIMS) or binding energy (XPS)
- **Summarises** the data with a large number of dependent variables using a **smaller** number of statistical variables

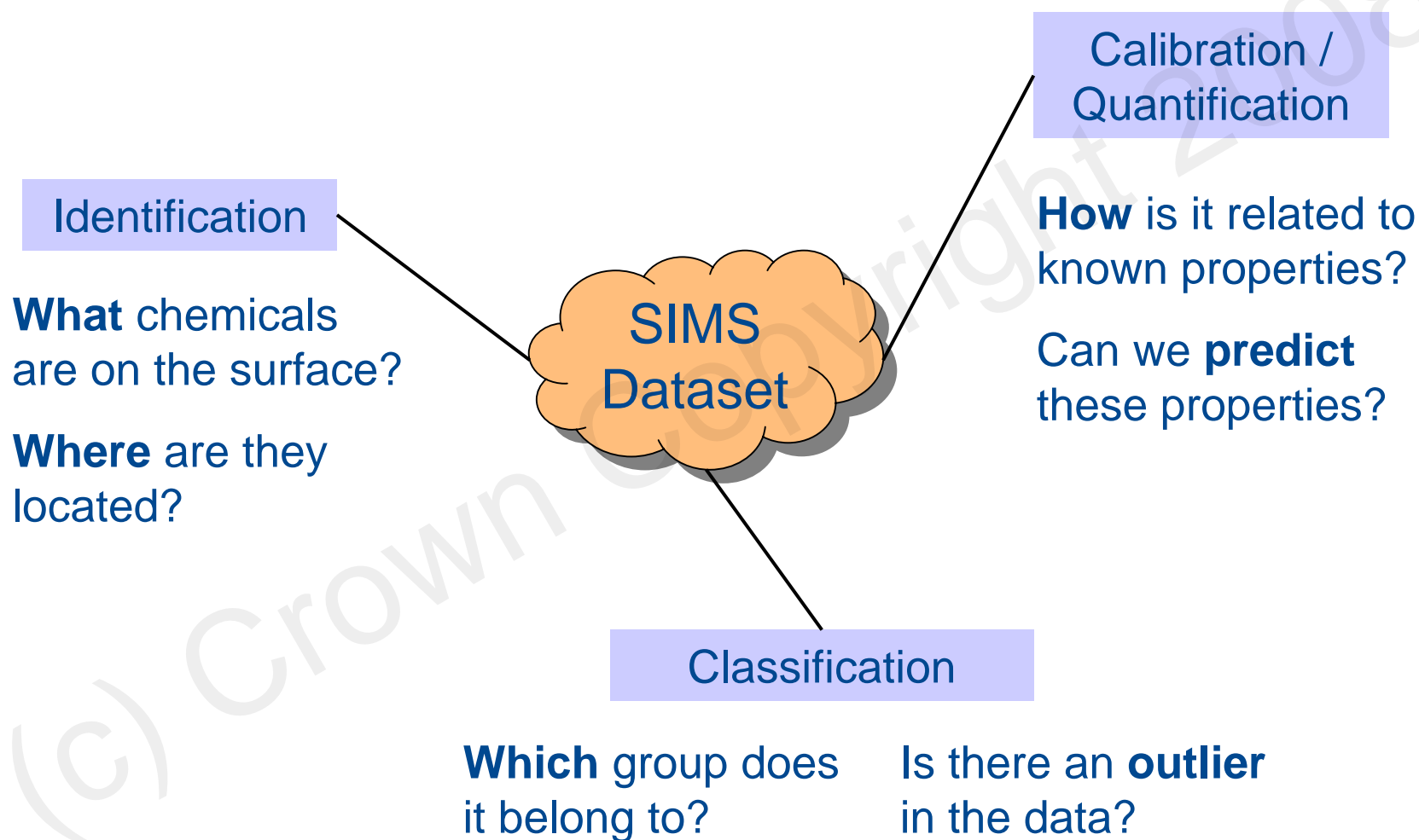
Multivariate analysis publications in surface chemical analysis



Multivariate analysis

- Advantages
 - Fast and efficient on modern computers
 - Statistically valid
 - Uses all information available
 - Removes potential bias
- Disadvantages
 - Lots of different methods, procedures, terminologies
 - Can be difficult to understand!





1. Introduction
2. Linear algebra
 - **Vector algebra**
 - **Matrix algebra**
 - **Rank and projections**
 - **Data matrix**
3. Factor analysis
4. Multivariate regression
5. Classification
6. Conclusion

Data matrix

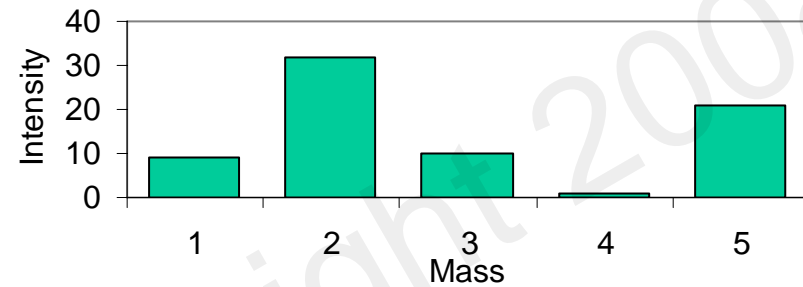
$X =$	9	32	10	1	21
	18	20	22	4	12
	24	12	30	6	6

Variables

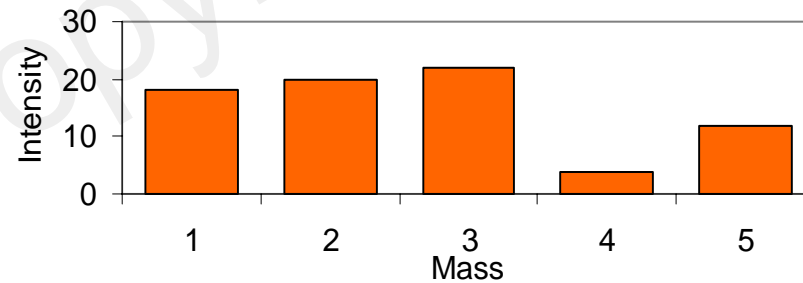
Samples

X has 3 row and 5 columns \rightarrow
 3×5 data matrix

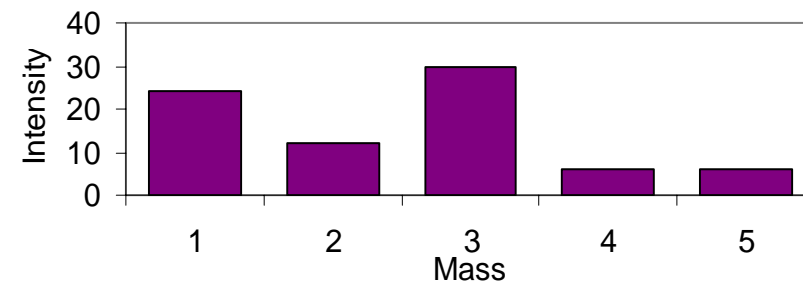
Mass spectrum of Sample 1



Mass spectrum of Sample 2



Mass spectrum of Sample 3



Vector inner product

$$\begin{aligned} a_x &= 1 \\ a_y &= 2 \\ a_z &= 4 \end{aligned} \quad \mathbf{a} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 4 \\ 2 \\ 2 \end{bmatrix}$$

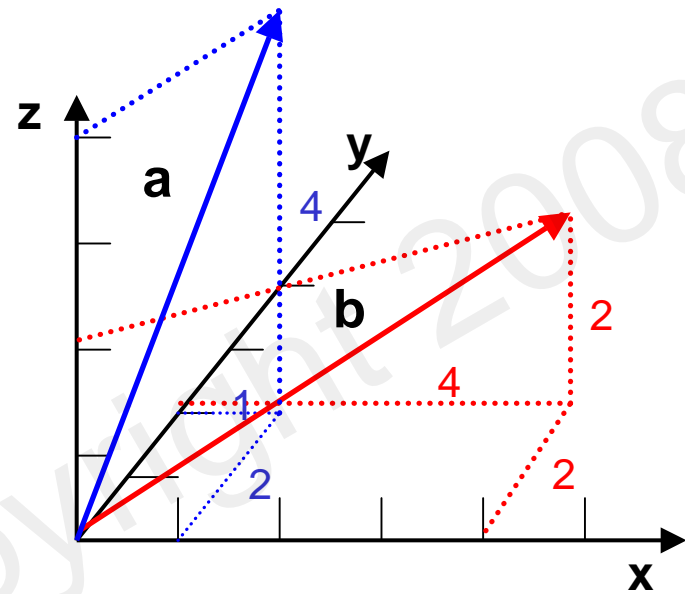
Vector Inner Product ('dot product')

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta$$

$$\begin{aligned} \mathbf{a} \cdot \mathbf{b} &= a_x b_x + a_y b_y + a_z b_z \\ &= 1 \times 4 + 2 \times 2 + 4 \times 2 \\ &= 16 \end{aligned}$$

$$\theta = 44.5^\circ$$

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}' \mathbf{b} = \begin{bmatrix} 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \\ 2 \end{bmatrix}$$



Transpose (to exchange rows and columns)

$$\mathbf{a}' = \begin{bmatrix} 1 & 2 & 4 \end{bmatrix}$$

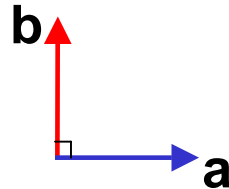
Vector length

$$|\mathbf{a}| = \sqrt{1^2 + 2^2 + 4^2}$$

Vector correlations

$$\mathbf{a}'\mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta$$

Orthogonality



If $\mathbf{a}'\mathbf{b} = 0$ then they are **orthogonal** i.e. at right angles $\theta = 90^\circ$

If they are also of unit length then they are **orthonormal** i.e. $|\mathbf{a}| = 1$ $|\mathbf{b}| = 1$

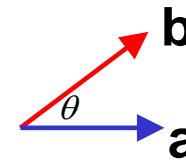
Orthogonal vectors are **uncorrelated**

Collinearity



If $\theta = 0^\circ$ then the vectors are **collinear**

Correlation



If $\theta \neq 0^\circ \neq 90^\circ$ then the vectors are neither orthogonal nor collinear – they are **correlated**

The smaller θ is the larger the correlation between **a** and **b**

Matrix addition

$$\mathbf{A} + \mathbf{B} = \mathbf{C}$$

$$(I \times K) + (I \times K) = (I \times K)$$

- **A** and **B** must be the same size
- Each corresponding element is added

$$\begin{bmatrix} 2 & 4 & 1 \\ 3 & 8 & 6 \end{bmatrix} + \begin{bmatrix} -1 & 2 & 0 \\ 0 & 1 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 6 & 1 \\ 3 & 9 & 4 \end{bmatrix}$$

(e.g. pure spectra + noise = experimental data)

Matrix multiplication

$$\mathbf{AB} = \mathbf{C}$$

$$(I \times N)(N \times K) = (I \times K)$$

- No. of columns of **A** must be equal no. of rows of **B**
- Row *i* of **A** times column *j* of **B** gives the row *i* and column *j* of the product matrix **AB**

$$\begin{bmatrix} 1 & 4 \\ 2 & 2 \\ 4 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \times 1 + 4 \times 3 & 1 \times 2 + 4 \times 2 \\ 2 \times 1 + 2 \times 3 & 2 \times 2 + 2 \times 2 \\ 4 \times 1 + 2 \times 3 & 4 \times 2 + 2 \times 2 \end{bmatrix} = \begin{bmatrix} 13 & 10 \\ 8 & 8 \\ 10 & 12 \end{bmatrix}$$

Matrix inverse

Identity matrix:
diagonal of 1s

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{AI} = \mathbf{A}$$

Matrix inverse
for a square matrix

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

(only exists if matrix
is 'full rank')

Matrix pseudoinverse
for a rectangular matrix

$$\mathbf{A}^{+} = \mathbf{A}'[\mathbf{AA}']^{-1}$$

$$\mathbf{A}^{+}\mathbf{A} = \mathbf{I}$$

We can now solve matrix equation

$$\mathbf{AB} = \mathbf{C}$$

If **A** is square

$$\mathbf{B} = \mathbf{A}^{-1}\mathbf{C}$$

If **A** is rectangular

$$\mathbf{B} = \mathbf{A}^{+}\mathbf{C}$$

Rank and singularity

Simultaneous equations of any size can be solved by matrices

Rank = number of unique equations. This matrix is rank 2

$$\begin{aligned} 1x + 2y &= 5 \\ 3x + 2y &= 7 \end{aligned} \Rightarrow \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

The matrix inverse. If it cannot be inverted the matrix is **singular**

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 7 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Row 3 is simply multiple of row 1 so rank = 2

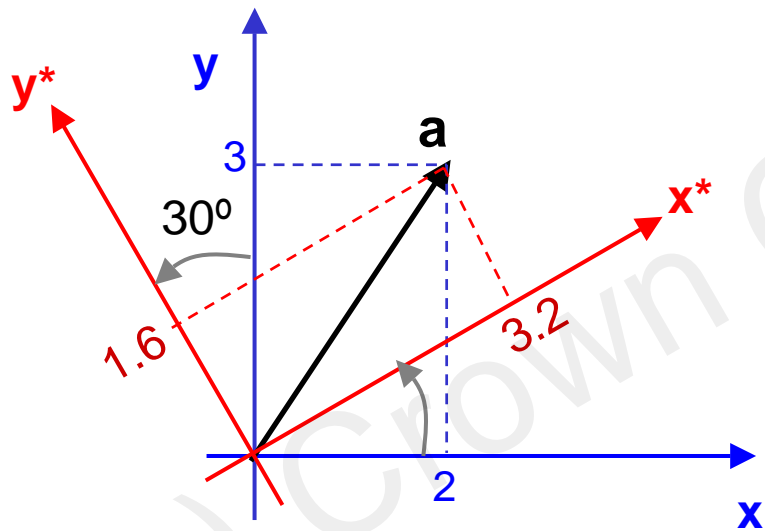
$$\begin{bmatrix} 1 & 2 \\ 3 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \\ 10 \end{bmatrix}$$

Rank is the number of rows or columns that are linearly independent
To obtain unique solution we require number of variables \leq rank

Matrix projections

To write \mathbf{a} in terms of \mathbf{x}^* and \mathbf{y}^* , we find its **projections** on the new axes

$$a = 2x + 3y \quad \mathbf{a} = \begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$



$$\mathbf{a} = \begin{bmatrix} 3.2 & 1.6 \end{bmatrix} \begin{bmatrix} x^* \\ y^* \end{bmatrix}$$

projections of \mathbf{a}
onto new axes

new axes

The new axes \mathbf{x}^* and \mathbf{y}^* can be written in terms of \mathbf{x} and \mathbf{y}

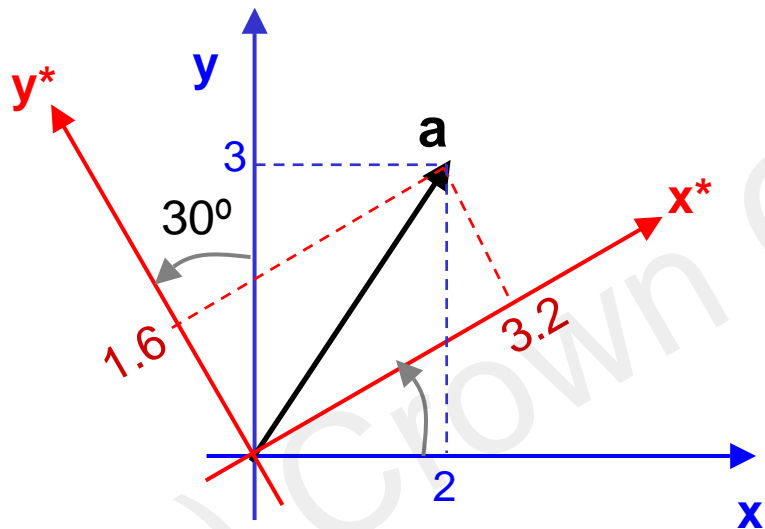
$$\begin{bmatrix} x^* \\ y^* \end{bmatrix} = \begin{bmatrix} 0.87 & 0.5 \\ -0.5 & 0.87 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

projections of new axes
onto old axes

old axes

Matrix projections

$$a = 2x + 3y \quad \mathbf{a} = \begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$



Putting these together,
we can write vector \mathbf{a} as

$$\mathbf{a} = \begin{bmatrix} 3.2 & 1.6 \end{bmatrix} \begin{bmatrix} 0.87 & 0.5 \\ -0.5 & 0.87 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

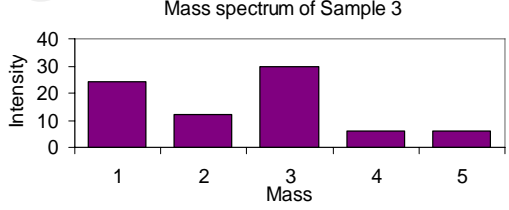
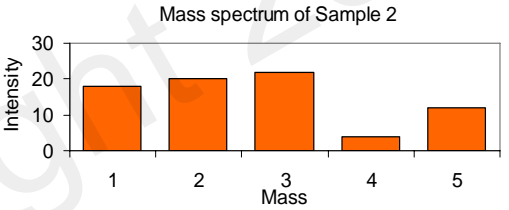
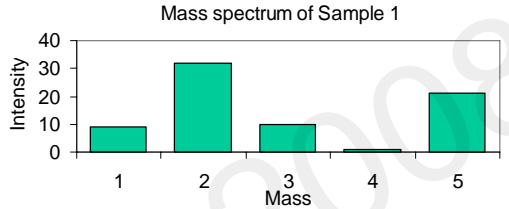
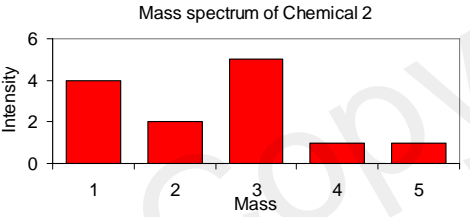
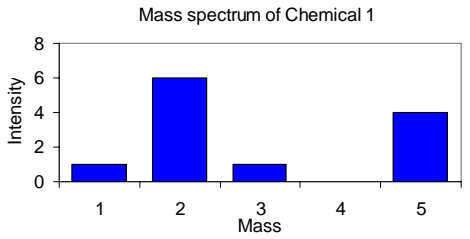
projections of \mathbf{a}
onto new axes

projections of
new axes
onto old axes

old axes

Data matrix

	Chemical 1	Chemical 2
Sample 1	5	1
Sample 2	2	4
Sample 3	0	6



×

=

Sample composition

Chemical spectra

Data matrix

Samples

$$\begin{bmatrix} 5 & 1 \\ 2 & 4 \\ 0 & 6 \end{bmatrix}$$

Chemicals

×

$$\begin{bmatrix} 1 & 6 & 1 & 0 & 4 \\ 4 & 2 & 5 & 1 & 1 \end{bmatrix}$$

Chemicals

Variables [mass]

=

$$\begin{bmatrix} 9 & 32 & 10 & 1 & 21 \\ 18 & 20 & 22 & 4 & 12 \\ 24 & 12 & 30 & 6 & 6 \end{bmatrix}$$

Samples

Variables [mass]

Data matrix

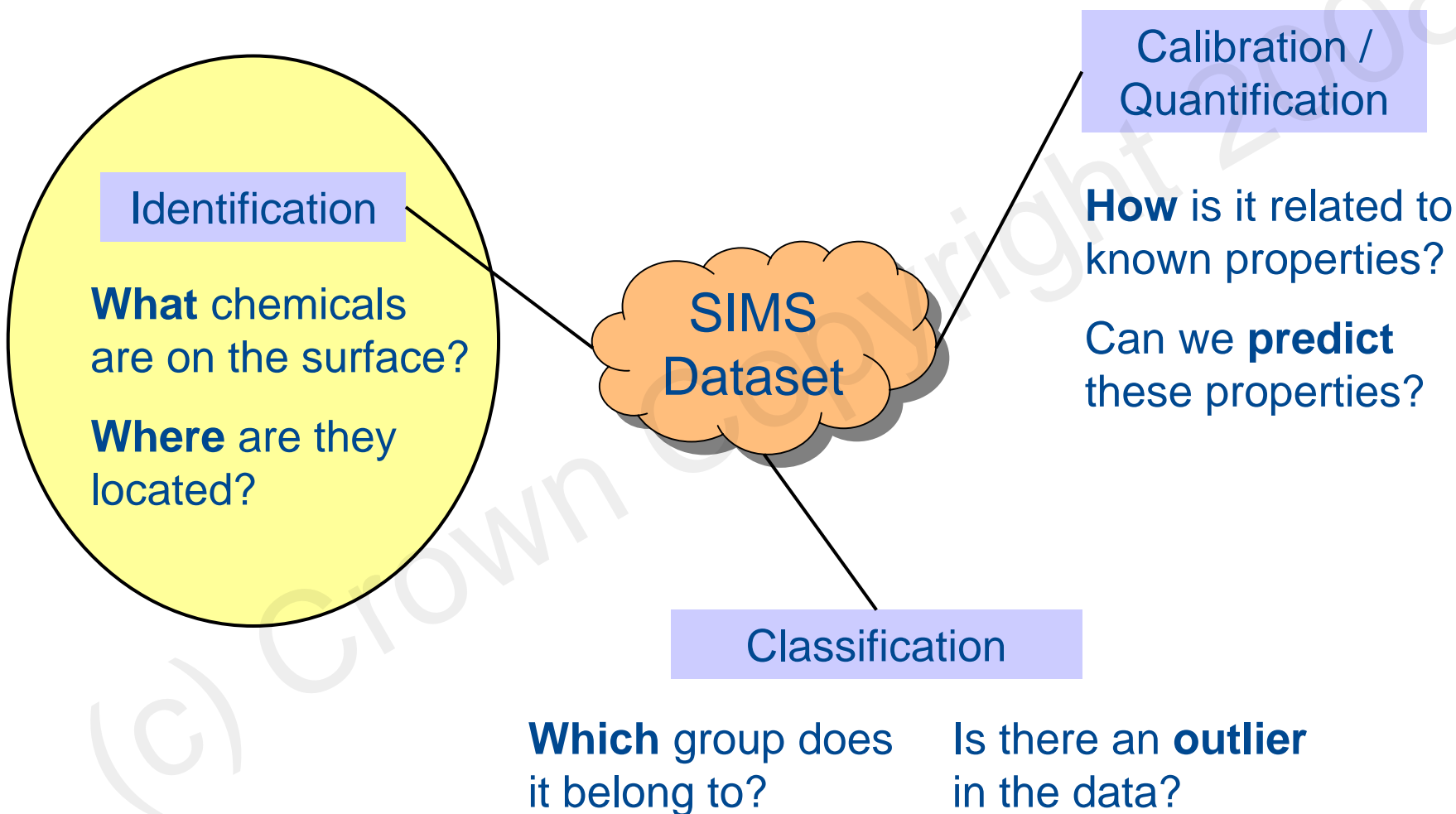
1. Each spectrum can be represented by a **vector**
2. Instead of x, y, z in 3D real space, the axes are *mass1, mass2, mass3...* etc in **variable space** (also '**data space**')
3. Without noise, rank of dataset = number of unique components
4. With random, uncorrelated noise, rank of dataset = number of samples or number of variables, whichever is smaller

$$\begin{array}{c} \text{Samples} \\ \left[\begin{array}{cc} 5 & 1 \\ 2 & 4 \\ 0 & 6 \end{array} \right] \\ \text{Chemicals} \end{array} \times \begin{array}{c} \text{Chemical spectra} \\ \left[\begin{array}{ccccc} 1 & 6 & 1 & 0 & 4 \\ 4 & 2 & 5 & 1 & 1 \end{array} \right] \\ \text{Variables [mass]} \\ \text{Chemicals} \end{array} = \begin{array}{c} \text{Data matrix} \\ \left[\begin{array}{ccccc} 9 & 32 & 10 & 1 & 21 \\ 18 & 20 & 22 & 4 & 12 \\ 24 & 12 & 30 & 6 & 6 \end{array} \right] \\ \text{Variables [mass]} \\ \text{Samples} \end{array}$$

Vector and matrix summary

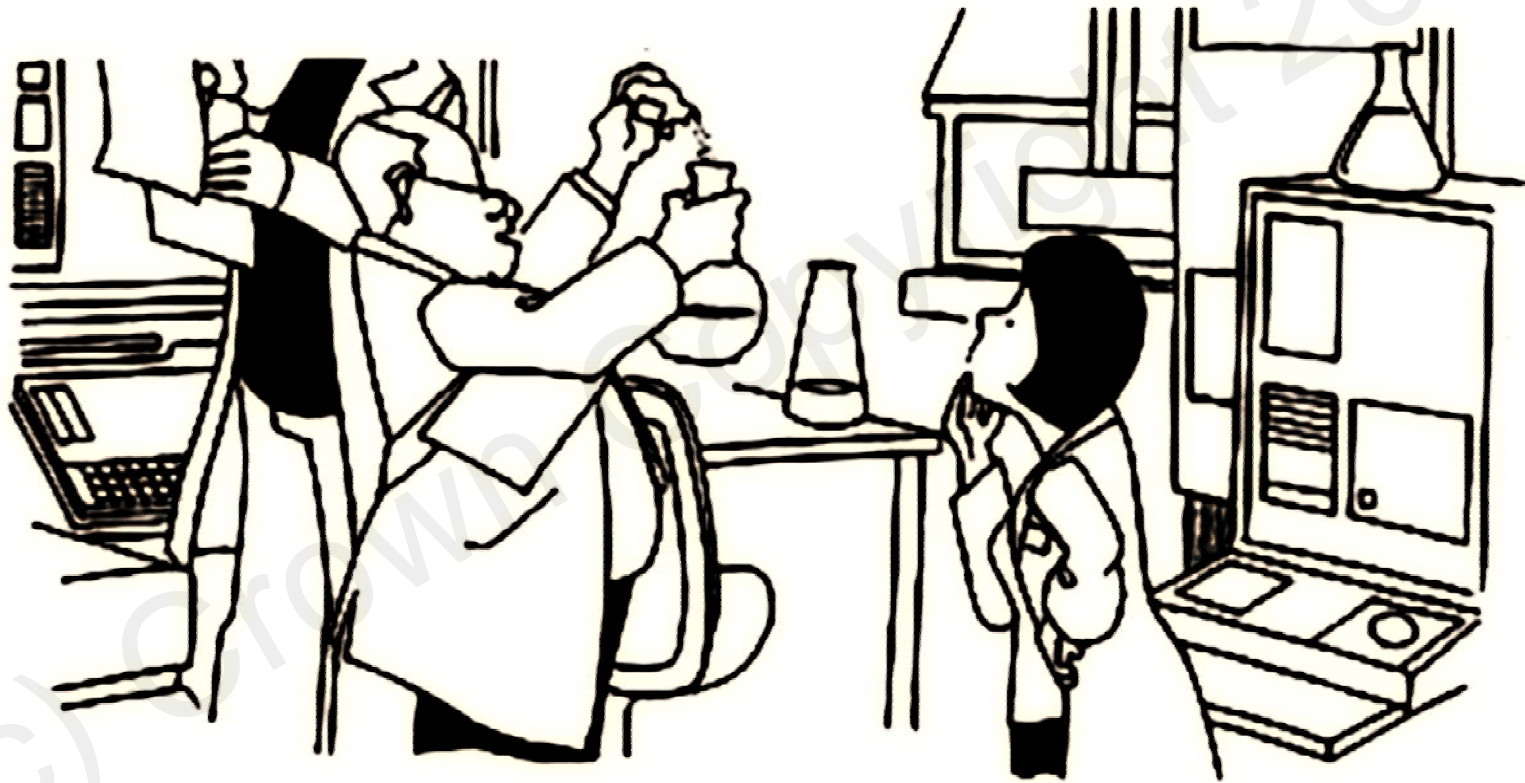
- Each row of the **data matrix** contains a spectrum that can be represented by a **vector** in K dimensional **data space** (K = no. of mass bins)
- Vectors can be **orthogonal** (90°), **collinear** (0°) or **correlated**
- Vectors can be described using a set of **rotated axes** by finding their **projections** onto the new axes
- The **rank** of a data set represents the number of **independent parameters** that are needed to fully describe the data

1. Introduction
2. Linear algebra
3. Factor analysis
 - **Principal component analysis (PCA)**
 - **Data preprocessing**
 - **PCA Examples**
 - Multivariate curve resolution (MCR)
 - MCR Examples
4. Multivariate regression
5. Classification
6. Conclusion



Terminology

A well-defined terminology is essential for ideas and practices to be communicated clearly and accurately



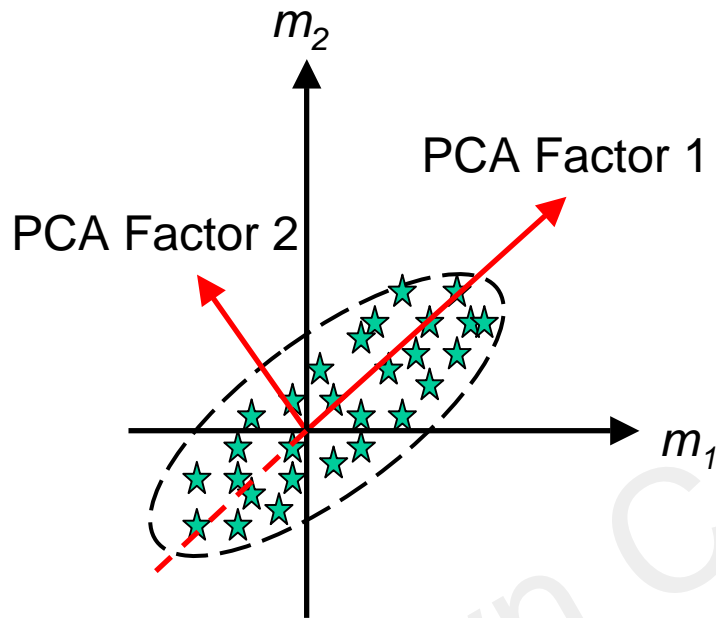
"... then we add a smidgin of this - that's less than a dollop, but more than a pinch..."

Terminology

In order to clarify existing terminology and emphasise the relationship between the different multivariate techniques, we are going to adopt the following terminology in this lecture

Terms Here	Symbol	Definition	PCA	MCR	PLS
Factor	-	An axis in the data space representing an underlying dimension that contributes to summarising or accounting for the original data set	Principal Component	Pure Component	Latent Vectors, Latent Variables
Loadings	P	Correlation between the original variables and the factors	Loadings, Eigenvector	Component Spectrum	Loadings
Scores	T	Projection of the samples onto the factors	Scores, Projections	Component Concentration	Scores

Principal component analysis (PCA)



PCA is a technique for reducing matrices of data to their **lowest dimensionality** by describing them using a small number of **factors**

- **Factors** are directions in the data space that contribute to summarising or accounting for the original data set
- Equivalent to a **rotation** in data space – **factors are new axes**
- Data described by their **projections** onto the factors
- **Factor analysis techniques** differ in the way the factors are extracted

Principal component analysis (PCA)

I = no. of samples
 K = no. of mass units
 N = no. of factors

PCA follows the **factor analysis** equation –

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}$$

$(I \times K) = (I \times N)(N \times K) + (I \times K)$

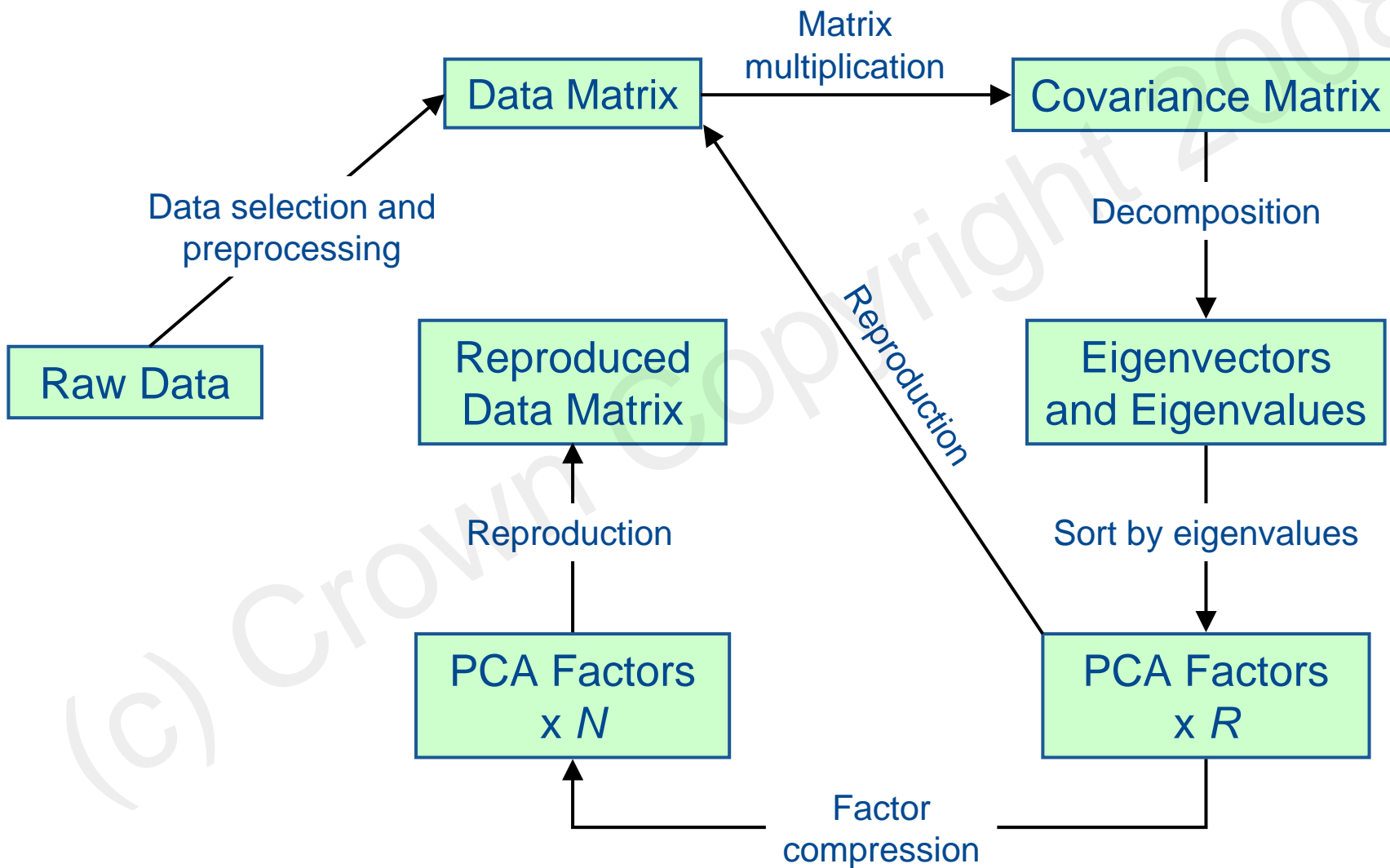
Data matrix \rightarrow \mathbf{X}
 Projection of samples onto factors (scores matrix) \rightarrow \mathbf{T}
 Projection of variables onto factors (loadings matrix) \rightarrow \mathbf{P}'
 Residuals (noise) \rightarrow \mathbf{E}

We describe the data \mathbf{X} (rank R) using N rotated axes (factors), where $N < R$. Each factor consists of two vectors, \mathbf{t}_n (scores vector), and \mathbf{p}_n (loadings vector)

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \sum_{n=1}^N \mathbf{t}_n \mathbf{p}'_n + \mathbf{E}$$

$(I \times 1)$ $(1 \times I)$

PCA outline



I = no. of samples
 K = no. of mass units
 N = no. of factors

Covariance matrix contains information about the variances of data points within the dataset, and is defined as

$$\mathbf{Z} = \mathbf{X}'\mathbf{X}$$

$(K \times K) = (K \times I)(I \times K)$

In PCA, \mathbf{Z} is *decomposed* into a set of **eigenvectors** \mathbf{p} and associated **eigenvalues** λ , such that

$$\mathbf{Z}\mathbf{p} = \lambda\mathbf{p}$$

$(K \times K)(K \times 1) = (K \times 1)$

Eigenvalues and eigenvectors have some special properties:

- Eigenvalues are **positive** or zero
- The number of **non-zero eigenvalues** = rank of data R
- Eigenvectors are **orthonormal**

PCA factors

- Because \mathbf{Z} is the covariance matrix, eigenvectors of \mathbf{Z} are special directions in the data space that is **optimal** in describing the **variance** of the data
- Eigenvalues are the **amount of variance** described by their associated eigenvector

$$\mathbf{X} = \mathbf{TP}' = \sum_{n=1}^R \mathbf{t}_n \mathbf{p}'_n$$

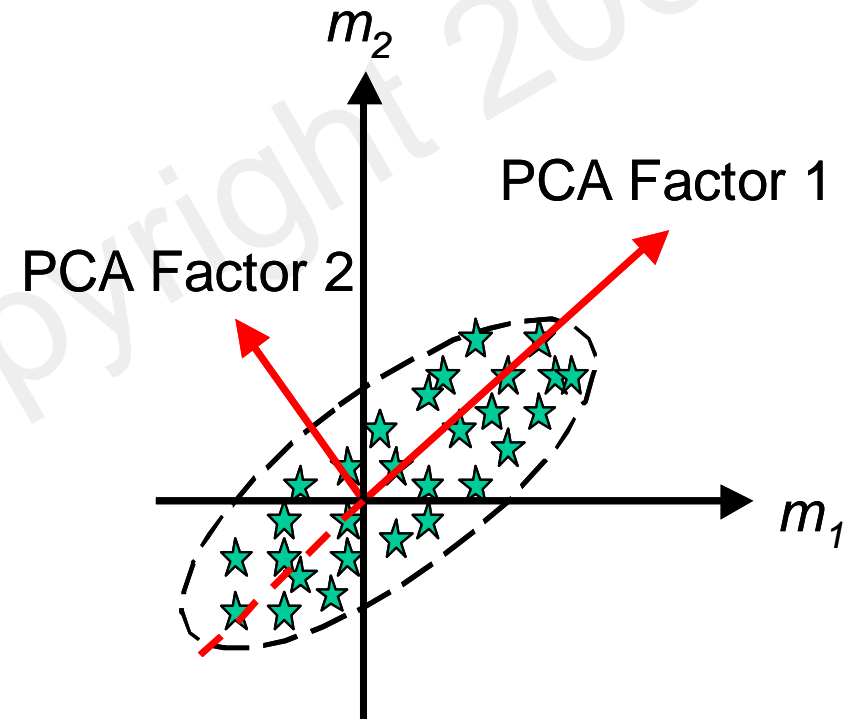
Projection of samples onto n^{th} factor (scores)

Projection of variables onto n^{th} factor (loadings)

- These eigenvectors are the **factors** PCA obtain for the factor analysis equation. They are sorted by their eigenvalues
- PCA factors **successively** capture the largest amount of variance (spread) within the dataset
- Projection of samples onto factors (scores) are **orthogonal**

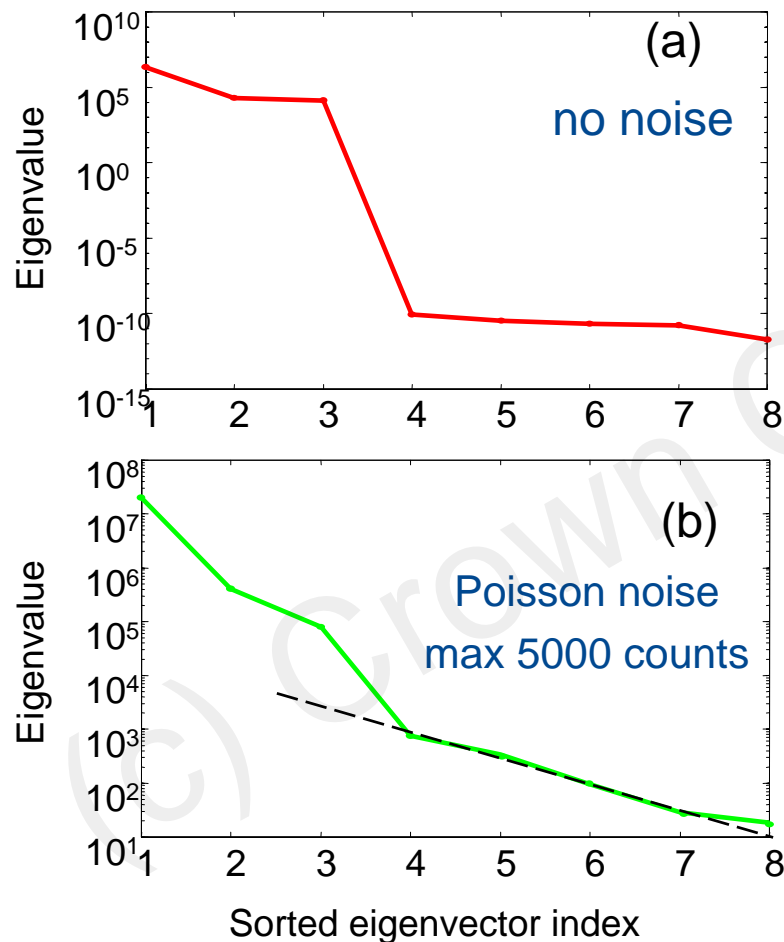
PCA – graphical representation

- The first factor lies along the major axis of ellipse and accounts for most variation
- Instead of describing the data using correlated variables m_1 and m_2 , we transform them onto a new basis (factors) which are uncorrelated
- By removing higher factors (variances due to noise) we can reduce dimensionality of data \Rightarrow 'factor compression'



Number of factors

Data set of 8 spectra from mixing 3 pure compound spectra



1. Prior knowledge of system
2. 'Scree test':
Eigenvalue plot levels off in a linearly decreasing manner after 3 factors
3. Percentage of variance captured by N^{th} eigenvector:

$$\frac{N^{\text{th}} \text{ eigenvalue}}{\text{sum of all eigenvalues}} \times 100\%$$

4. Percentage of total variance captured by N eigenvectors:

$$\frac{\text{sum of eigenvalues up to } N}{\text{sum of all eigenvalues}} \times 100\%$$

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \sum_{n=1}^N \mathbf{t}_n \mathbf{p}'_n + \mathbf{E}$$

$$\bar{\mathbf{X}} = \mathbf{X} - \mathbf{E} = \mathbf{TP}'$$

\mathbf{E} is the matrix of **residuals**

- should contain noise only
- useful for judging quality of PCA model
- may show up unexpected features!

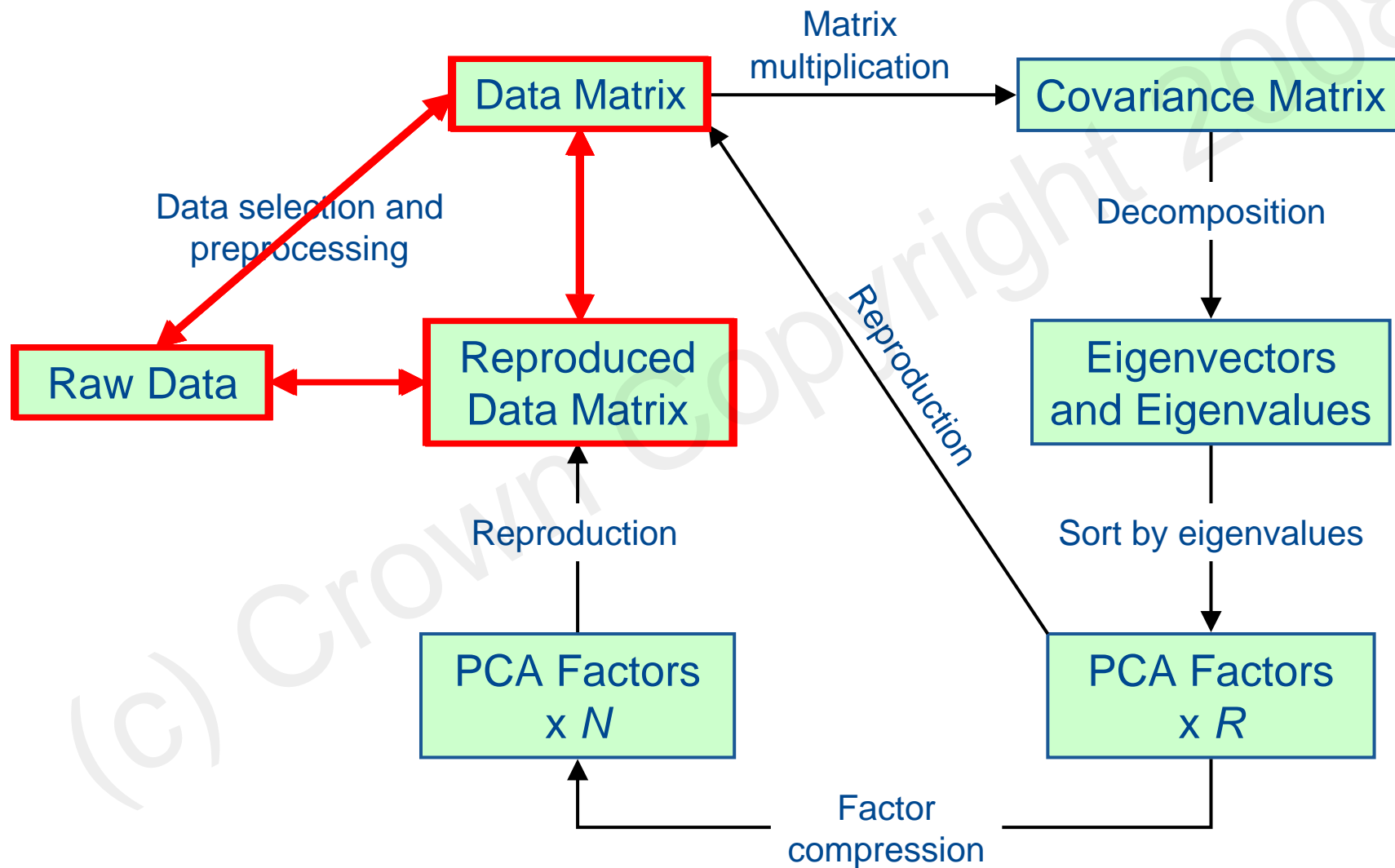
$\bar{\mathbf{X}}$ is the **reproduced data matrix**

- reproduced from N selected factors
- noise filtered by removal of higher factors that describe noise variations
- useful for MCR

$$\mathbf{E} = \mathbf{X} - \bar{\mathbf{X}}$$

$$\mathbf{E} = \mathbf{X} - \sum_{n=1}^N \mathbf{t}_n \mathbf{p}'_n$$

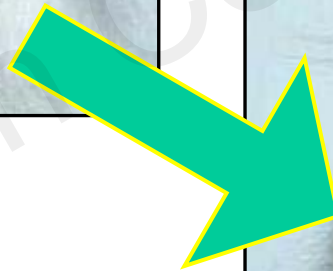
PCA outline



Data preprocessing



Data preprocessing is the manipulation of data prior to data analysis...



Data preprocessing

- Enhances PCA by bringing out important variance in dataset
- Makes assumption about nature of variance in data
- Can distort interpretation and quantification
- Includes:
 - mass binning
 - peak selection
 - mean centering
 - normalisation
 - variance scaling
 - Poisson scaling
 - Binomial scaling
 - Logarithmic transformation

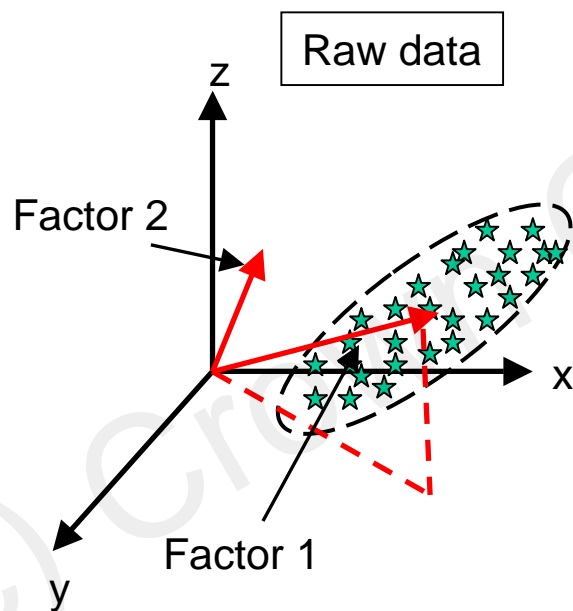
More details in the following slides

Mean centering

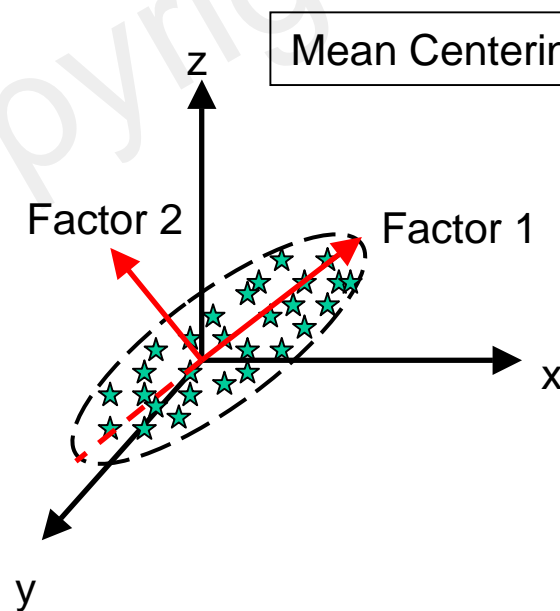
$$\tilde{\mathbf{x}}_{i,k} = \mathbf{x}_{i,k} - \text{mean}(\mathbf{x}_{:,k})$$

Preprocessed data sample i , mass k Raw data sample i , mass k Mean intensity of mass k

- Subtract mean spectrum from each sample
- PCA describes variations from the mean



1st factor goes from origin to centre of gravity of data

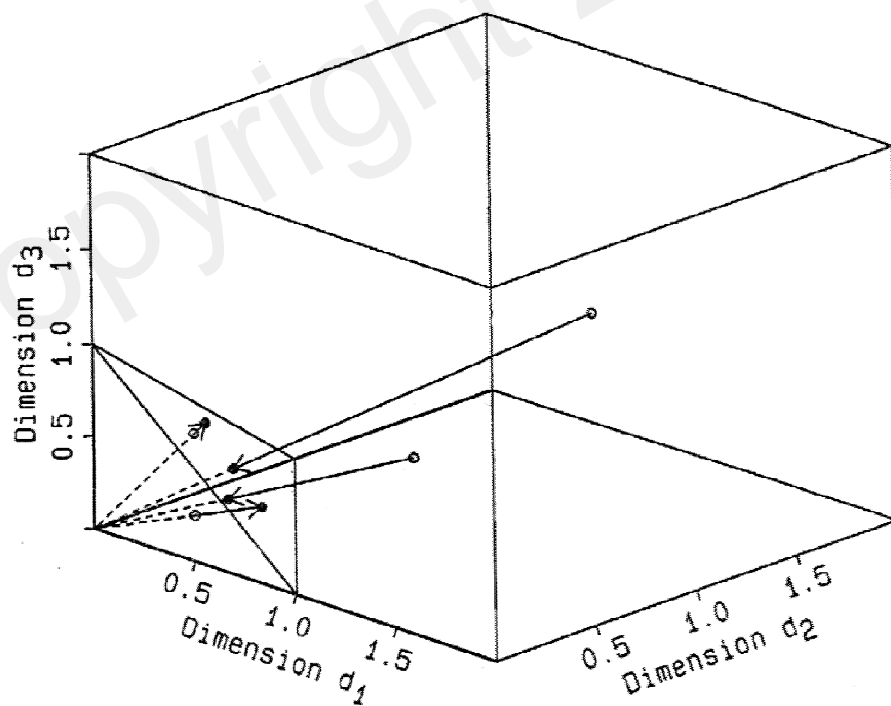


1st factor goes from origin and accounts for the highest variance

$$\tilde{\mathbf{X}}_{i,k} = \mathbf{X}_{i,k} \times \frac{1}{\text{sum}(\mathbf{X}_{i,:})}$$

Preprocessed data sample i , mass k Raw data sample i , mass k Total intensity of sample i

- Divide each spectrum by its total ion intensity
- Reduces effects of topography, sample charging, drift in primary ion current
- Assumes chemical variances can be described by relative changes in ion intensities
- Reduces rank of data by 1



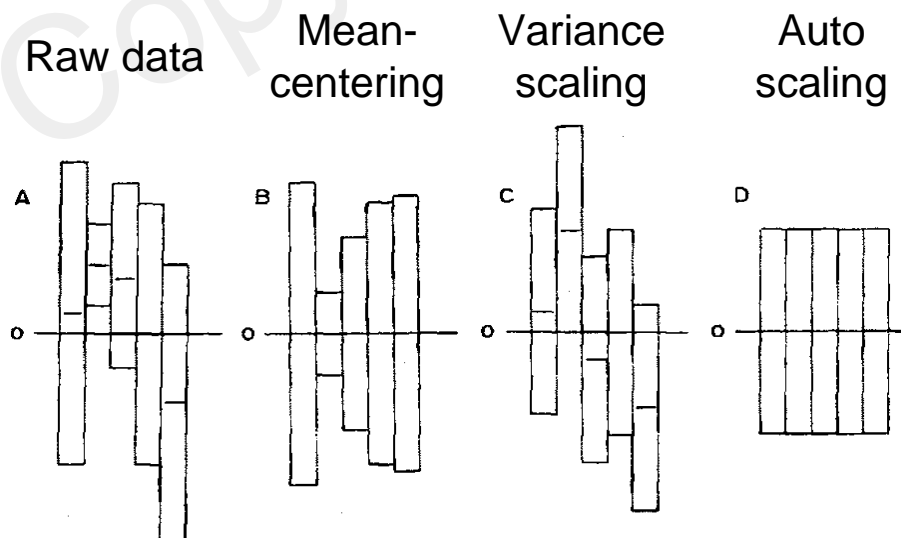
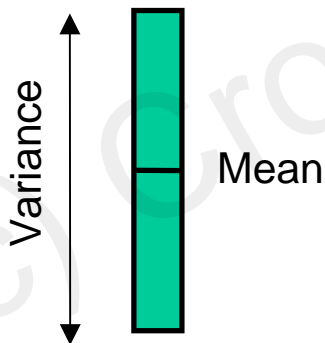
Variance scaling

$$\tilde{\mathbf{X}}_{i,k} = \mathbf{X}_{i,k} \times \frac{1}{\text{var}(\mathbf{X}_{:,k})}$$

Preprocessed data sample i , mass k Raw data sample i , mass k Variance of mass k

- Divide each variable by its variance in the dataset
- Equalises importance of each variable (i.e. mass)
- Problematic for weak peaks – usually used with peak selection
- Called ‘auto scaling’ if combined with mean centering

For each variable (mass, in SIMS spectrum)

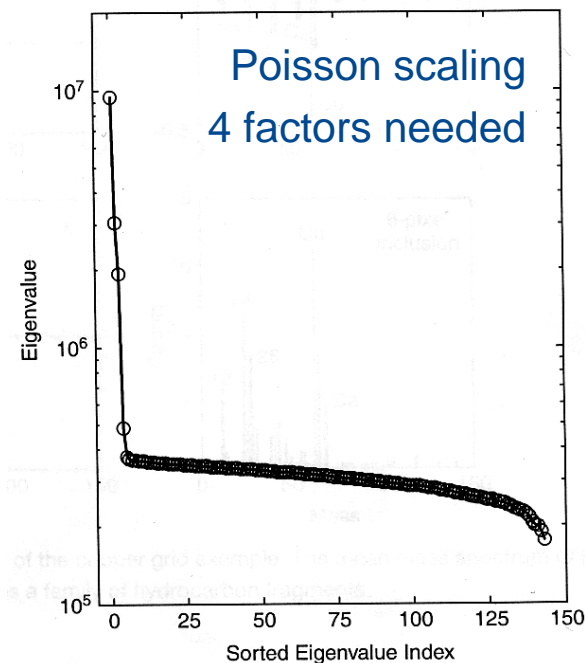
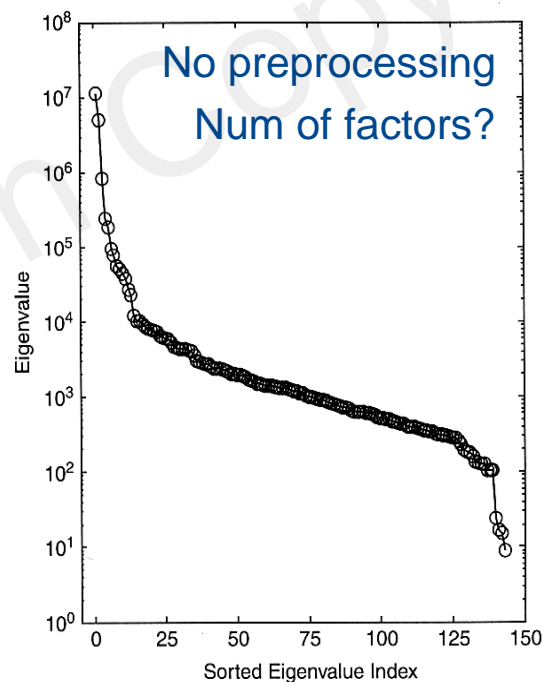


Poisson scaling

$$\tilde{\mathbf{X}}_{i,k} = \mathbf{X}_{i,k} \times \frac{1}{\sqrt{\text{mean}(\mathbf{X}_{i,:})}} \times \frac{1}{\sqrt{\text{mean}(\mathbf{X}_{:,k})}}$$

Preprocessed data sample i , mass k Raw data sample i , mass k Mean intensity of sample i Mean intensity of mass k

- PCA assumes the error associated with each data point is equal
- But SIMS data is dominated by Poisson counting noise – noise variance of a peak is proportional to its intensity
- Divide each data point by the square root of the mean sample intensity and the square root of the mean spectrum
- Provides improved noise rejection in PCA

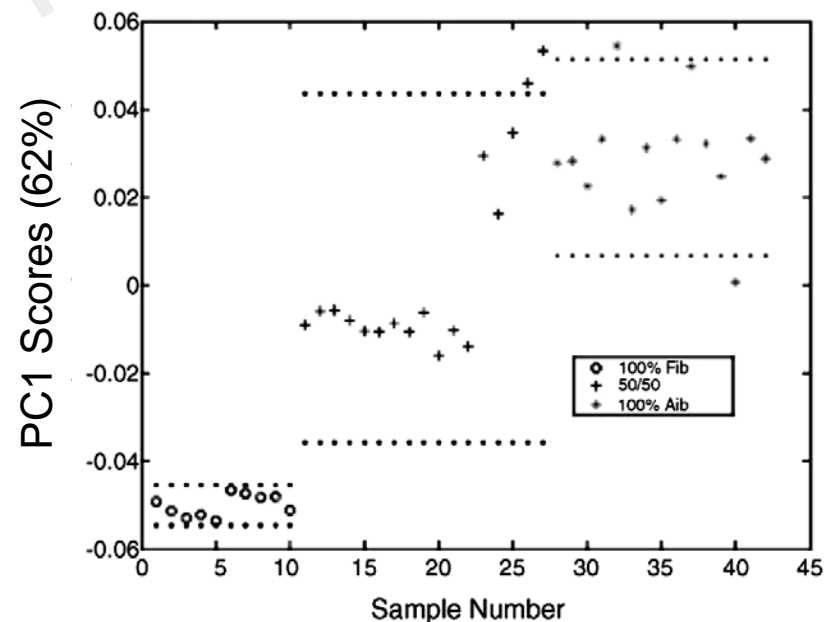
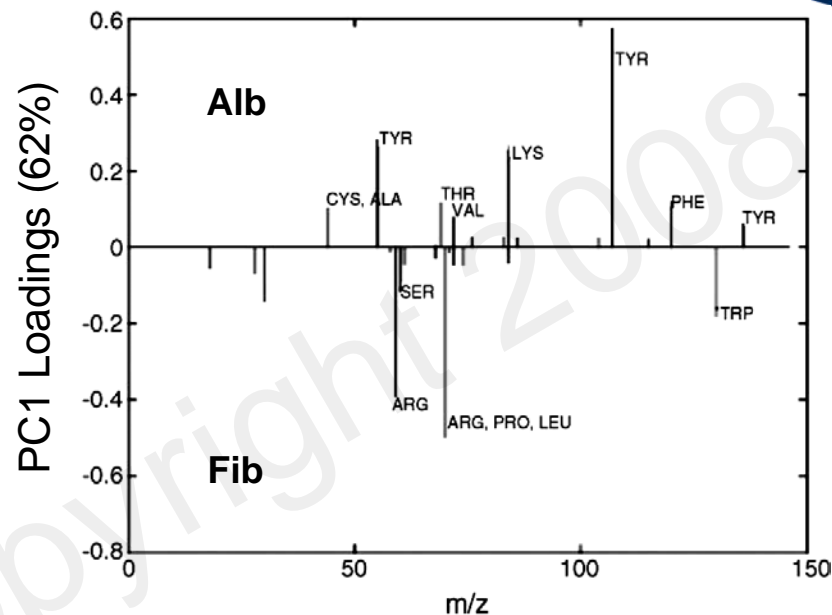


Data preprocessing summary

Method of preprocessing	Effect of preprocessing
No preprocessing	First factor goes from origin to mean of data
Mean centering	All factors describe variations from the mean
Normalisation	Equalises total ion yield of each sample and emphasise relative changes in ion intensities
Variance scaling	Equalises variance of every peak regardless of intensity. Best with peak selection.
Poisson scaling	Equalises <i>noise</i> variance of each data point. Provides greater noise rejection.

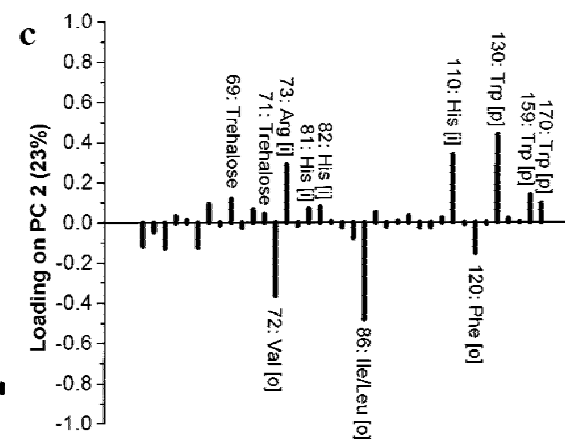
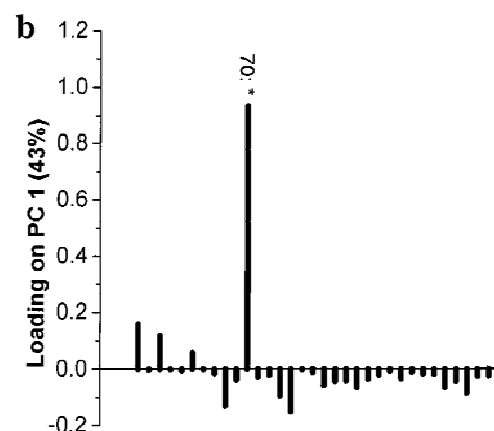
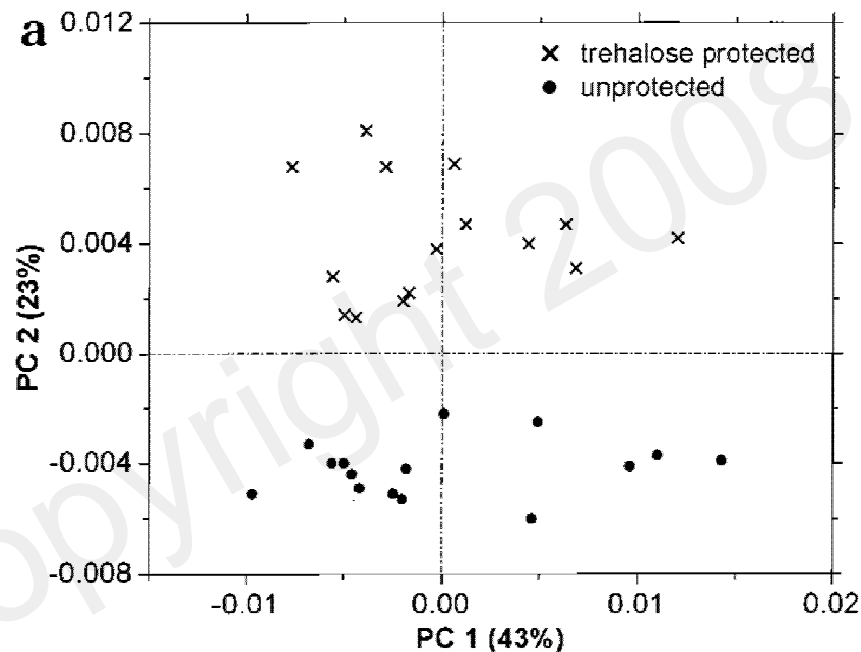
PCA example (1)

- Three protein compositions (100% fibrinogen, 50% fibrinogen / 50% albumin, 100% albumin) adsorbed onto poly(DTB suberate)
- Loadings on first factor (PC1) shows relative abundance of amino acid peaks of two proteins
- Scores on PC1 separates samples based on protein composition



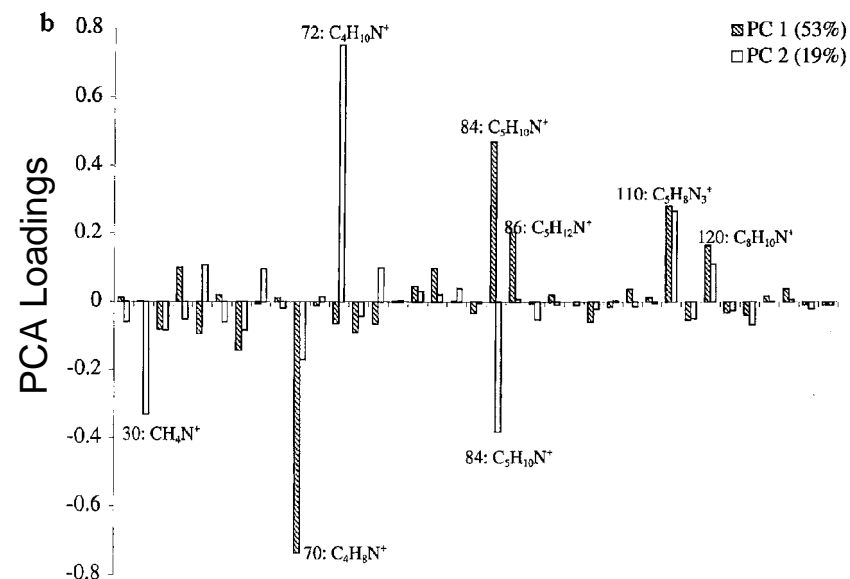
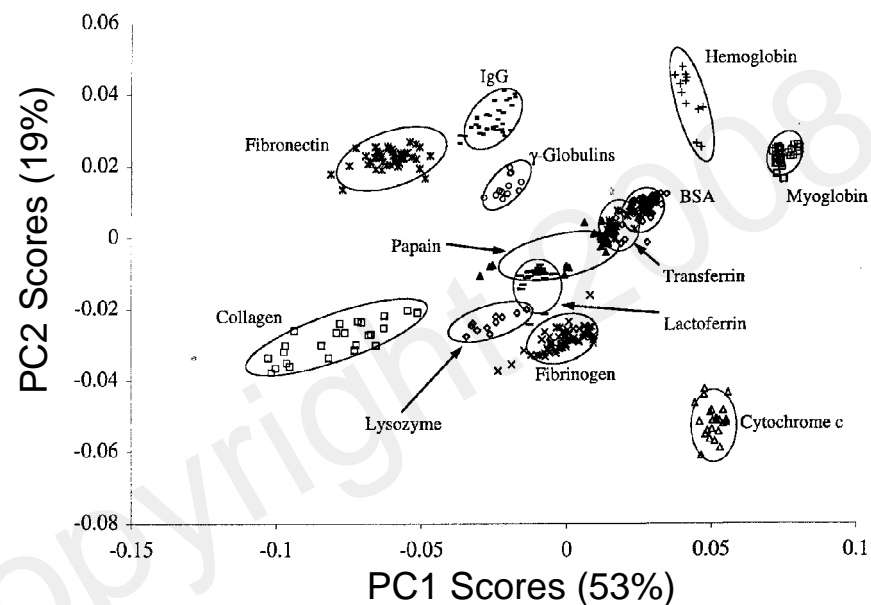
PCA example (2)

- SIMS spectra acquired for antiferritin with or without trehalose coating
- Largest variance (PC 1) arises from sample heterogeneity
- PC 2 distinguishes samples protected by trehalose – higher intensities of polar and hydrophilic amino acid fragments
- Trehalose preserves protein conformation in UHV



PCA example (3)

- 16 different single protein films adsorbed on mica
- Excellent classification of proteins using only 2 factors
- Factors consistent with total amino acid composition of various proteins
- 95% confidence limits provide means for identification / classification



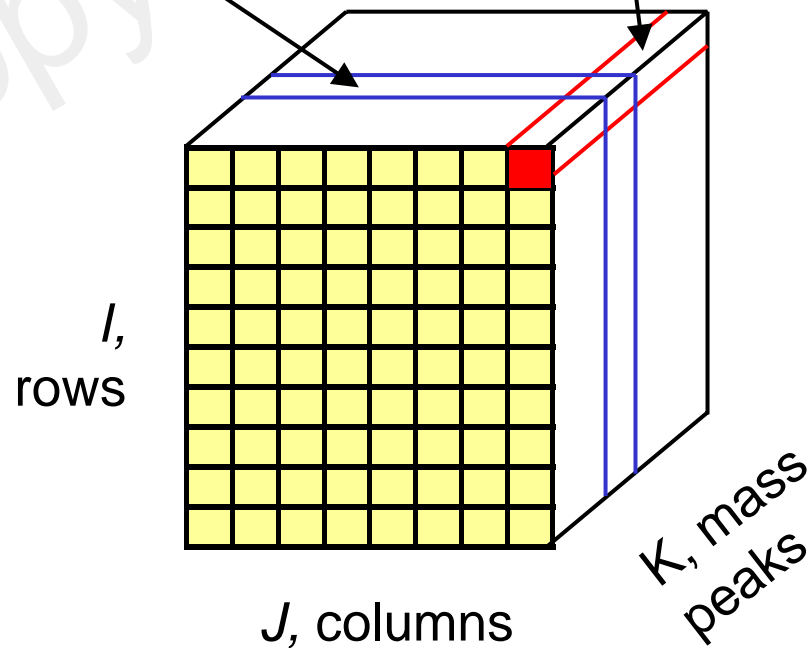
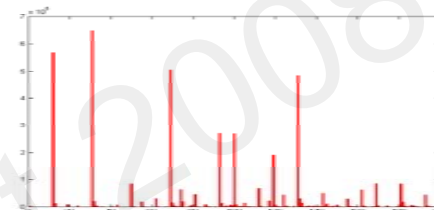
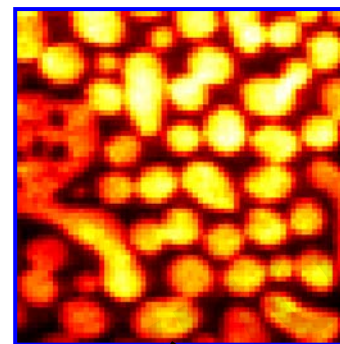
PCA image analysis

- 'Datacube' contains a raster of $I \times J$ pixels and K mass peaks
- The datacube is rearranged into 2D data matrix with dimensions $[(I \times J) \times K]$ prior to PCA – 'unfolding'
- PCA results are folded to form scores images prior to interpretation

1	4	7
2	5	8
3	6	9

unfold →

1
2
3
4
5
6
7
8
9

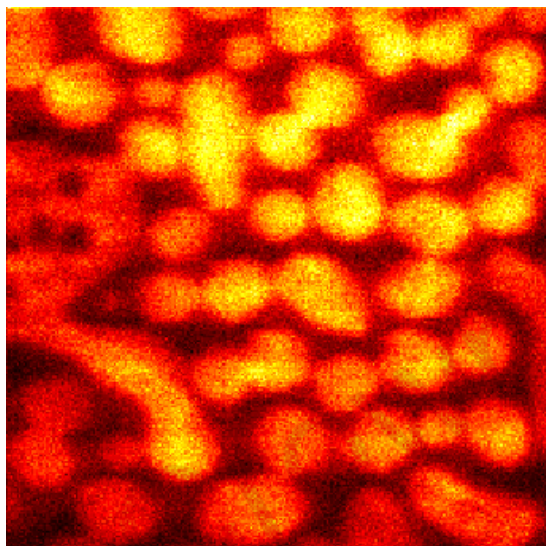


PCA image example (1)

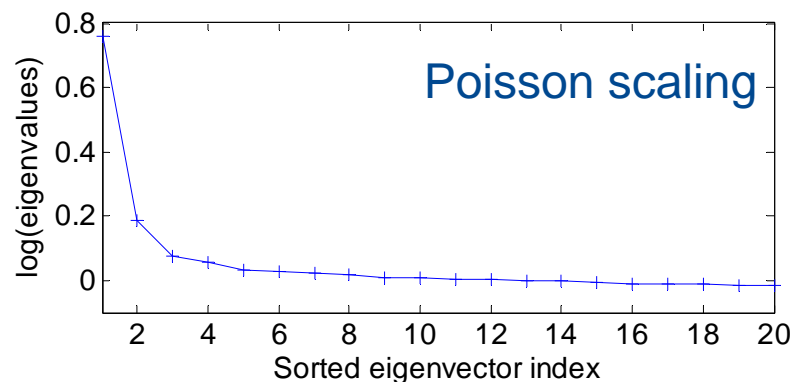
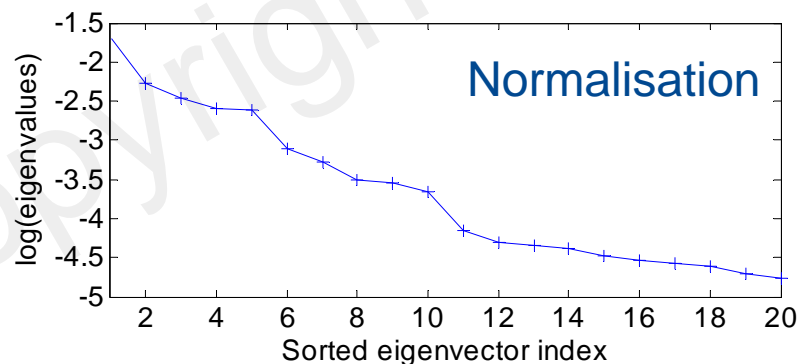
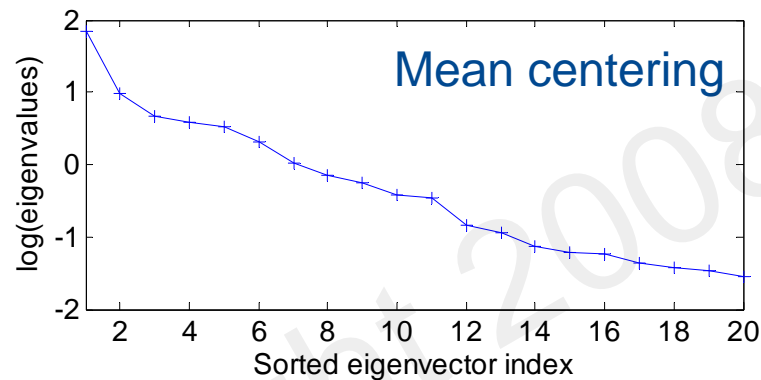
Immiscible PC / PVC polymer blend

42 counts per pixel on average

Total ion image

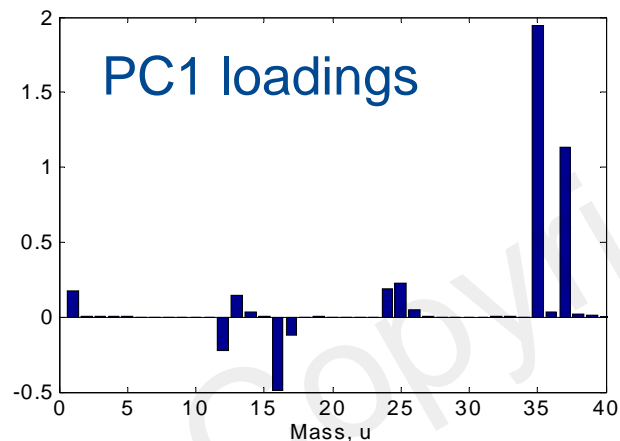
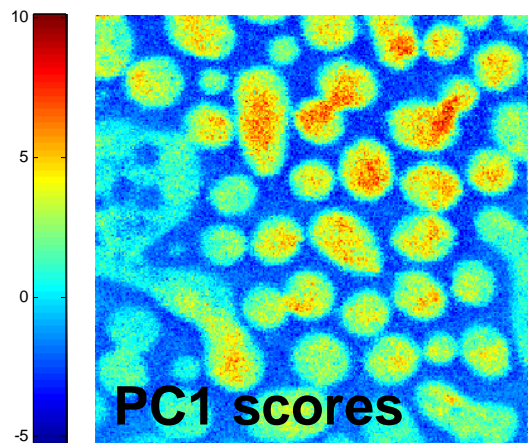


Only 2 factors needed –
dimensionality of image reduced
by a factor of 20!

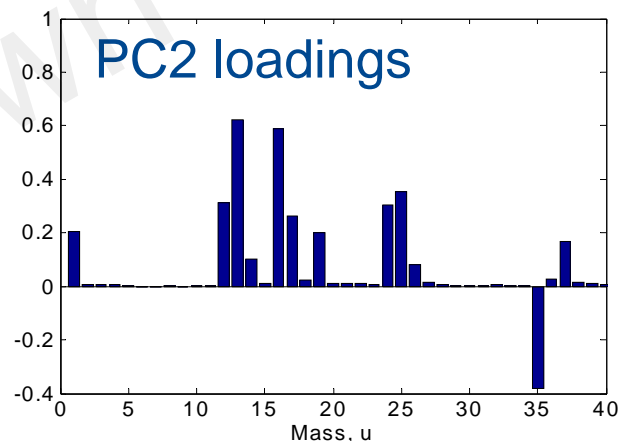
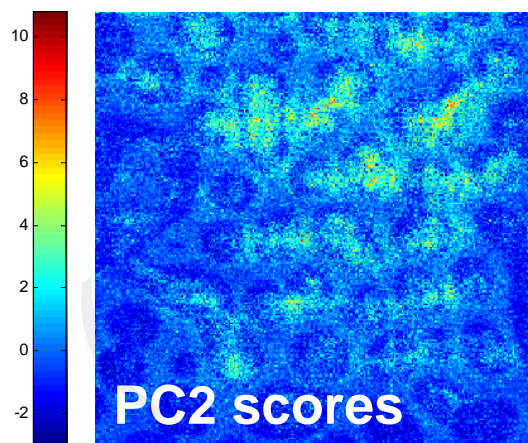


PCA image example (1)

PCA results after Poisson scaling and mean centering



1st factor distinguishes PVC and PC phases



2nd factor shows detector saturation for intense ³⁵Cl peak

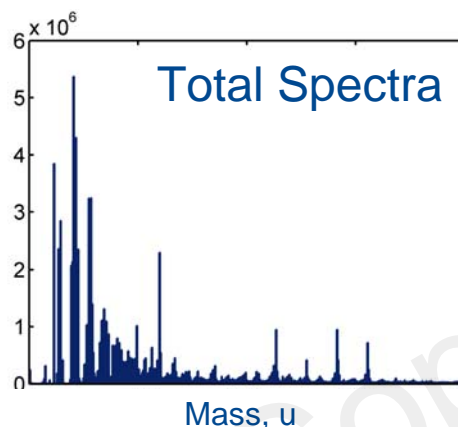
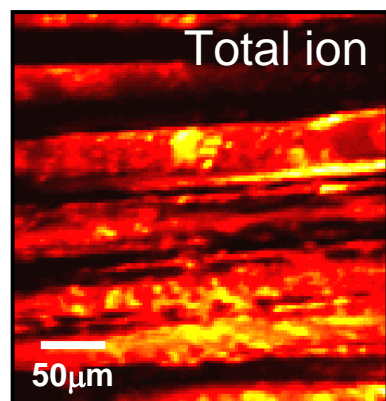
PCA image example (2)

Hair fibre with multi-component pretreatment



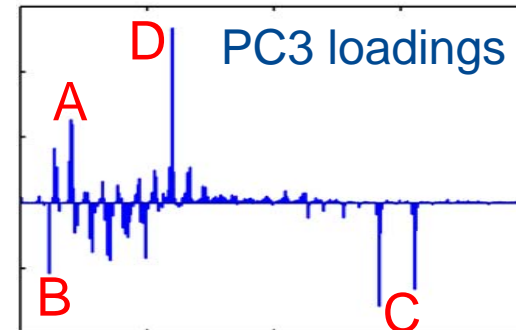
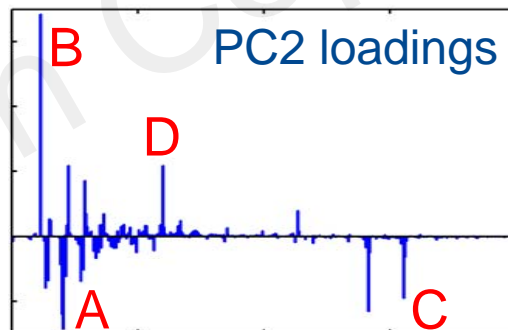
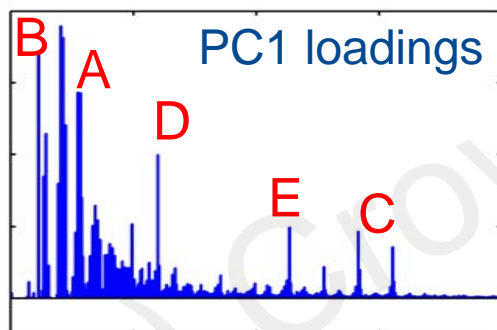
PCA image example (2)

Hair fibre with multi-component pretreatment



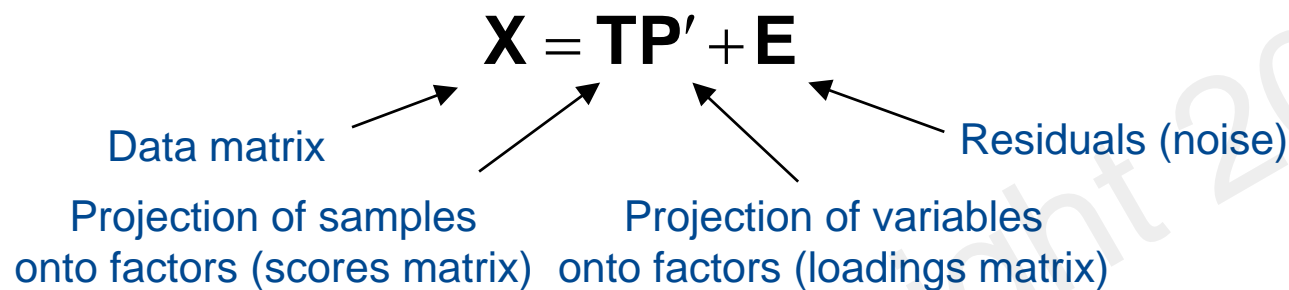
Intertek

Image courtesy of Dr
Ian Fletcher, Intertek MSG



PCA factors are linear combinations of chemical components and optimally describe *variance* – PCA results can be difficult to interpret!

I = no. of samples
 K = no. of mass units
 N = no. of factors



- PCA describes the original data using **factors**, consisting of **loadings** and **scores** which efficiently accounts for variance in the data
- **Eigenvalues** give the variance captured by the corresponding factors
- **Data preprocessing** method needs to be selected with care
- PCA is excellent for **discrimination** and **classification** based on differences in spectra, and for **identifying important mass peaks**
- PCA factors optimally describe variance – PCA results **may be difficult to interpret**

1. Introduction
2. Linear algebra
- 3. Factor analysis**
 - Principal component analysis (PCA)
 - Data preprocessing
 - PCA Examples
 - **Multivariate curve resolution (MCR)**
 - **MCR Examples**
4. Multivariate regression
5. Classification
6. Conclusion

Multivariate curve resolution (MCR)

- PCA factors are directions that describes variance
 - positive and negative peaks in the loadings
 - can be difficult to interpret
- We want to resolve original chemical spectra and reverse the following process:

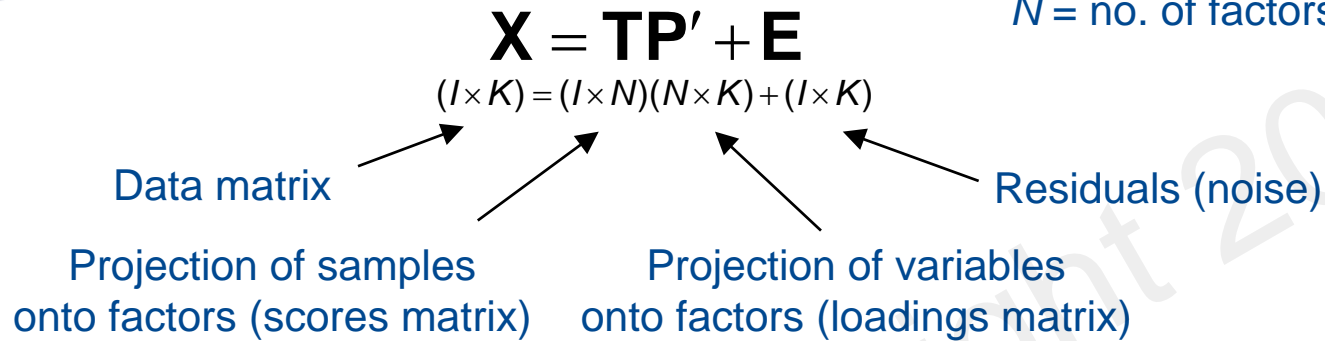
$$\begin{array}{c} \text{Samples} \end{array} \begin{array}{c} \text{Sample} \\ \text{composition} \end{array} \times \begin{array}{c} \text{Chemical spectra} \\ \text{Variables} \end{array} = \begin{array}{c} \text{Data matrix} \\ \text{Variables} \end{array}$$

$$\begin{array}{c} \text{Samples} \\ \left[\begin{array}{cc} 5 & 1 \\ 2 & 4 \\ 0 & 6 \end{array} \right] \begin{array}{c} \text{Chemicals} \\ \left[\begin{array}{ccccc} 1 & 6 & 1 & 0 & 4 \\ 4 & 2 & 5 & 1 & 1 \end{array} \right] \end{array} = \begin{array}{c} \text{Samples} \\ \left[\begin{array}{ccccc} 9 & 32 & 10 & 1 & 21 \\ 18 & 20 & 22 & 4 & 12 \\ 24 & 12 & 30 & 6 & 6 \end{array} \right] \end{array}$$

- Use multivariate curve resolution (MCR)

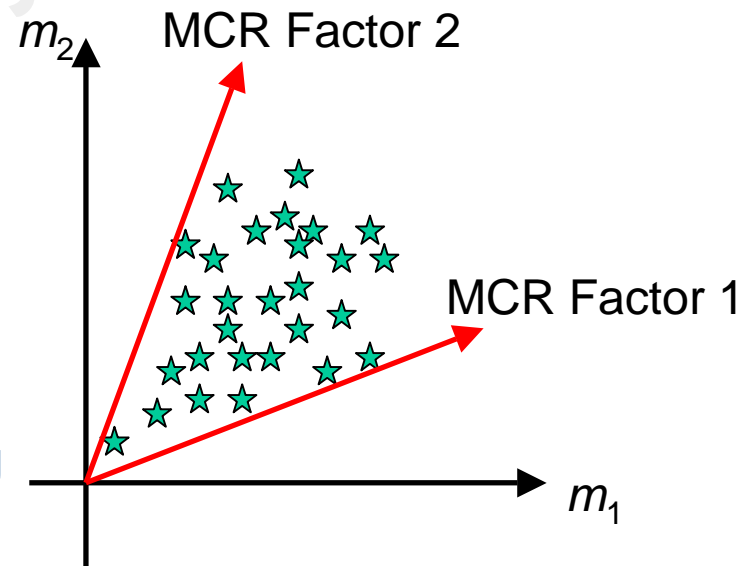
Multivariate curve resolution (MCR)

I = no. of samples
 K = no. of mass units
 N = no. of factors



MCR is designed for recovery of **chemical spectra** and **contributions** from a multi-component mixture, when little or no prior information about the composition is available

MCR uses an **iterative least-squares algorithm** to extract solutions, while applying suitable constraints e.g. **non-negativity**



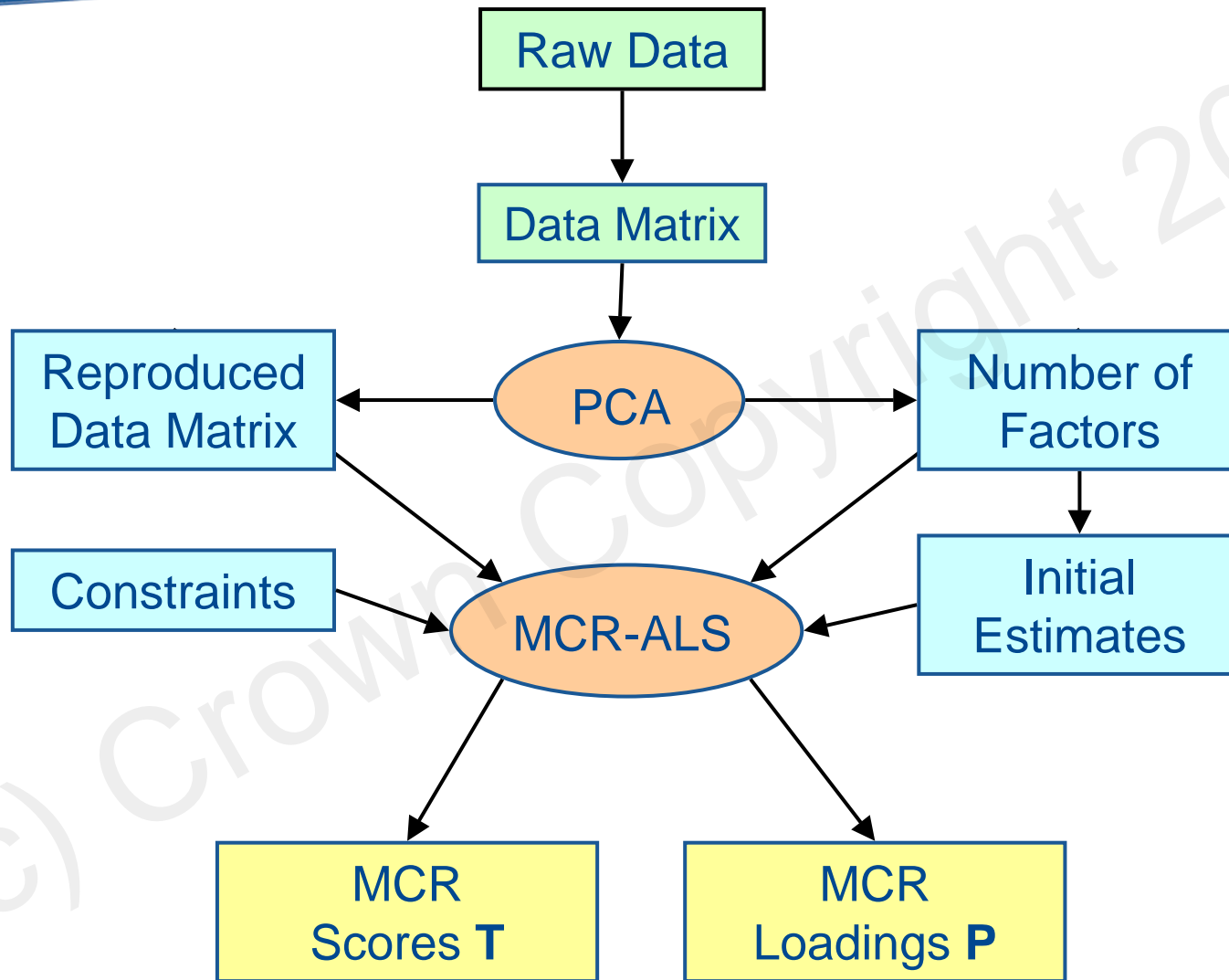
Multivariate curve resolution (MCR)

Six Steps to MCR Results

1. Determine number of factors N via eigenvalue plot
2. Obtain PCA reproduced data matrix for N factors
3. Obtain initial estimates of spectra (loadings) or contributions (scores)
 - Random initialisation
 - PCA loadings or scores
 - Varimax rotated PCA loadings or scores
 - Pure variable detection algorithm e.g. SIMPLISMA
4. Constraints
 - Non-negativity
 - Equality
5. Convergence criterion
6. Alternating least squares (ALS) optimisation



Outline of MCR



MCR-ALS algorithm

- Start with PCA reproduced data matrix

$$\bar{\mathbf{X}} = \mathbf{T}\mathbf{P}'$$

- Assume initial estimate of loadings \mathbf{P}

$$\mathbf{T} = \bar{\mathbf{X}}(\mathbf{P}')^+$$

(1) Find estimate of \mathbf{T} using \mathbf{P} , applying constraints

$$\mathbf{P}' = \mathbf{T}^+\bar{\mathbf{X}}$$

(2) Find new estimate of \mathbf{P} using \mathbf{T} , applying constraints

$$\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}'$$

(3) Compute MCR reproduced matrix

$$\mathbf{E} = \hat{\mathbf{X}} - \bar{\mathbf{X}}$$

(4) Compare results and check convergence

⋮

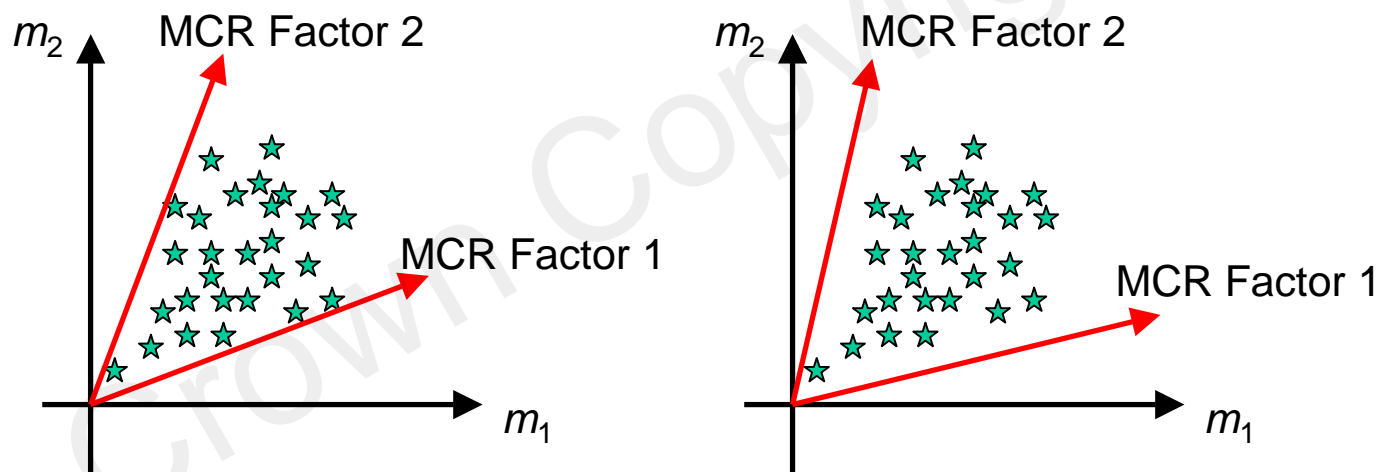
- Steps (1) – (4) are repeated until MCR loadings \mathbf{P} and scores \mathbf{T} are able to reconstruct reproduced data matrix $\bar{\mathbf{X}}$ within acceptable error specified in convergence criterion

Pseudoinverse of rectangular matrix

$$\mathbf{A}^+ = \mathbf{A}'[\mathbf{A}\mathbf{A}']^{-1}$$

Rotational ambiguity

- MCR solutions are not unique!
- Accuracy of resolved spectra depends on the existence of pixels or samples where there is only contribution from one chemical component

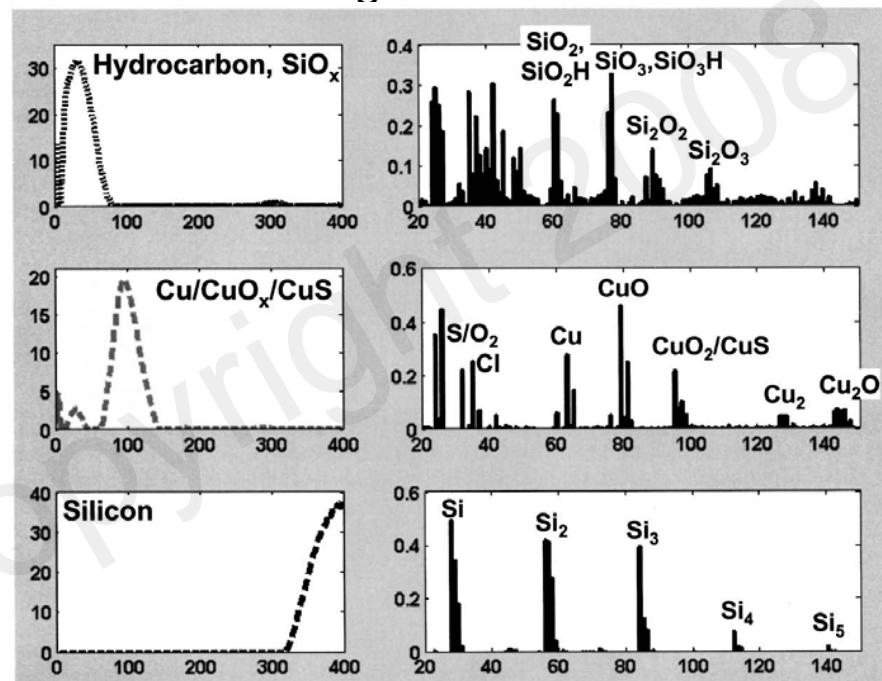


- Peaks for the intense components may appear in spectra resolved for weak components
- Good initial estimates and suitable data preprocessing are essential

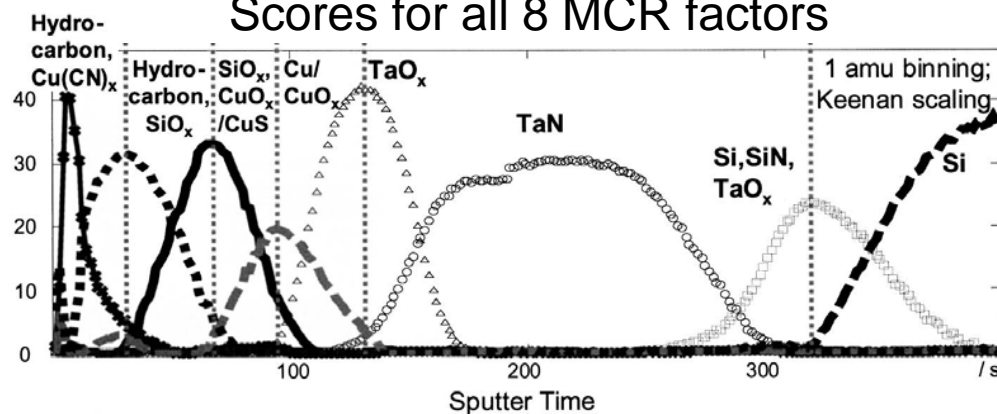
MCR example (1)

- ToF-SIMS depth profiling of copper film grown on TaN coated silicon wafer
- Manual analysis is difficult, e.g. Si^- can arise from SiO_x^- , SiN^- or silicon substrate
- MCR resolves 8 factors. Loadings resemble SIMS spectra of individual phases and scores resemble their contribution to the depth profile
- Improve signal to noise and correlation of related peaks

Scores and loadings for 3 of the MCR factors

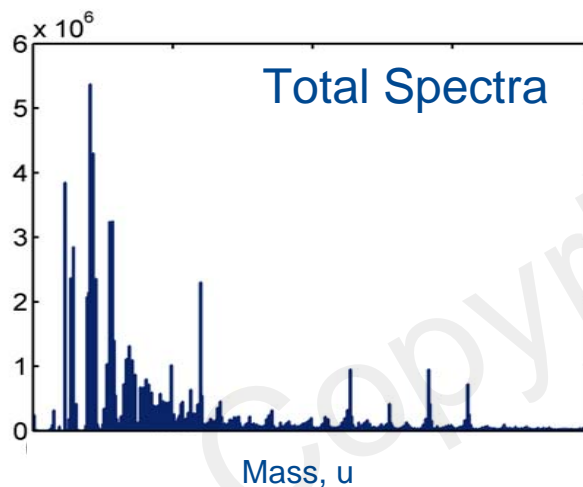
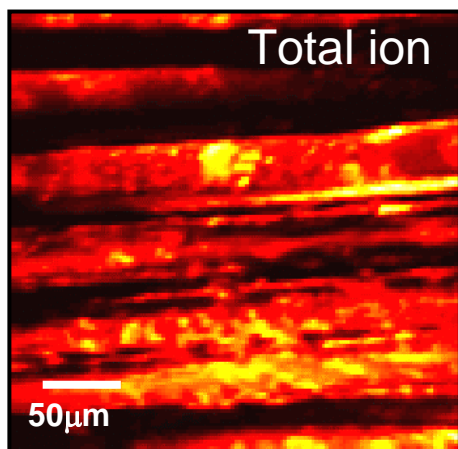


Scores for all 8 MCR factors



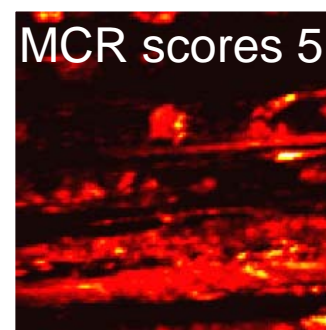
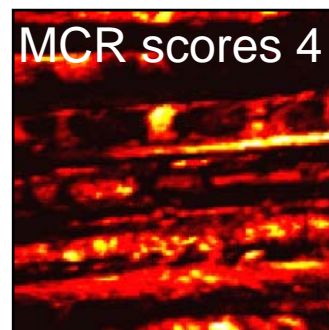
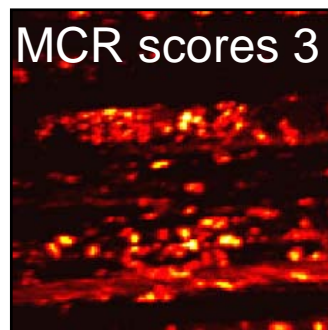
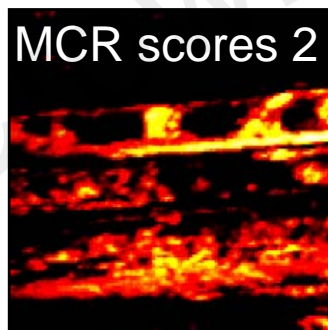
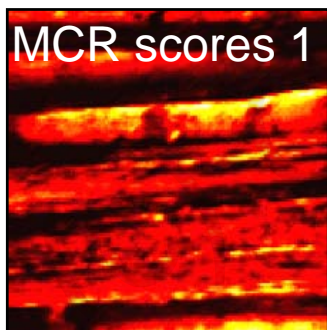
MCR image example (1)

Hair fibre with multi-component pretreatment



Intertek

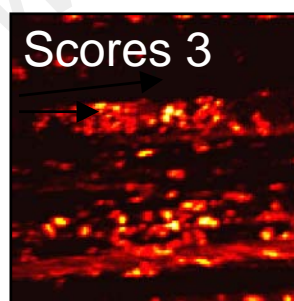
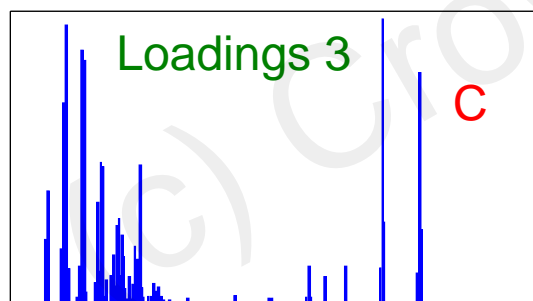
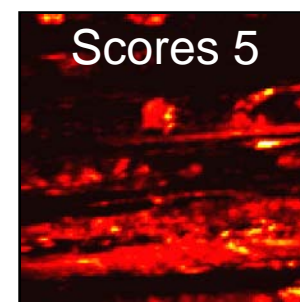
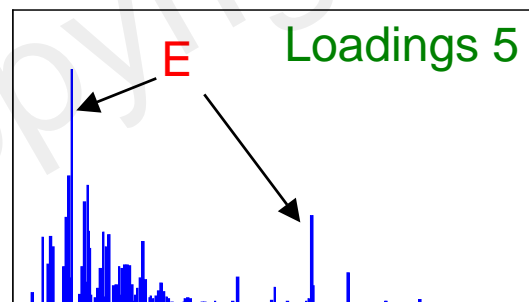
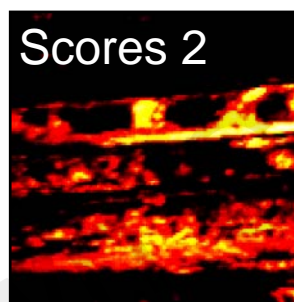
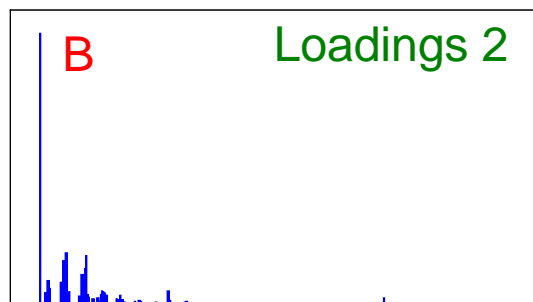
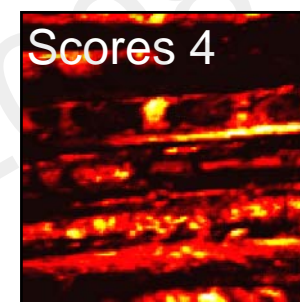
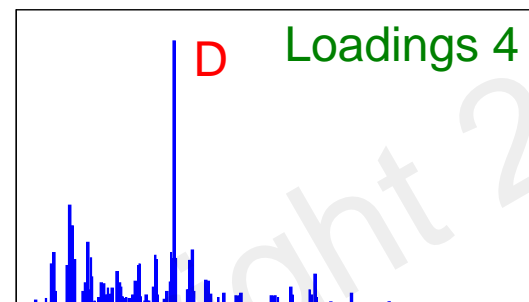
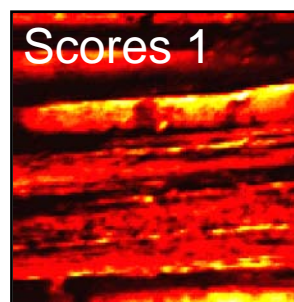
Image courtesy of Dr
Ian Fletcher, Intertek MSG



MCR image example (1)

Intertek

Image courtesy of Dr
Ian Fletcher, Intertek MSG



Mass, u

Distribution and characteristic peaks are obtained for hair fibre and four surface chemicals

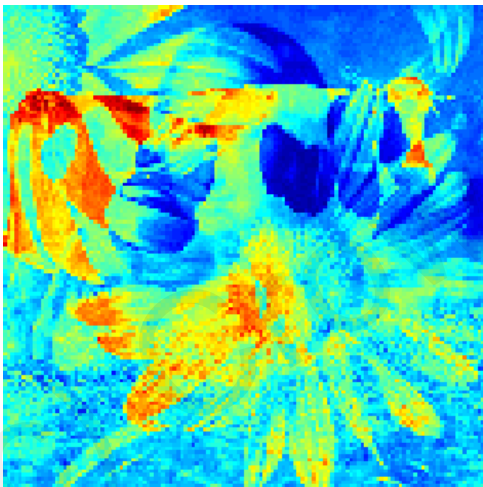
Mass, u

© Crown Copyright 2008

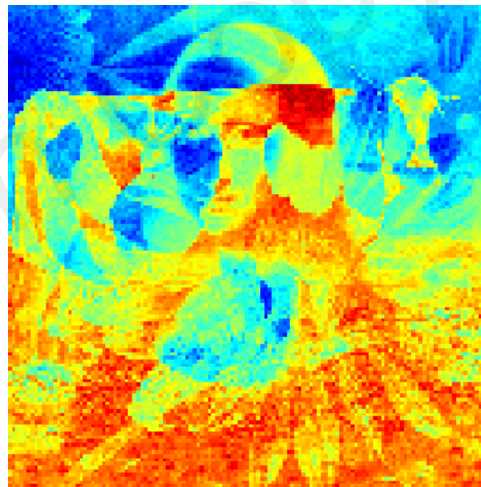
MCR image example (2)

- We take three pictures and assign each with a SIMS spectra (PBC, PC, PVT)
- The pictures are combined to form a multivariate image dataset
- Poisson noise are added to the image (avg ~50 counts per pixel)

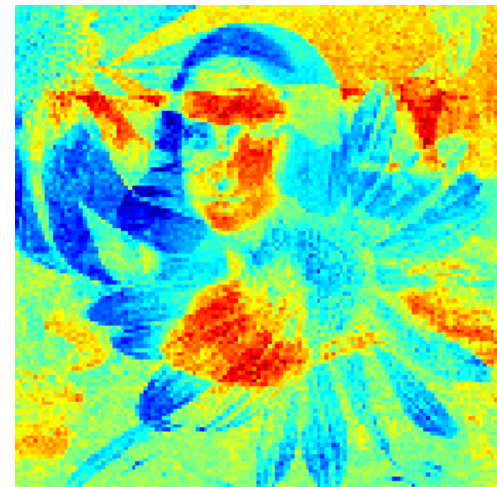
PCA Scores 1



PCA Scores 2



PCA Scores 3



MCR image example (2)

- We take three pictures and assign each with a SIMS spectra (PBC, PC, PVT)
- The pictures are combined to form a multivariate image dataset
- Poisson noise are added to the image (avg ~50 counts per pixel)

MCR Scores 1



MCR Scores 2

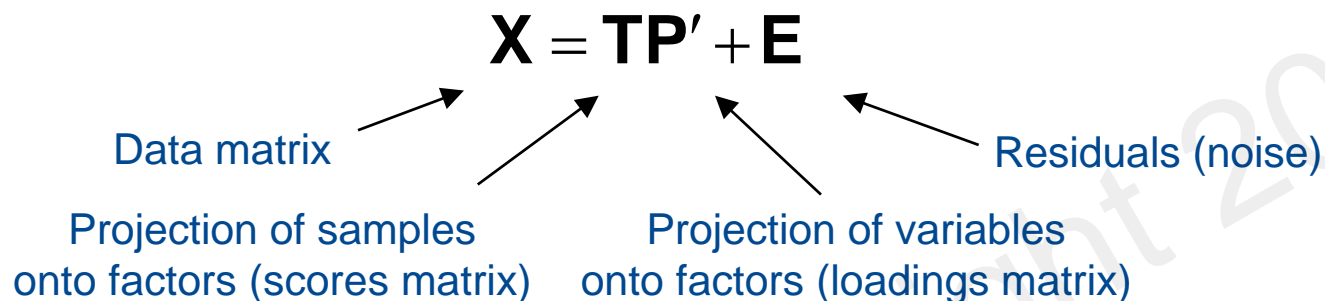


MCR Scores 3



MCR resolves the original images unambiguously!

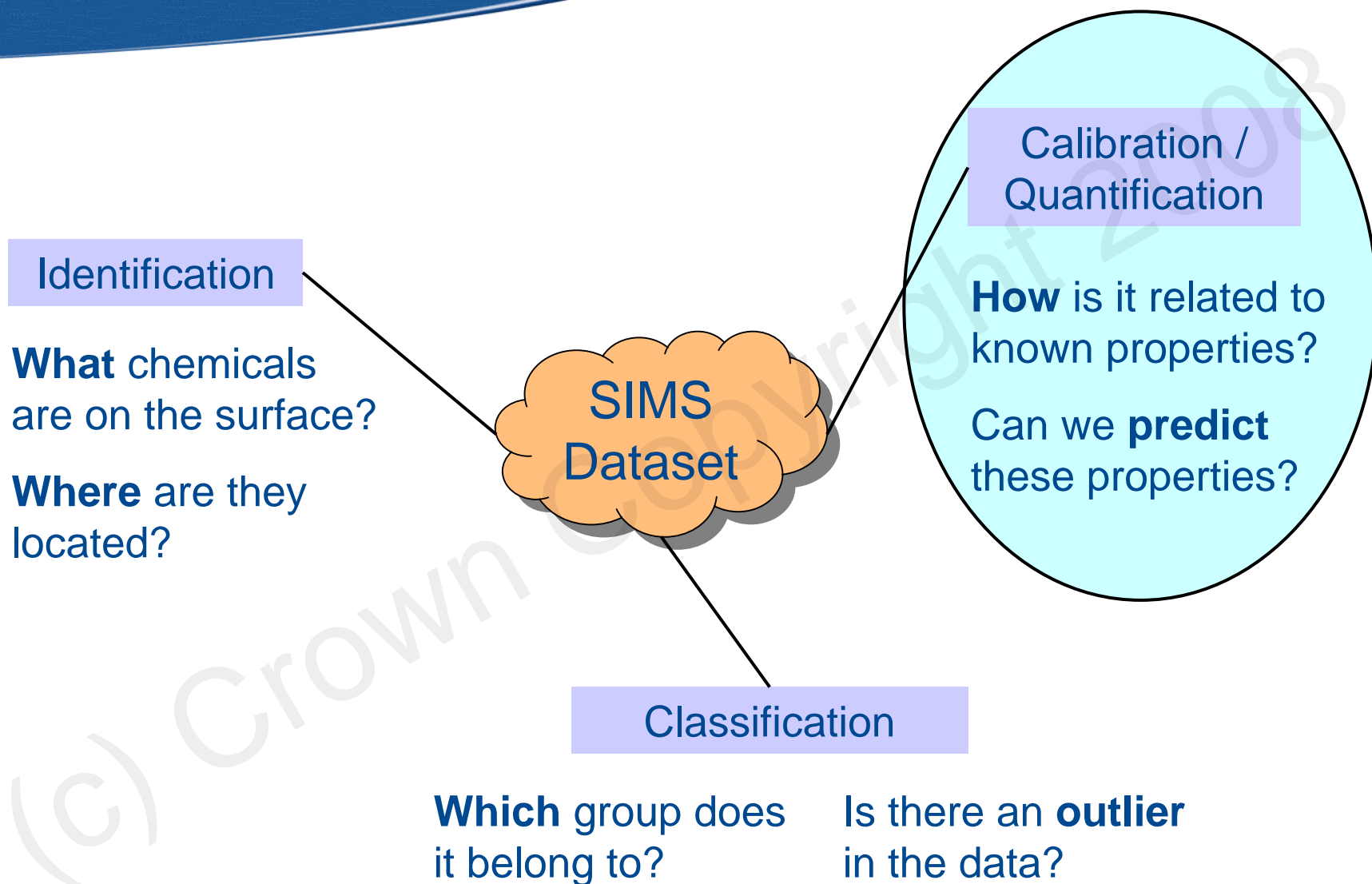
I = no. of samples
 K = no. of mass units
 N = no. of factors



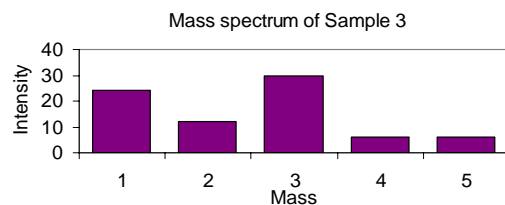
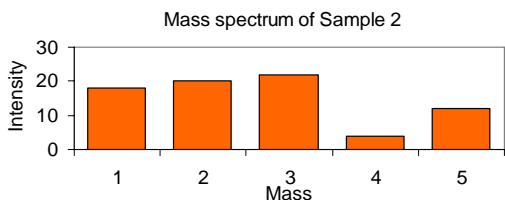
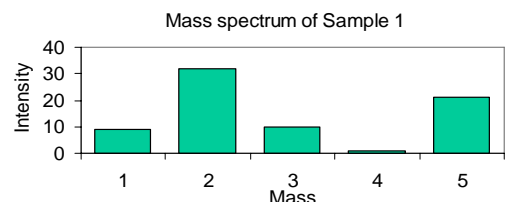
- MCR describes the original data using **factors**, consisting of **loadings** and **scores** which which resembles **chemical spectra** and **contributions** from a multi-component mixture, respectively
- MCR uses an **iterative algorithm** to extract solutions, while applying suitable constraints e.g. **non-negativity**
- Good **initial estimates** and suitable **data preprocessing** are essential
- MCR is excellent for **identification** and **localisation** of chemicals in complex mixtures and allows for **direct interpretation**

Contents

1. Introduction
2. Linear algebra
3. Factor analysis
4. Multivariate regression
 - **Multiple linear regression (MLR)**
 - **Principal component regression (PCR)**
 - **Partial least squares regression (PLS)**
 - **PLS Examples**
 - **Calibration, validation and prediction**
5. Classification
6. Conclusion



Regression analysis



Measured properties

	XPS measurement	Molecular weight	Concentration ratio
Sample 1	5	1	3
Sample 2	2	4	7
Sample 3	1	6	4

Can we predict the properties of similar materials from their SIMS spectra?

$$y = f(\mathbf{x}) + e$$

$$y = b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots + b_m x_m + e$$

'Response' variable
i.e. measured property

Regression coefficient

'Predictor' variable
i.e. intensity at mass m

Multiple linear regression (MLR)

I = no. of samples
 K = no. of mass units
 M = no. of response variables

- Extending to I samples and M response variables

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

$(I \times M) = (I \times K)(K \times M) + (I \times M)$

Response variables \rightarrow \mathbf{Y}

SIMS data matrix \rightarrow \mathbf{X}

Regression matrix \rightarrow \mathbf{B}

Residuals (noise) \rightarrow \mathbf{E}

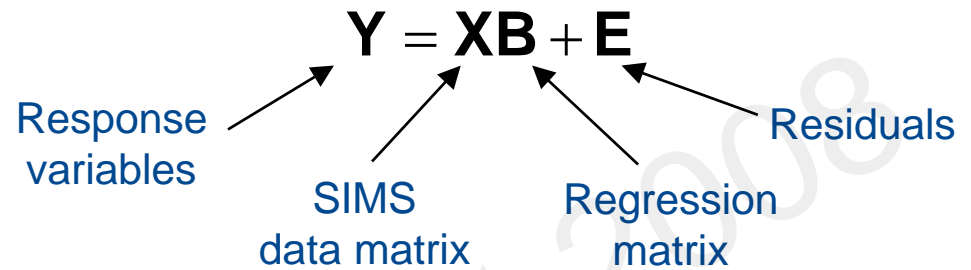
- Least squares solution (MLR solution)

$$\mathbf{B} = \mathbf{X}^+ \mathbf{Y} \quad \text{or} \quad \mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

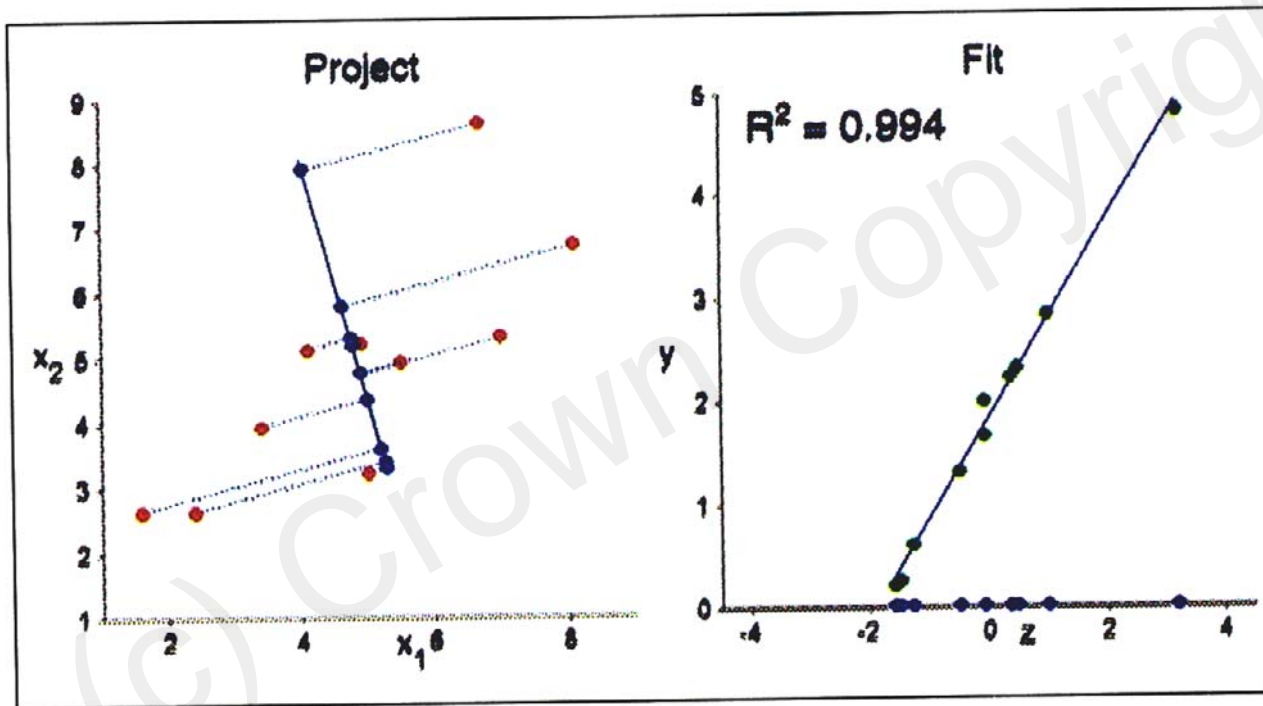
$$\left[\begin{array}{l} \mathbf{X}^+ = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \\ \text{is the pseudoinverse of } \mathbf{X} \end{array} \right]$$

This is the **covariance matrix** of \mathbf{X} ! In SIMS this is likely to be close to **singular** and a well defined **inverse matrix cannot be found**. This is due to the problem of **collinearity**, caused by **linearly dependent** rows or columns in the matrix.

MLR - graphical representation



We relate Y to the projection of X onto B –



MLR finds the least squares solution that minimises E i.e. the best R^2 correlation between Y and the projections of data onto the regression vector XB

Large number of correlated variables (e.g. mass) \rightarrow Risk of overfitting!

Principal component regression (PCR)

I = no. of samples
 N = no. of PCA factors
 M = no. of response variables

- PCA reduces dimensionality of data and reduces effect of noise
- PCA scores matrix is the coordinates of data points in reduced factor space
- Hence we can use PCA scores matrix \mathbf{T} in our linear regression

$$\mathbf{Y} = \mathbf{TB} + \mathbf{E}$$

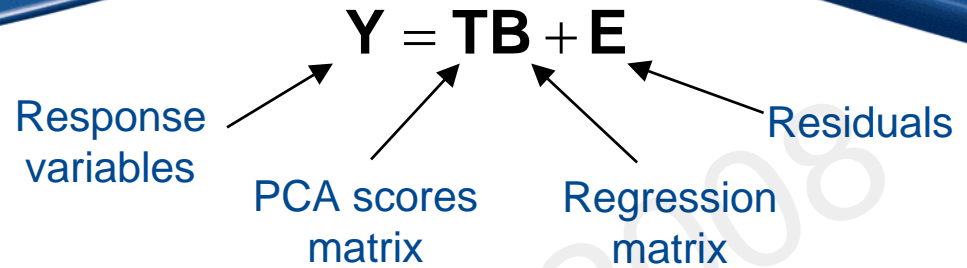
$(I \times M) = (I \times N)(N \times M) + (I \times M)$

Response variables → \mathbf{Y}
PCA Scores Matrix → \mathbf{T}
Regression matrix → \mathbf{B}
Residuals → \mathbf{E}

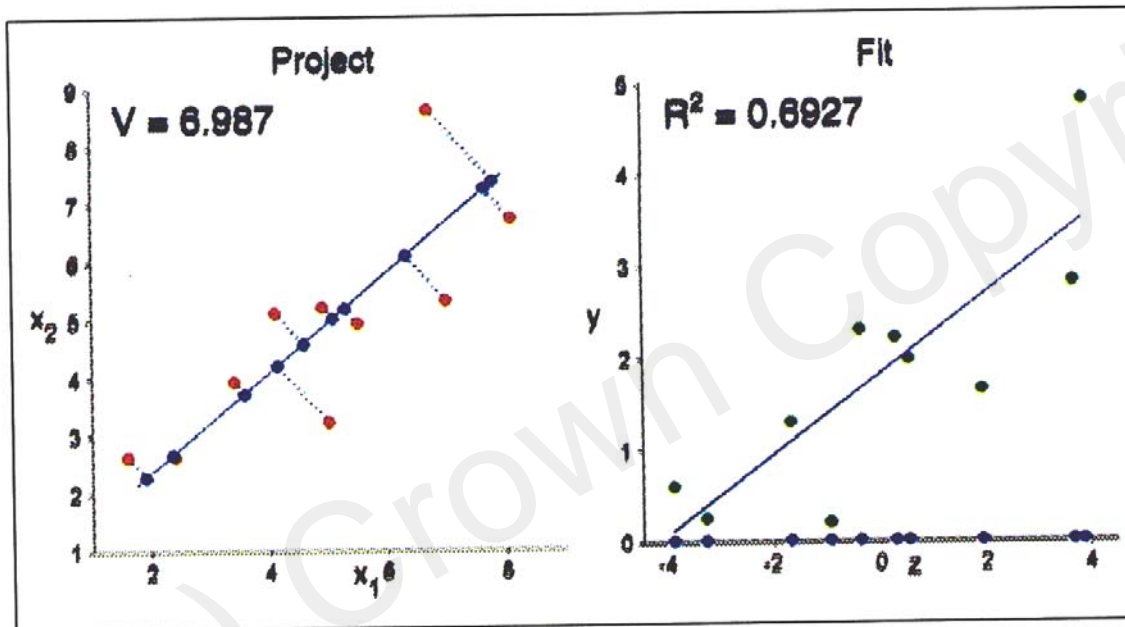
$$\mathbf{B} = \mathbf{T}^+ \mathbf{Y}$$
$$\mathbf{B} = (\mathbf{T}'\mathbf{T})^{-1} \mathbf{T}'\mathbf{Y}$$

These are now guaranteed to be invertible since the rows of PCA scores matrix are orthogonal

PCR – graphical representation



One factor PCR example –



PCR finds correlation between Y and projection of data onto first PCA factor (scores T).

For more than one factor, PCR finds **linear combinations of scores T** on each PCA factor that are best for predicting Y

Important to determine appropriate number of factors to include in PCR model

Partial least squares regression (PLS)

X = SIMS data matrix

Y = Response variables

The problem with PCR

- PCR uses PCA scores T are computed to model **variations in X** only!
- By choosing directions that **maximise the variance in data X** we hope to include important information which relates the original variables to Y
- First few PCA factors of X may contain only matrix, topographical or other effects, and may have **no relation to quantities Y** which we want to predict

Introducing PLS!

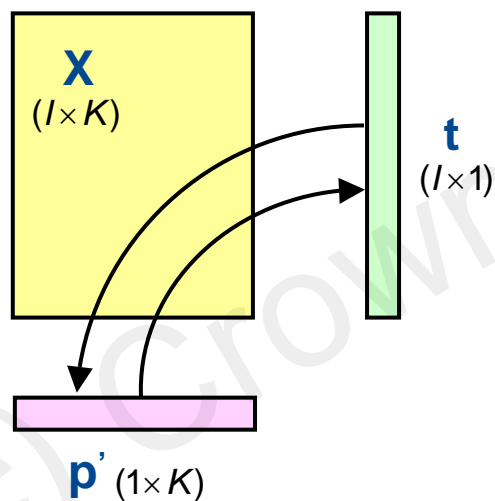
- PLS extracts scores T that are **common** to both X and Y , using **simultaneous decomposition** of X and Y
- It finds factors describing large amounts of **covariance** between X and Y
- It **removes redundant information** from the regression i.e. factors describing X that has no correlation with Y
- More **viable, robust** solution using **fewer** number of factors

PLS (NIPALS) algorithm

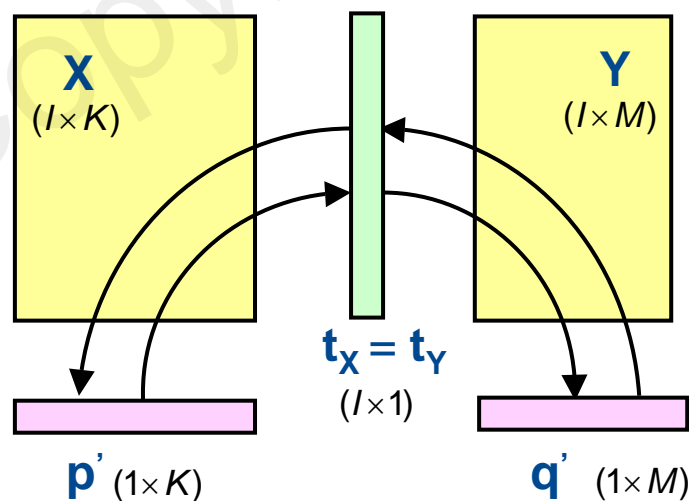
I = no. of samples
 K = no. of mass units
 M = no. of response variables
 N = no. of PCA factors

For decomposition of single matrix \mathbf{X} in PCA, NIPALS calculate \mathbf{t}_1 and \mathbf{p}_1 alternately until convergence. The next set of factors \mathbf{t}_2 and \mathbf{p}_2 are calculated by fitting the residuals (data not explained by \mathbf{p}_1)

(1) PCA decomposition



(2) PLS decomposition



For simultaneous decomposition of \mathbf{X} and \mathbf{Y} , PLS finds a mutual set of scores common to \mathbf{X} and \mathbf{Y} so $\mathbf{t}_X = \mathbf{t}_Y$

PLS formulation

\mathbf{X} = SIMS data matrix

\mathbf{Y} = Response variables

We can now write

$$\begin{array}{l} \mathbf{X} = \mathbf{TP}' + \mathbf{E} \\ \mathbf{Y} = \mathbf{TQ}' + \mathbf{F} \end{array} \quad \left. \vphantom{\begin{array}{l} \mathbf{X} = \mathbf{TP}' + \mathbf{E} \\ \mathbf{Y} = \mathbf{TQ}' + \mathbf{F} \end{array}} \right\}$$

scores residuals

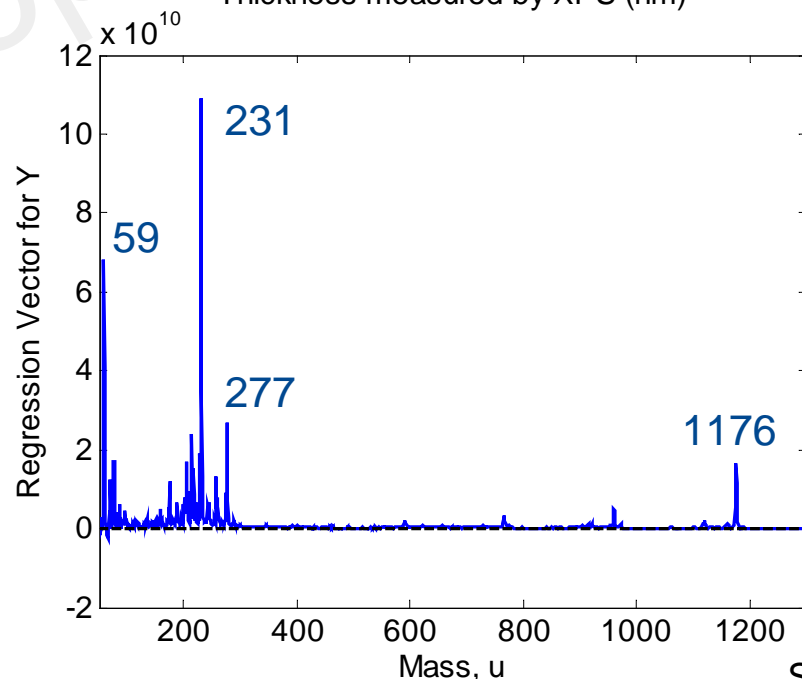
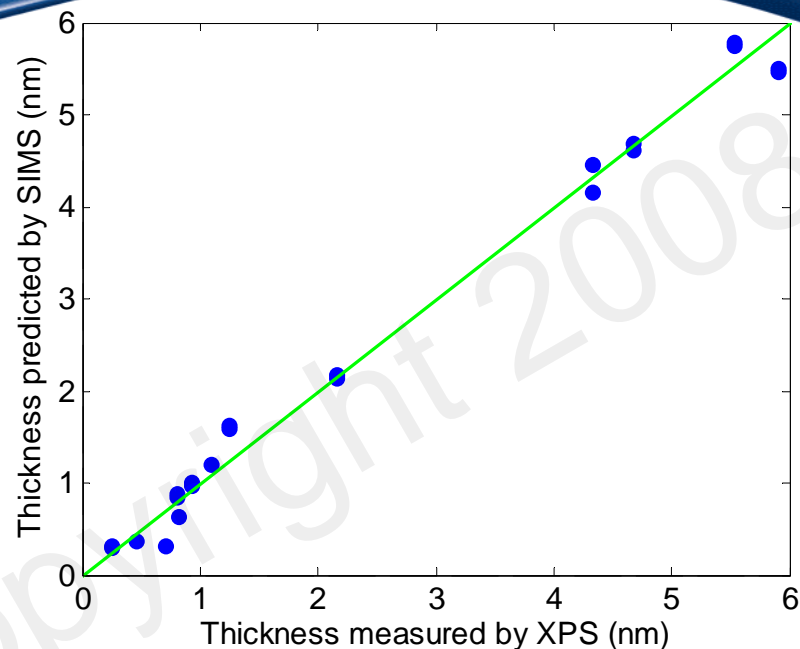
$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$
$$\mathbf{B} = \mathbf{X}^+ \mathbf{Y} = (\mathbf{P}')^+ \mathbf{Q}' = \mathbf{WQ}'$$

regression matrix weights matrix

- \mathbf{W} is the weights matrix and reflects covariance structure between \mathbf{X} and \mathbf{Y}
- \mathbf{T} are PLS scores used to predict \mathbf{Y} from \mathbf{X} . Columns of \mathbf{T} are orthogonal.
- \mathbf{P} and \mathbf{Q} are not orthogonal matrices due to constraint on finding common scores \mathbf{T} . They are sometimes called 'x-loadings' and 'y-loadings' respectively
- In literature '**latent variable**' refers to the set of quantities \mathbf{t} , \mathbf{p} and \mathbf{q} associated with each **PLS factor**

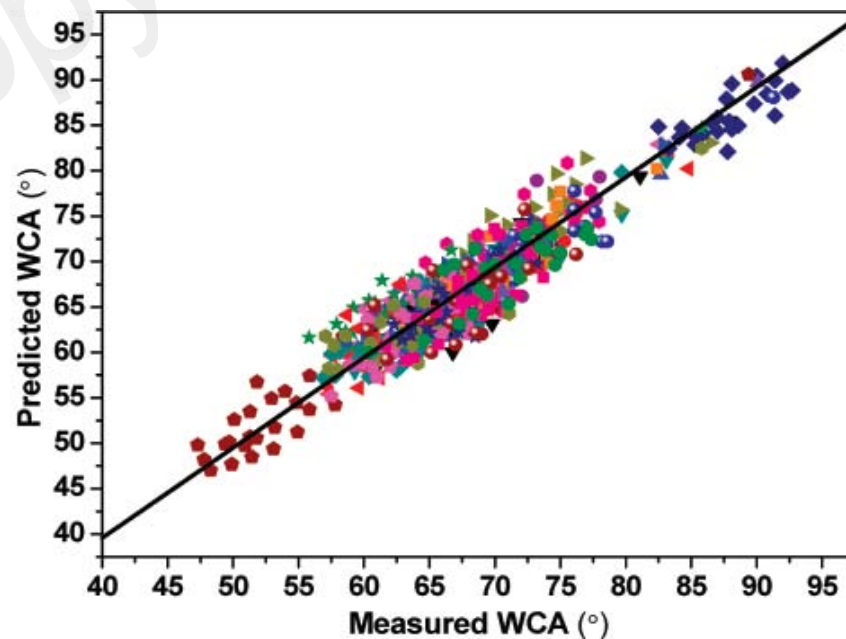
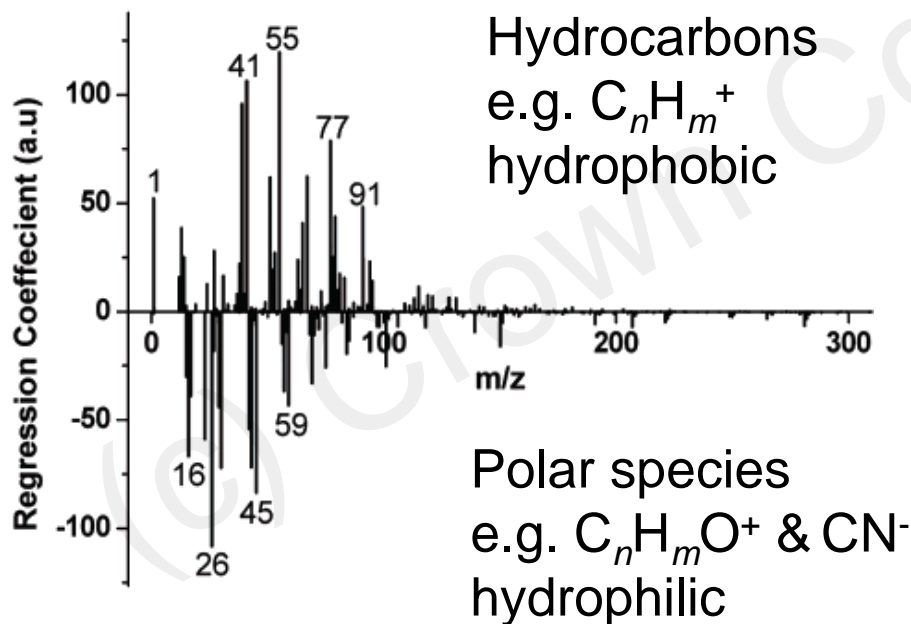
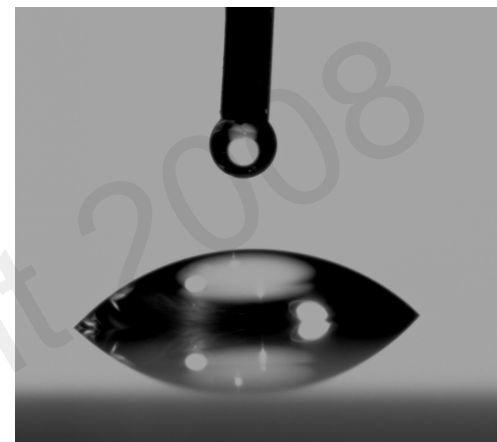
PLS example (1)

- SIMS spectra of thin films of Irganox were compared with their thicknesses measured with XPS
- Two PLS factors are retained, explaining 99.8% of the variance in **X** (SIMS data) and 98.8% of the variance in **Y** (thicknesses)
- PLS model able to predict thicknesses for $t < 6$ nm
- PLS regression vector shows us the SIMS peaks most correlated with thickness



PLS example (2)

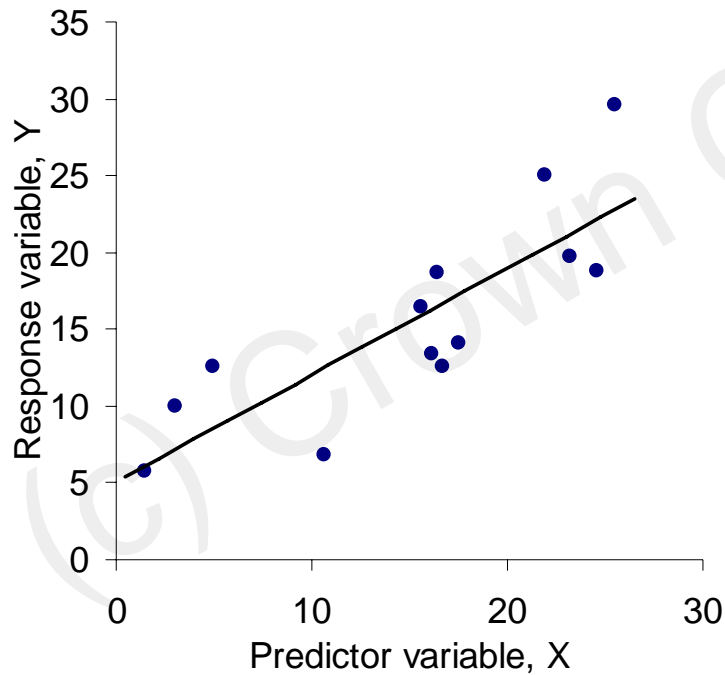
- ToF-SIMS spectra of 576 copolymers are related to their experimental water contact angles (WCA)
- Positive and negative ion spectra are normalised separately, then concatenated (combined) into single data matrix \mathbf{X}



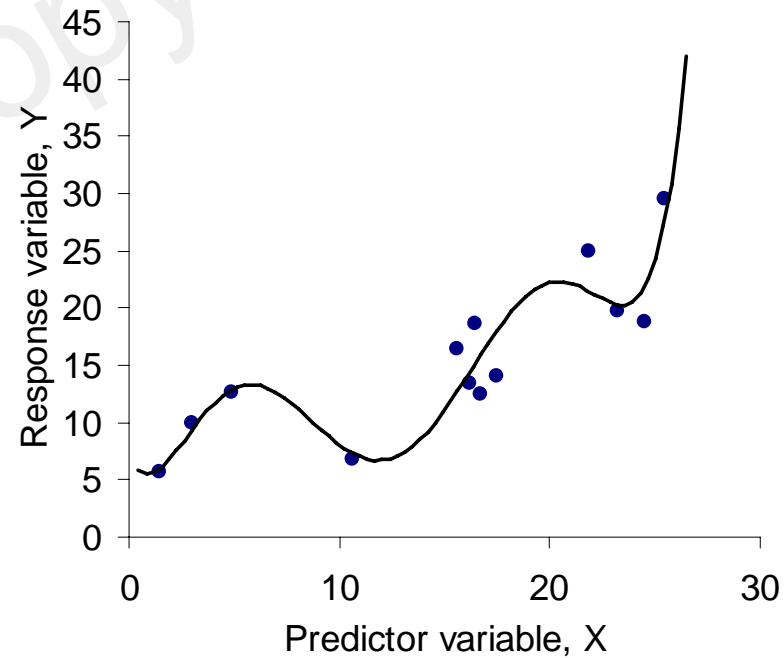
PLS validation

- PLS can be used to build predictive models (calibration)
- Validation is needed to guard against over-fitting
- Without enough data for validation set, cross validation can be useful

Good predictive model

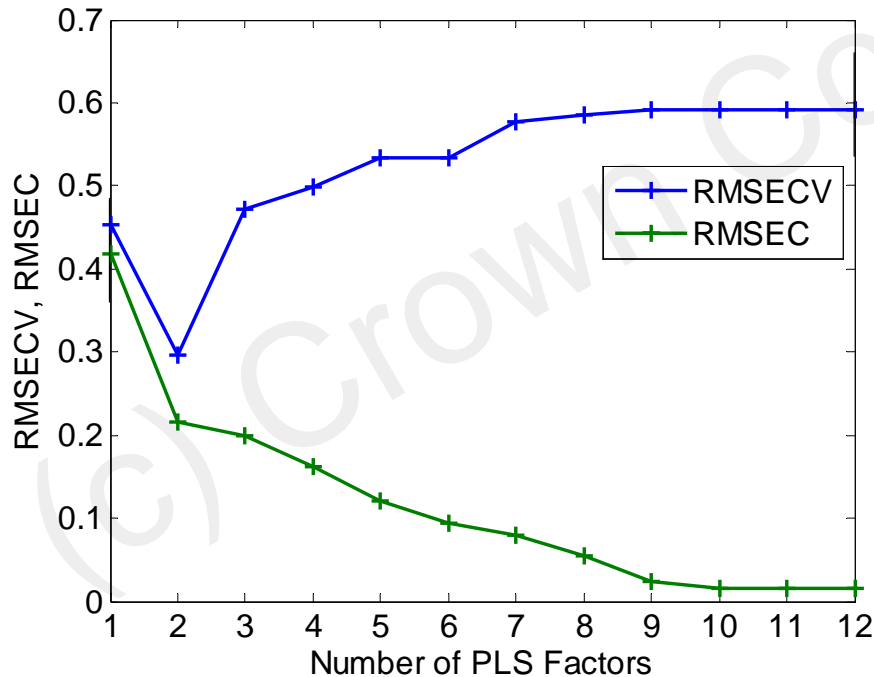


Data is overfitted!



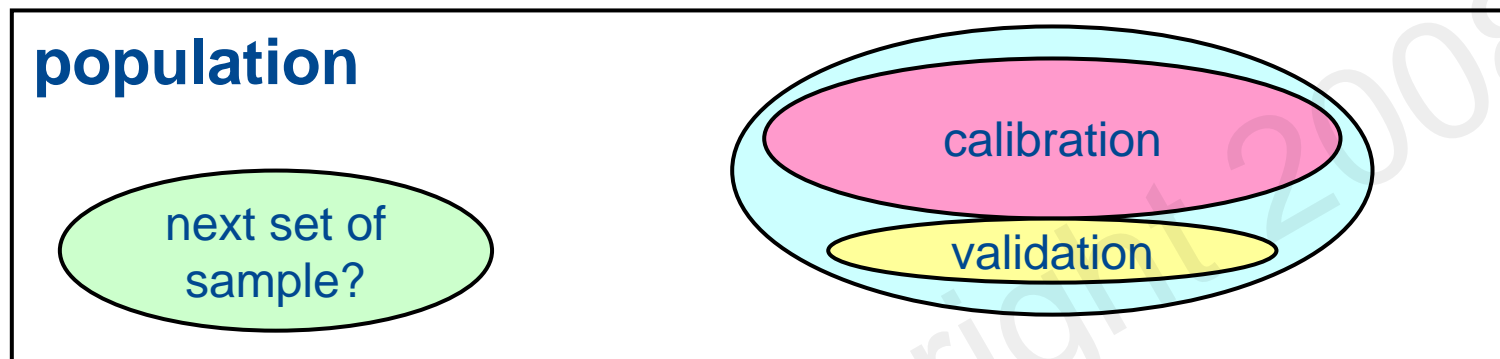
PLS validation

- 'Leave one out' cross validation most popular
 - Calculate PLS model excluding sample i
 - Predict sample i and calculate error
 - Repeat for all different samples
 - Calculate root mean square error of cross validation (RMSECV)



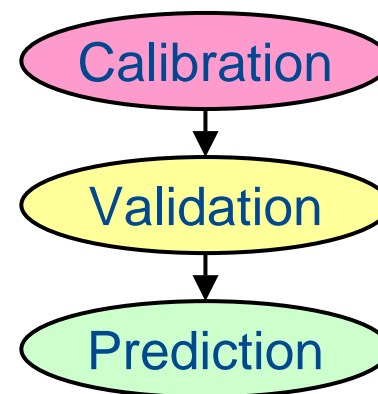
RMSEC (Root Mean Square Error of Calibration) goes down with increasing number of factors

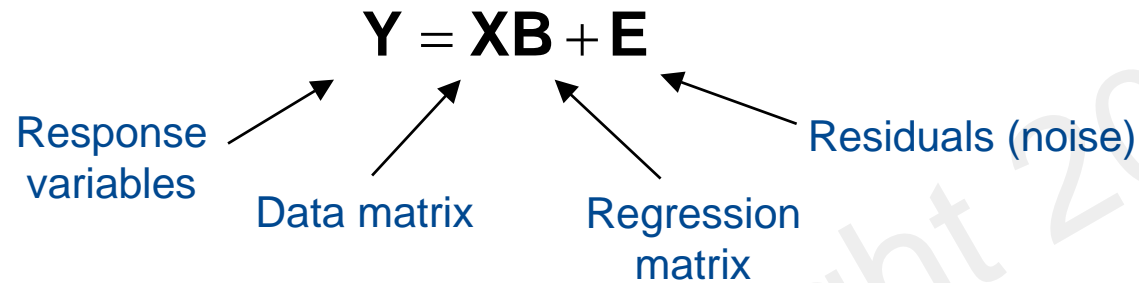
To decide optimal number of factors use **minimum of RMSECV** (Root Mean Square Error of Cross Validation) or PRESS (Prediction Residual Sum of Squares)



- If dataset is large enough, split into calibration and validation sets
- Rule of thumb – 2/3 calibration set, 1/3 validation set
- Validation data should be statistically independent from calibration data i.e. NOT repeat spectra of same sample!

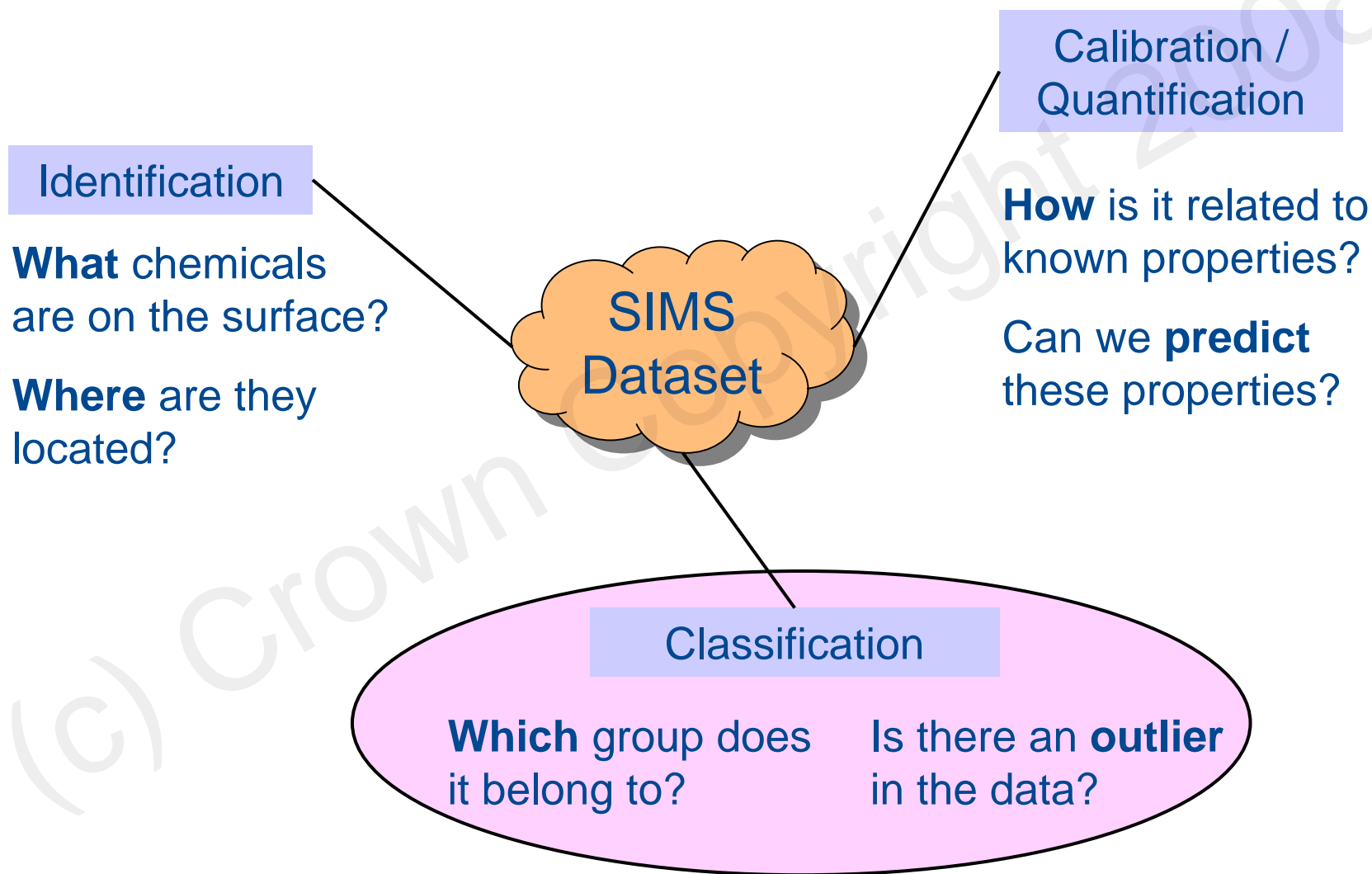
Independent validation set is **essential** if we want to use model to predict new samples!





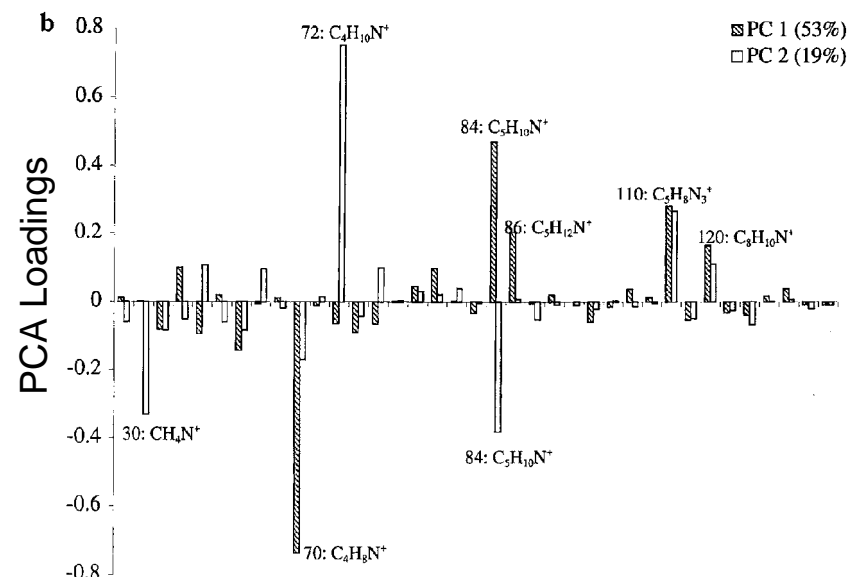
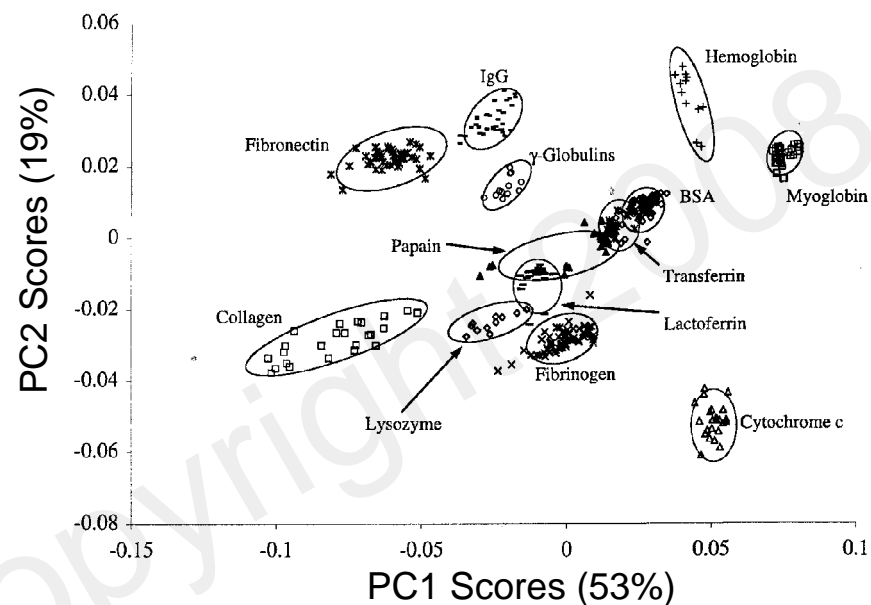
- PLS is a **multivariate linear regression** technique
- PLS decomposes matrices \mathbf{X} (predictors) and \mathbf{Y} (responses) **simultaneously**, in order to find factors that best describe the structure of **covariance** between \mathbf{X} and \mathbf{Y}
- **Data preprocessing** method needs to be selected with care
- PLS is excellent for **calibration** and **quantification**, and for studying the relationship between SIMS data and other measured properties
- Properly **validated** PLS models can be used for **predictions** of these properties using SIMS data

1. Introduction
2. Linear algebra
3. Factor analysis
4. Multivariate regression
5. Classification
 - **PCA classification**
 - **PC-DFA**
 - **PLS-DA**
- Conclusion



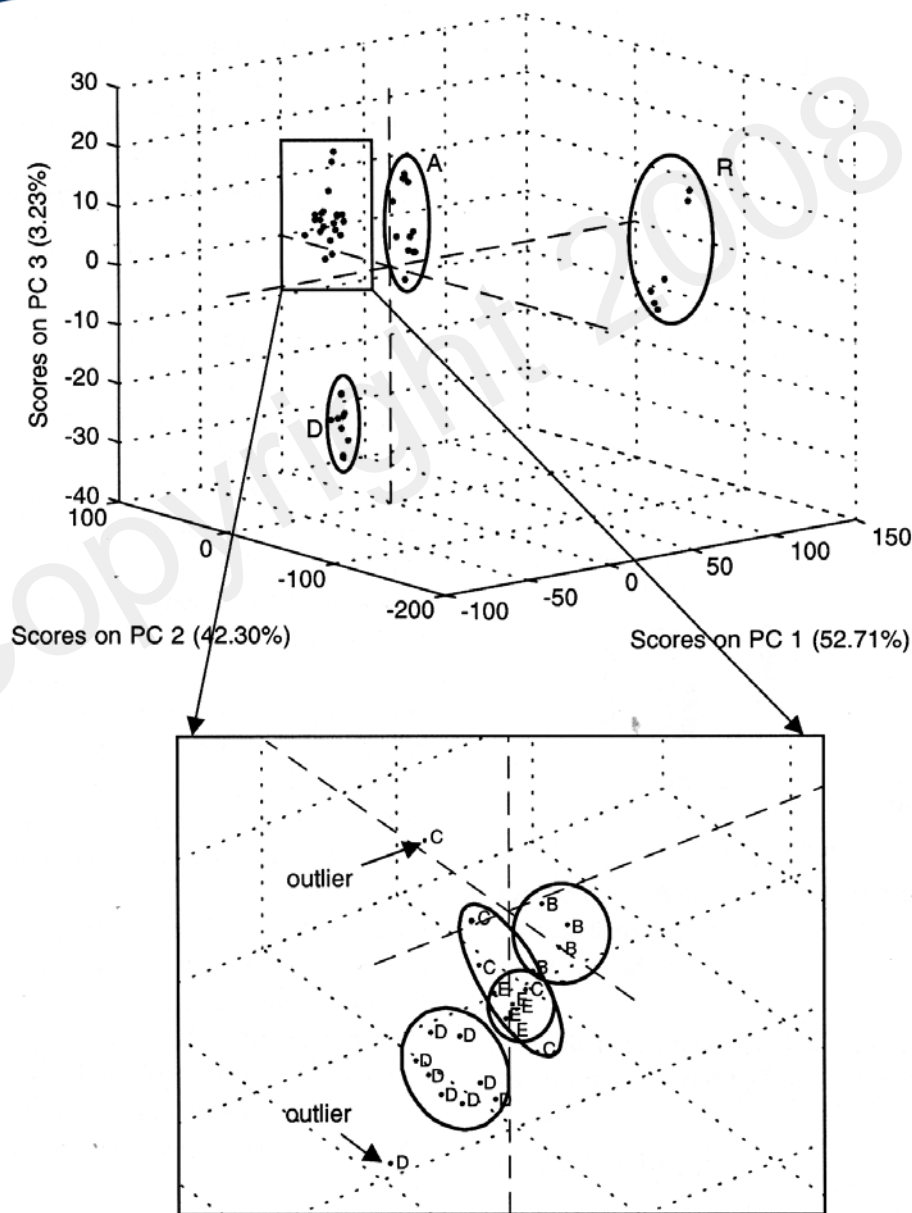
PCA classification (1)

- 16 different single protein films adsorbed on mica
- Excellent classification of proteins using only 2 factors
- Factors consistent with total amino acid composition of various proteins
- 95% confidence limits provide means for identification / classification



PCA classification (2)

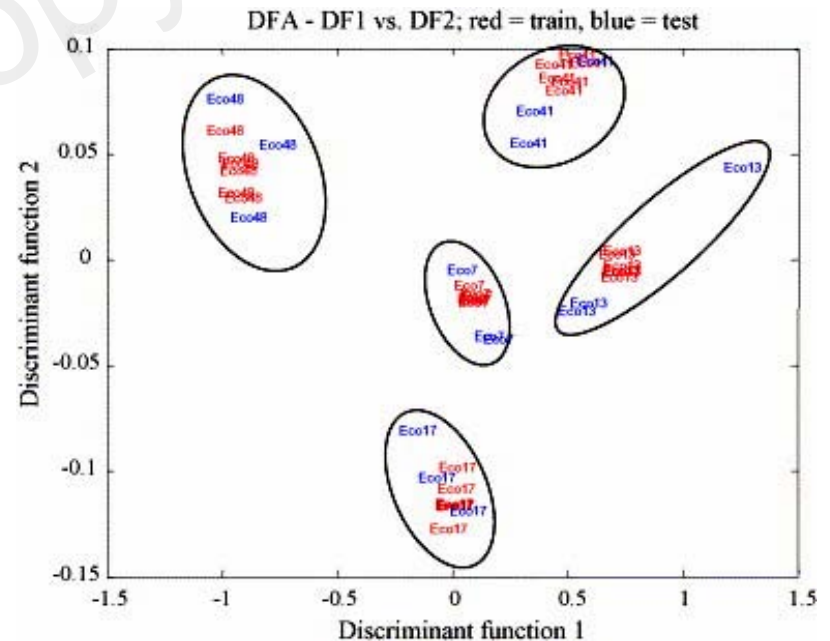
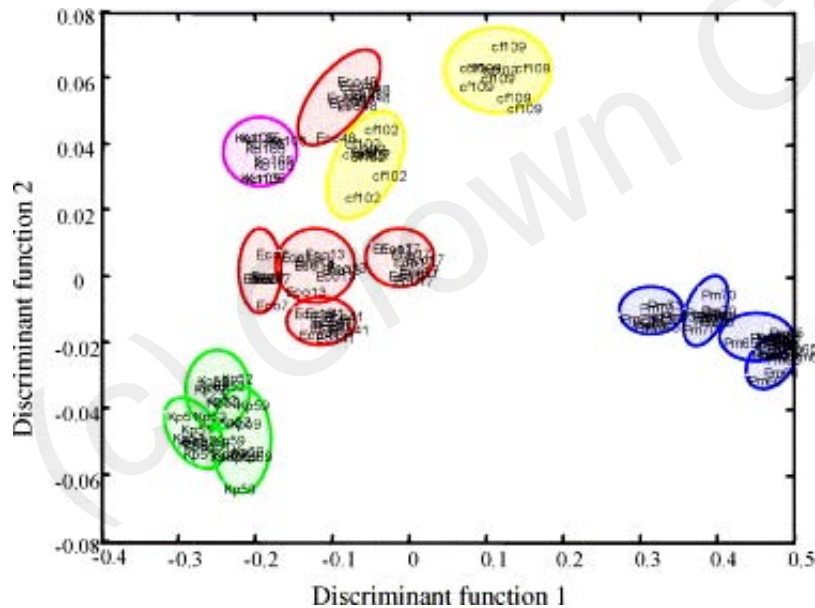
- Octadecanethiol self-assembled monolayers on gold substrates, exposed to different allylamine plasma deposition times
- Four clusters of objects are observed when the scores are on different PCA factors are plotted
- Magnification of framed cluster reveals further clustering
- Outliers can also be located

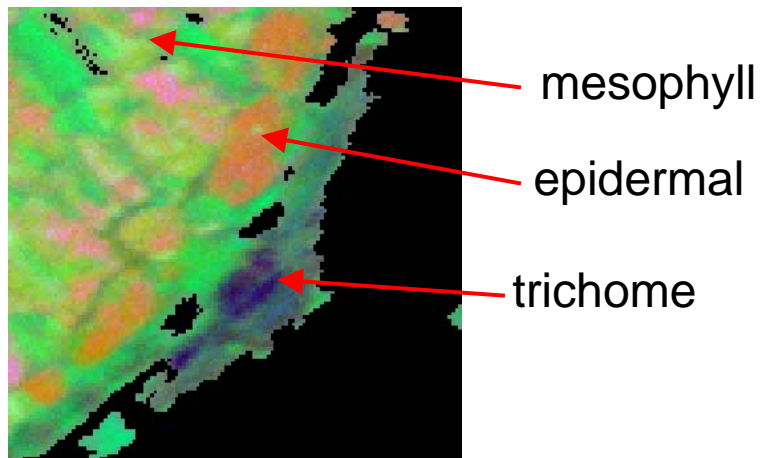


- PC-DFA = “Principal Component – Discriminant Function Analysis”
- ‘Discriminant functions’ maximizes the Fisher’s ratio between groups

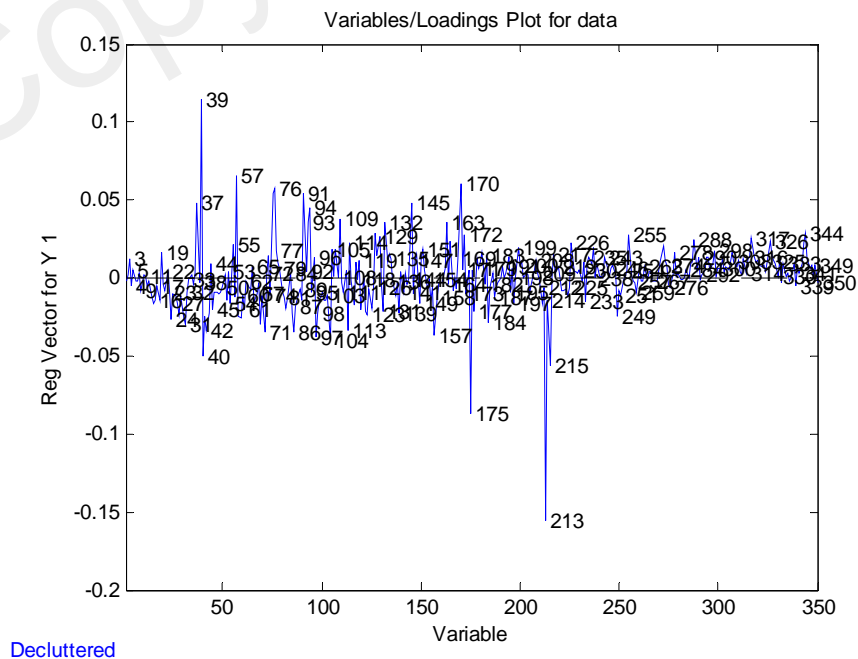
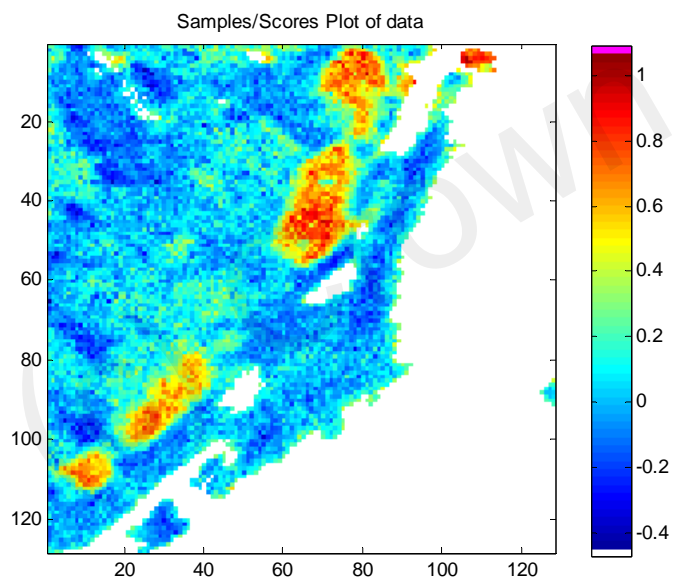
$$\text{Fisher's ratio} = \frac{(\text{mean}_1 - \text{mean}_2)^2}{\text{var}_1 + \text{var}_2}$$

- Used to distinguish strains of bacteria



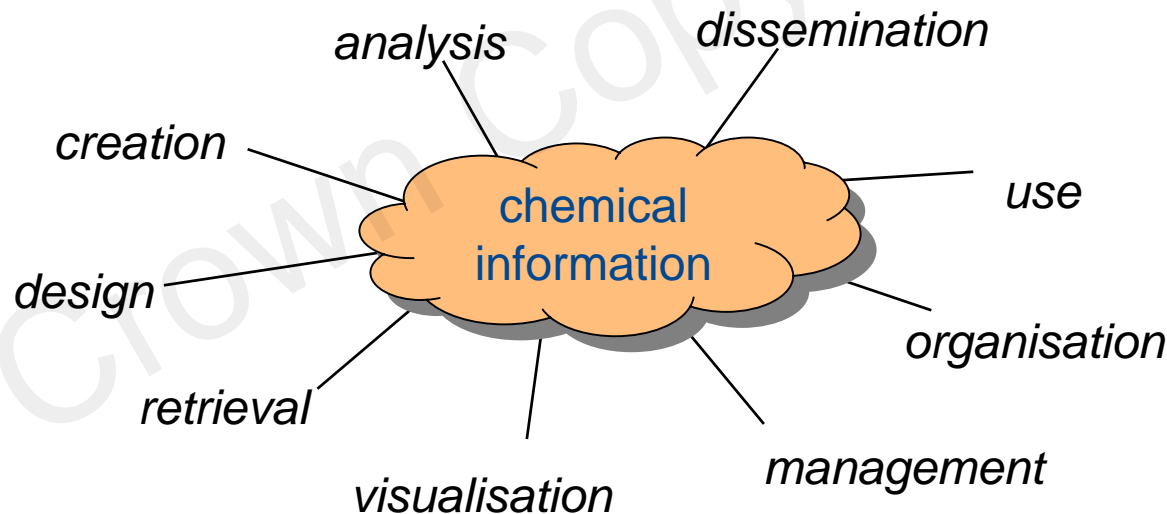


- Partial Least Squares Discriminant Analysis
- PLS finds factors that describes the biggest co-variance between the data X and the group assignments (e.g. 0 and 1) Y .
- Regression vector shows linear combination of peaks that maximally distinguishes epidermal and other cells

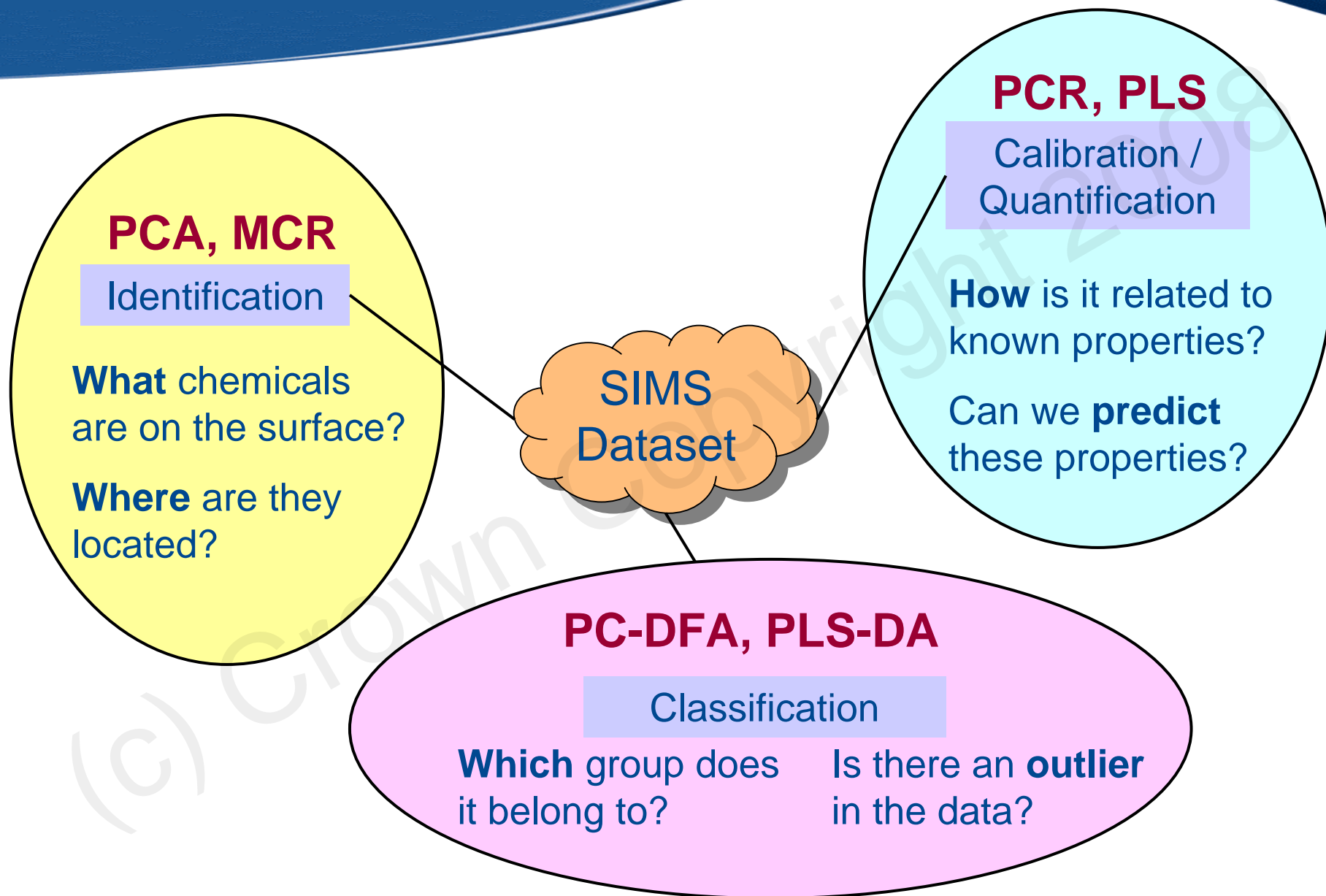


Classification summary

- PCA allows for quick grouping of samples based on their similarities
- PC-DFA and PLS-DA are **supervised** classification methods – prior knowledge about groups are required
- Properly **validated** classification models are needed for **predictions**
- There also exists **unsupervised clustering** methods, e.g. hierarchal cluster analysis, K-nearest-neighbours, artificial neural networks.....

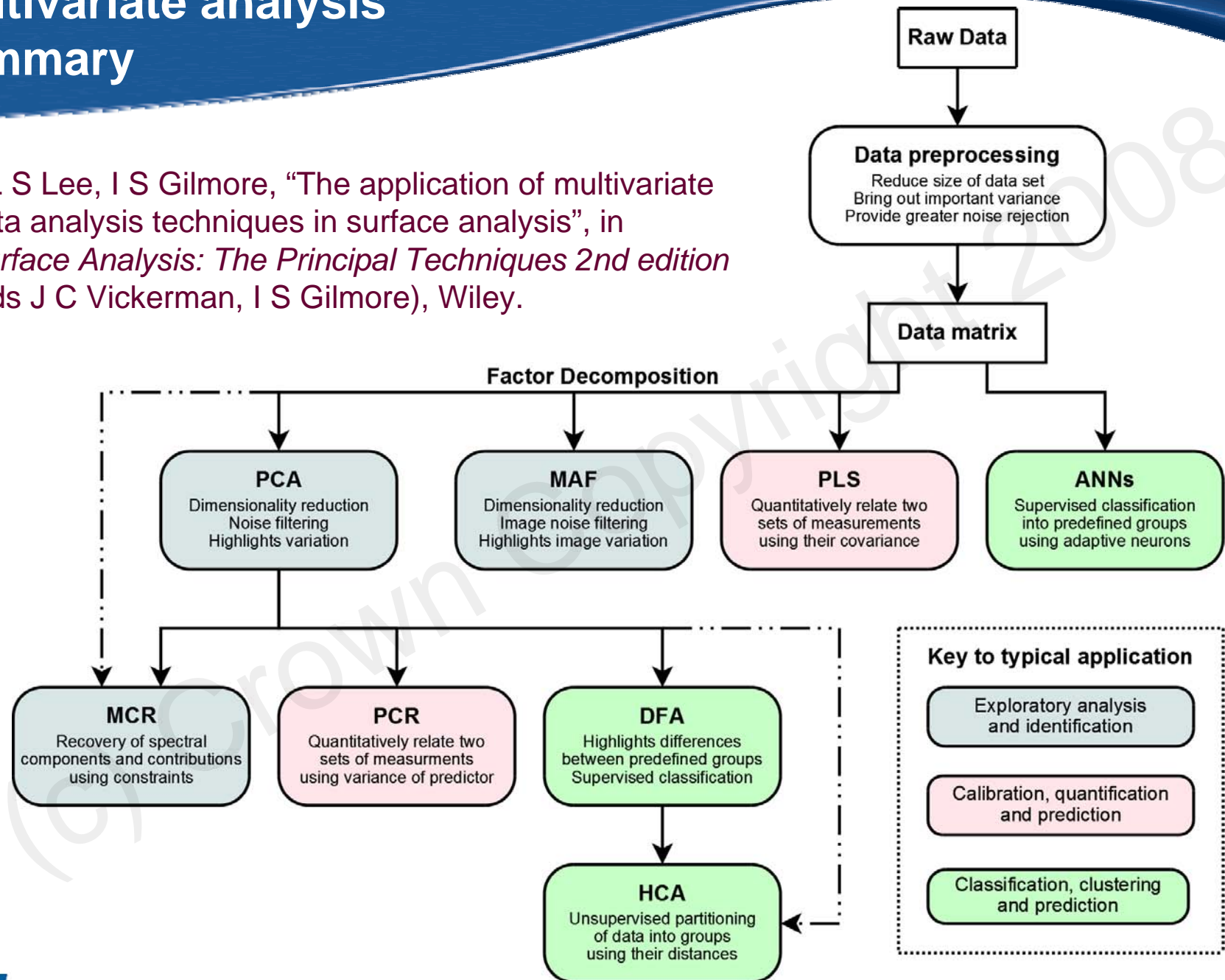


All these (and much, much more) belong to the wider field of **chemoinformatics!**



Multivariate analysis summary

J L S Lee, I S Gilmore, "The application of multivariate data analysis techniques in surface analysis", in *Surface Analysis: The Principal Techniques 2nd edition* (eds J C Vickerman, I S Gilmore), Wiley.



Conclusion

In this tutorial we have looked at

- Identification using PCA and MCR
- Quantification using MLR, PCR and PLS
- Classification using PC-DFA, PLS-DA
- Importance of validation for predictive models
- Data preprocessing techniques and their effects
- Matrix and vector algebra
- Newly defined multivariate analysis terminology

Terms Here	Symbol	Definition	PCA	MCR	PLS
Factor	-	An axis in the data space representing an underlying dimension that contributes to summarising or accounting for the original data set	Principal Component	Pure Component	Latent Vectors, Latent Variables
Loadings	P	Correlation between the original variables and the factors	Loadings, Eigenvector	Component Spectrum	Loadings
Scores	T	Projection of the samples onto the factors	Scores, Projections	Component Concentration	Scores

Bibliography

General

- J. L. S. Lee *et al*, "The application of multivariate data analysis techniques in surface analysis", in *Surface Analysis: The Principal Techniques 2nd edition* (eds J C Vickerman, I S Gilmore), Wiley.
- S. Wold, *Chemometrics; what do we mean with it, and what do we want from it?*, Chemom. Intell. Lab. Syst. **30** (1995) 109
- E. R. Malinowski, *Factor analysis in Chemistry*, John Wiley and Sons (2002)
- P. Geladi *et al*, *Multivariate image analysis*, John Wiley and Sons (1996)
- J. L. S. Lee *et al*, *Quantification and methodology issues in multivariate analysis of ToF-SIMS data for mixed organic systems*, Surf. Interface Anal. **40** (2008) 1
- D. J. Graham, *NESAC/BIO ToF-SIMS MVA web resource*, <http://nb.engr.washington.edu/nb-sims-resource/>

PCA

- D. J. Graham *et al*, *Information from complexity: challenges of ToF-SIMS data interpretation*, Appl. Surf. Sci. **252** (2006) 6860
- M. R. Keenan *et al*, *Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images*, Surf. Interface Anal. **36** (2004) 203
- M. R. Keenan *et al*, *Mitigating dead-time effects during multivariate analysis of ToF-SIMS spectral images*, Surf. Interface Anal. **40** (2008) 97

MCR

- N. B. Gallagher *et al*, *Curve resolution for multivariate images with applications to TOF-SIMS and Raman*, Chemom. Intell. Lab. Syst. **73** (2004) 105
- J. A. Ohlhausen *et al*, *Multivariate statistical analysis of time-of-flight secondary ion mass spectrometry using AXSIA*, Appl. Surf. Sci. **231-232** (2004) 230
- R. Tauler, A. de Juan, *MCR-ALS Graphic User Friendly Interface*, <http://www.ub.edu/mcr/>

PLS

- P. Geladi *et al*, *Partial Least-Squares Regression: A Tutorial*, Analytica Chimica Acta **185** (1986) 1
- A. M. C. Davies *et al*, *Back to basics: observing PLS*, Spectroscopy Europe **17** (2005) 28

Acknowledgements

The work is supported by UK Department of Innovation, Universities and Skills (DIUS)'s Chemical and Biological Metrology Programme

Department for
**Innovation,
Universities &
Skills**

We would like to thank Dr Ian Fletcher (Intertek MSG) and Prof Chris Grovenor (University of Oxford) for images, and Dr Martin Seah (NPL) for helpful comments

Intertek

For further information of Surface and Nanoanalysis at NPL please visit
<http://www.npl.co.uk/nanoanalysis>

