Quadram Institute Science Health Food Innovation

Multivariate analysis in analytical chemistry

("Chemometrics")

Prof E K Kemsley

Head of Core Science Resources

Quadram Institute Bioscience

Norwich Research Park

kate.kemsley@quadram.ac.uk

In this talk

What is multivariate data?

Why take a multivariate approach to data analysis?

Introducing ‘Chemometrics’

Avoiding pitfalls – how the incorrect use of multivariate methods can lead to nonsense results

What is multivariate data?

A simple (non-chemical) example: variables that might influence a child's weight

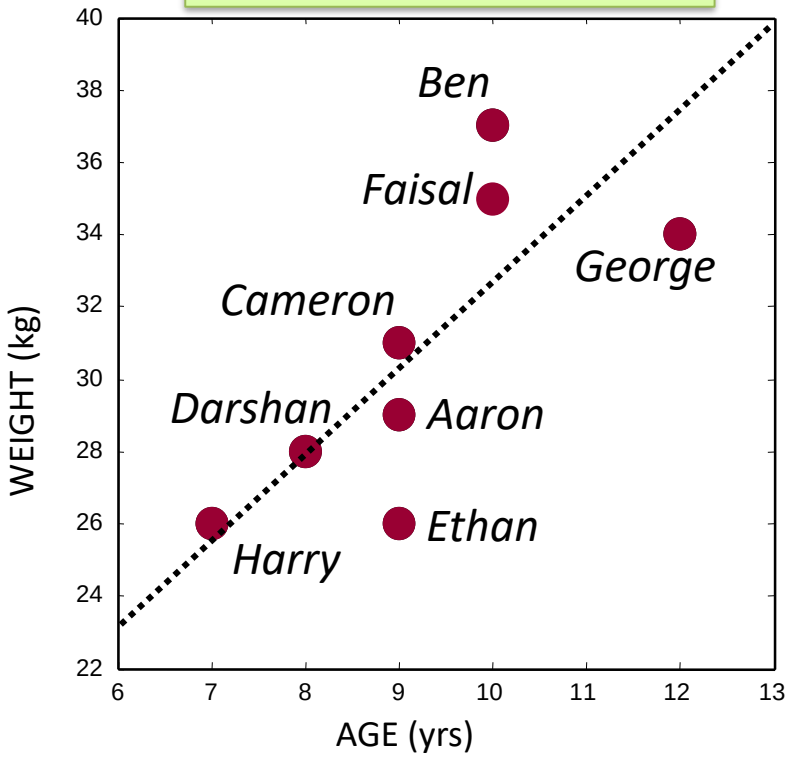
	Weight (kg) ***('dependent variable')***	Age (years) ***('predictor #1')***	Height (cm) ***('predictor #2')***
Aaron	29	9	146
Ben	35	10	138
Cameron	31	9	143
Darshan	28	8	127
Ethan	26	9	141
Faisal	37	10	132
George	34	12	155
Harry	26	7	128

n = 8 observations

How could we analyse this data?

- Dependent variable vs. each predictor:

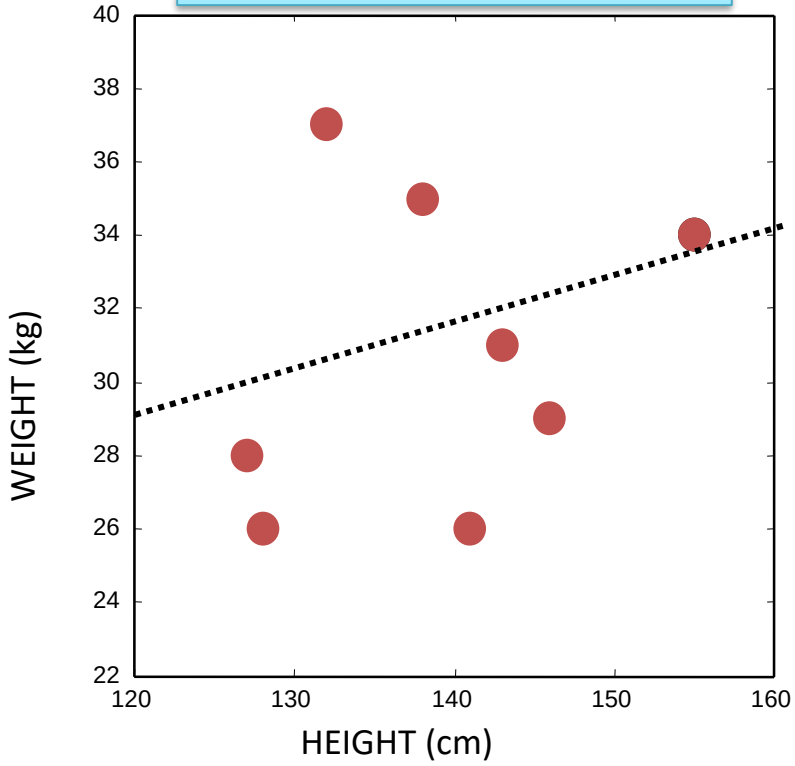
WEIGHT vs. AGE



$$y = m x_1 + c$$

WEIGHT = slope . AGE + intercept

WEIGHT vs. HEIGHT



$$y = m x_2 + c$$

WEIGHT = slope . HEIGHT + intercept

What about a multivariate approach?

- Make use of both predictor variates:

$$WEIGHT = m_1 . HEIGHT + m_2 . AGE + c$$
$$y = m_1 . x_1 + m_2 . x_2 + c$$

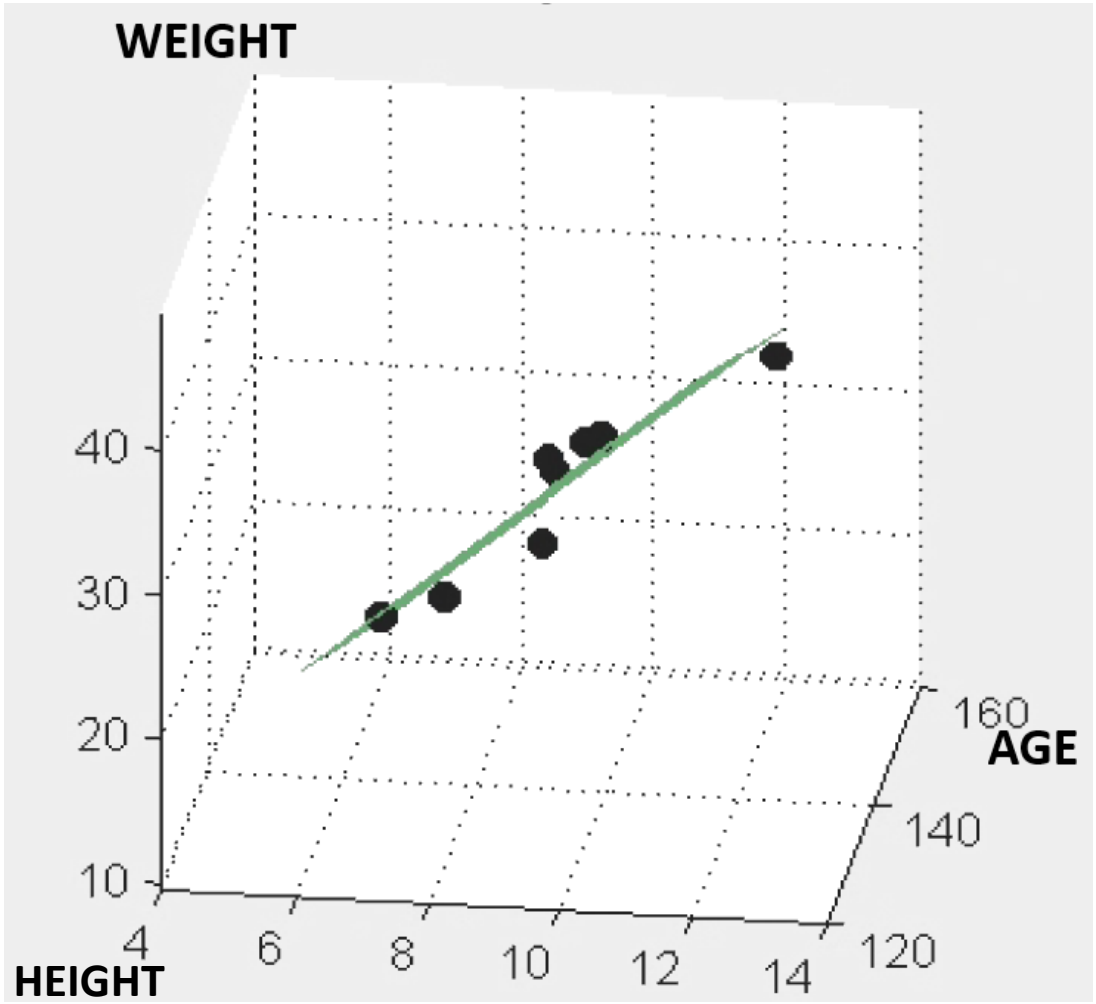
- Regression no longer describes a line, but a **plane** in a 3-d coordinate system

- Before widespread use of computers in the last 20 years, this was laborious - even for datasets of this size

- Multivariate analysis of ‘big data’ is still an evolving field

What about a multivariate approach?

- Make use of both predictor variates:

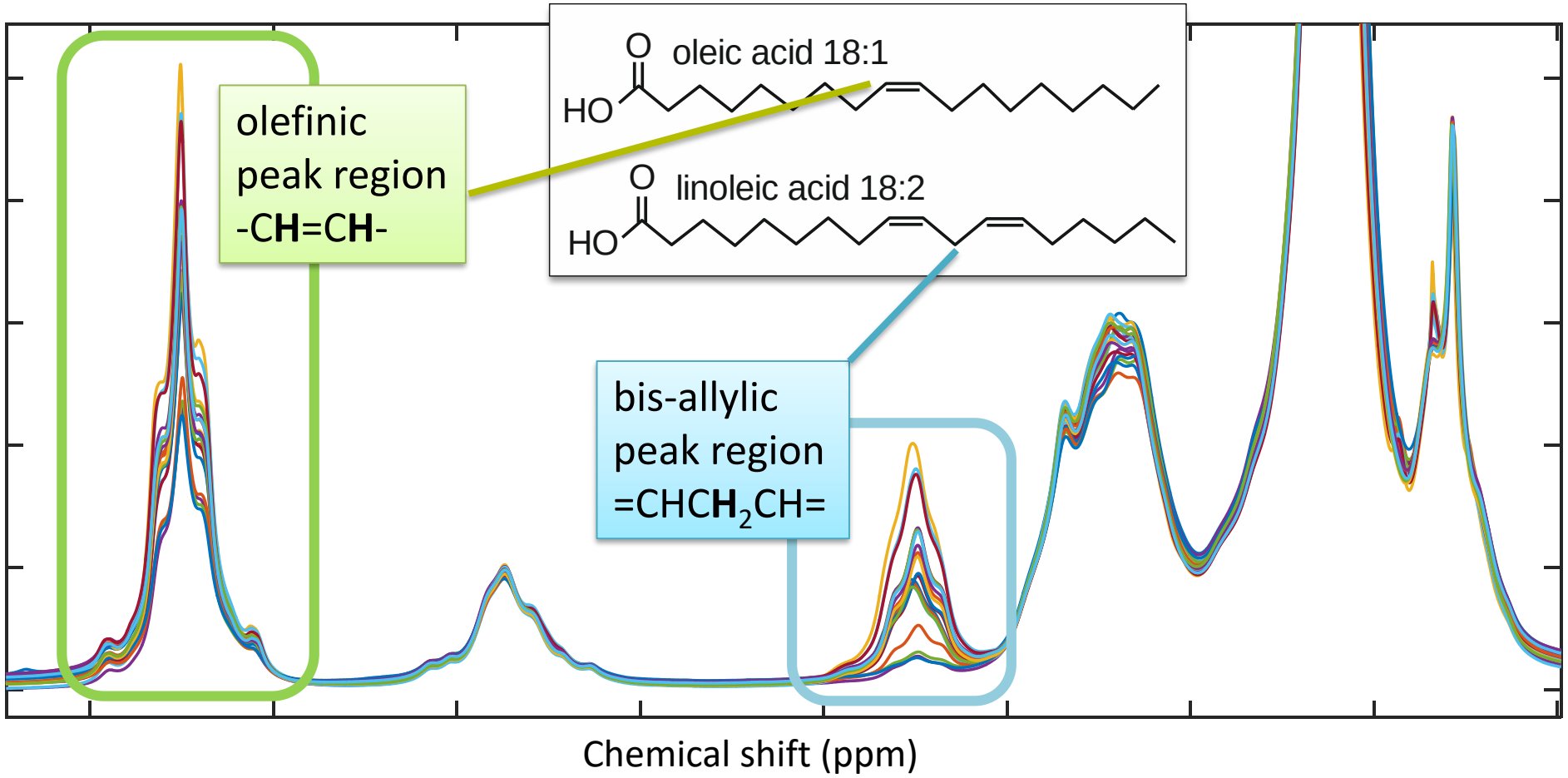


An example from chemistry…

- **60MHz proton NMR spectra of edible oils**
  (sunflower, corn, olive, rapeseed, sesame, walnut,….)
- Why? Compositional information for labels

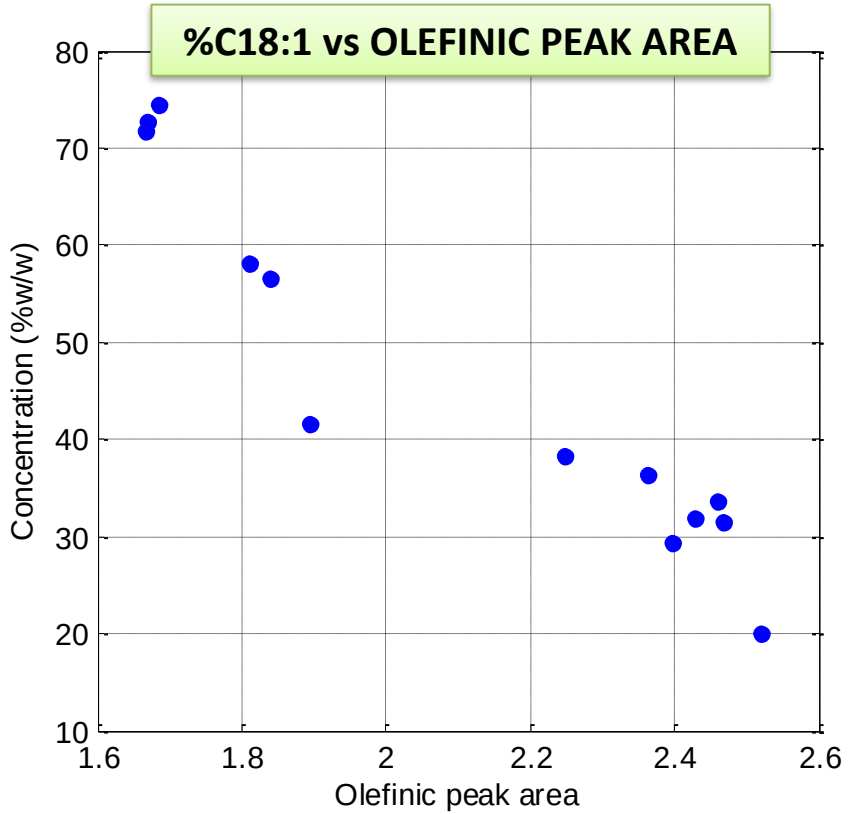
An example from chemistry…

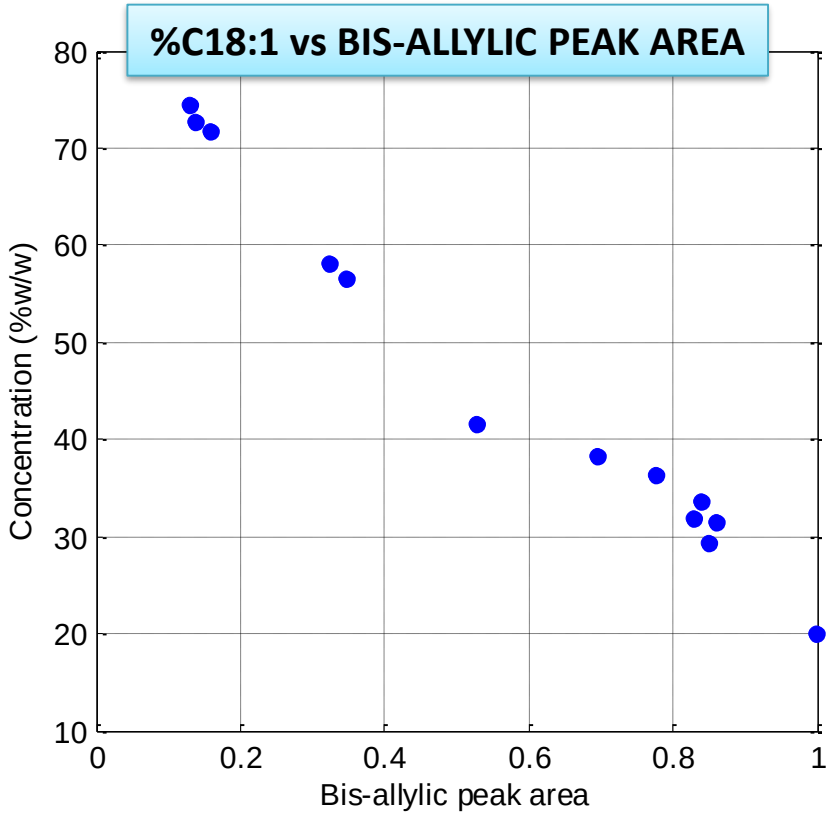
- **60MHz proton NMR spectra of edible oils**
  (sunflower, corn, olive, rapeseed, sesame, walnut,….)



An example from chemistry…

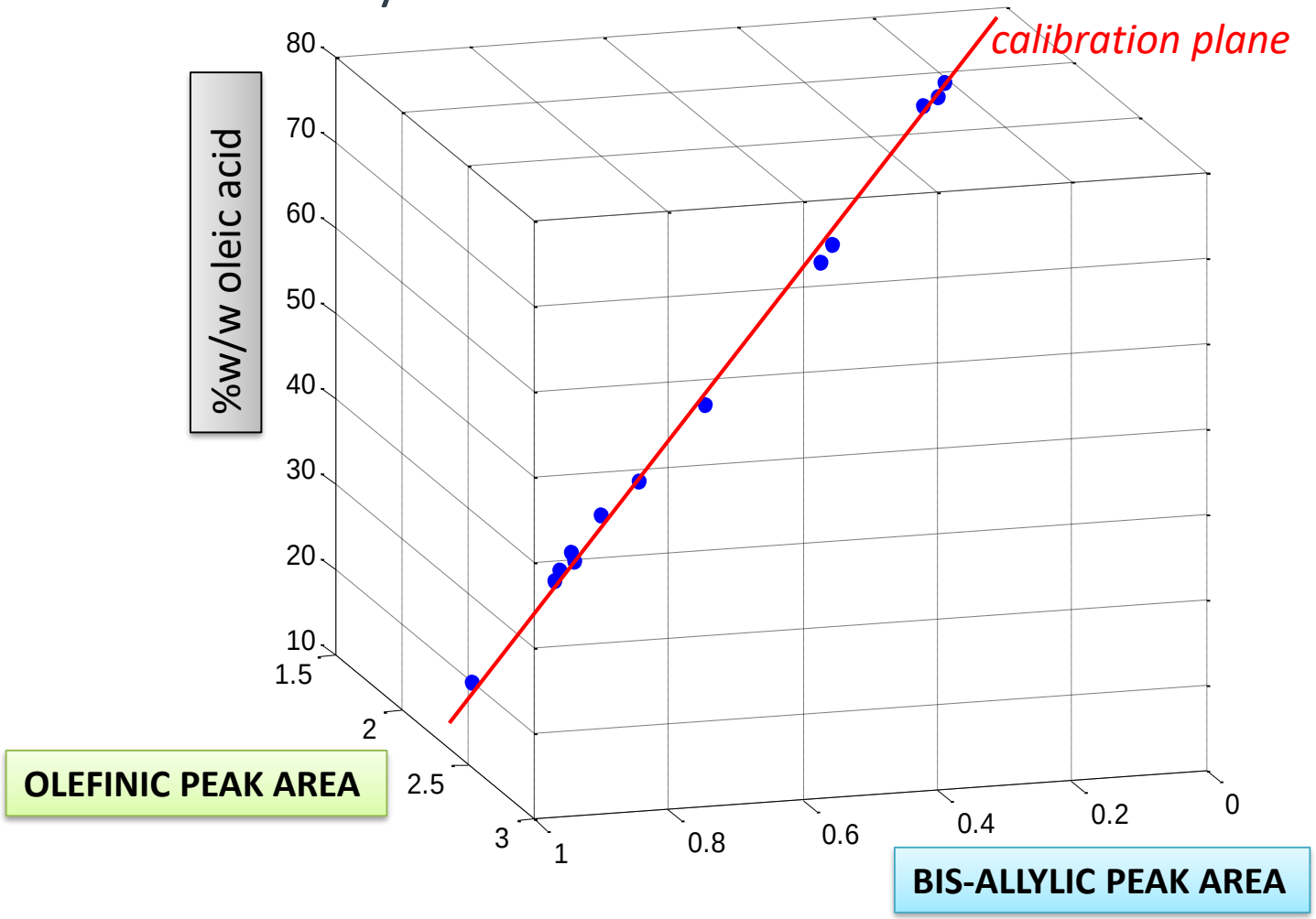
- The oleic acid content (C18:1) was also measured by a GC-MS reference method
- How could we obtain a calibration of GC-MS %w/w versus NMR peak areas for analysing future samples?



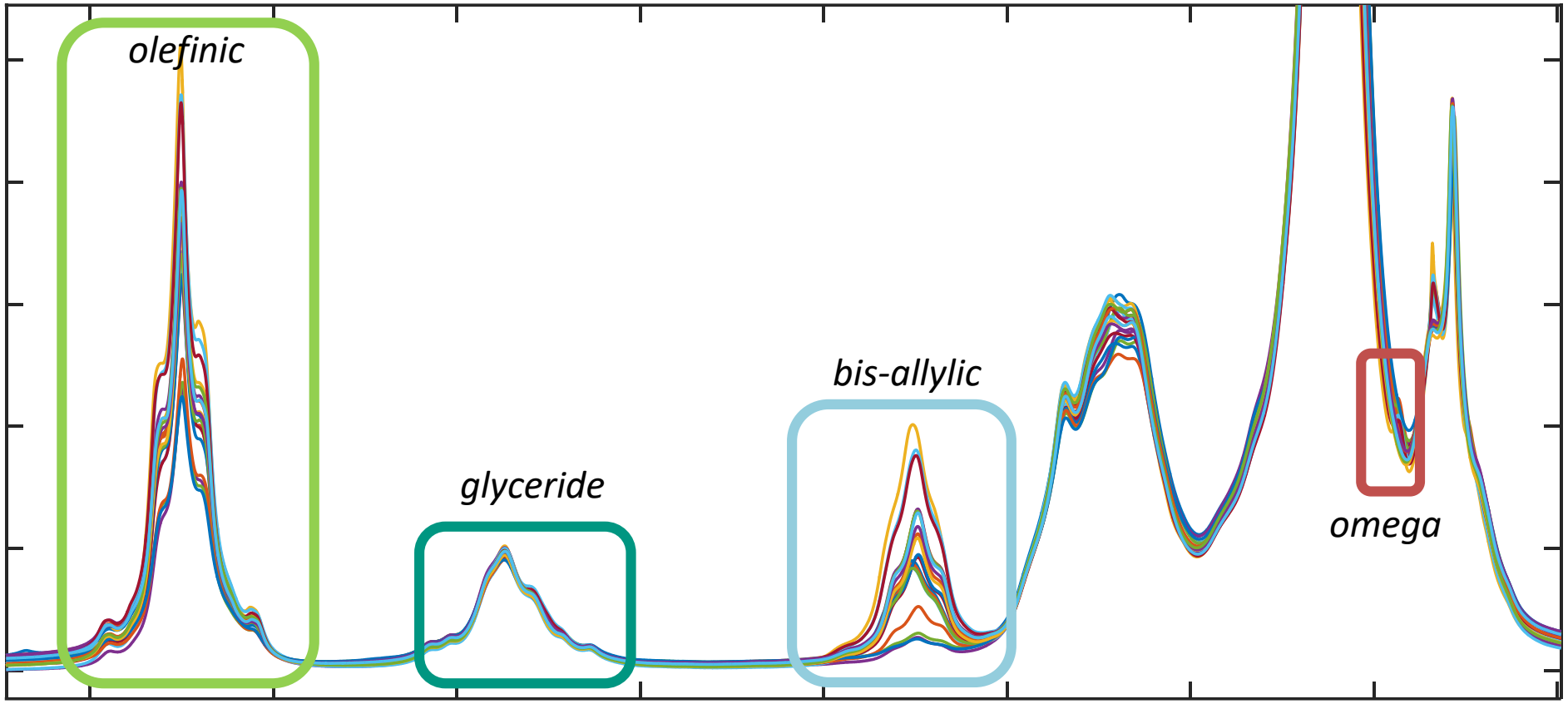


An example from chemistry…

- …but we could also plot %w/w oleic acid versus both peak areas simultaneously:



An example from chemistry…



- 4 peak areas plus the %w/w data to deal with
- **Can no longer plot the complete data easily on ordinary axes**
- **Need to leave behind “2-d thinking”**

Strategies for handling multivariate data

Focus on outcome: 'Actual' versus 'Predicted by regression'

- Multivariate regression of GC-MS %w/w C18:1 onto [olefinic, glyceride, bis-allylic, omega] NMR peak areas:

$$MONO\ \%w/w = m_1 . OLEFINIC + m_2 . BIS\text{-}ALLYLIC + m_3 . GLYCERIDE + m_4 . OMEGA + c$$

$$y = m_1 . x_1 + m_2 . x_2 + m_3 . x_3 + m_4 . x_4 + c$$

$$\underset{(n \times 1)}{\mathbf{y}} = \underset{(n \times 4)(4 \times 1)}{\mathbf{X}\ \mathbf{m}}$$

(vector & matrix notation; to remove the need for c, **X** *is mean-centered)*

$$\hat{\mathbf{m}} = (\mathbf{X}^T\mathbf{X})^{-1}\ \mathbf{X}^T\ \mathbf{y}$$

Least squares estimate of m

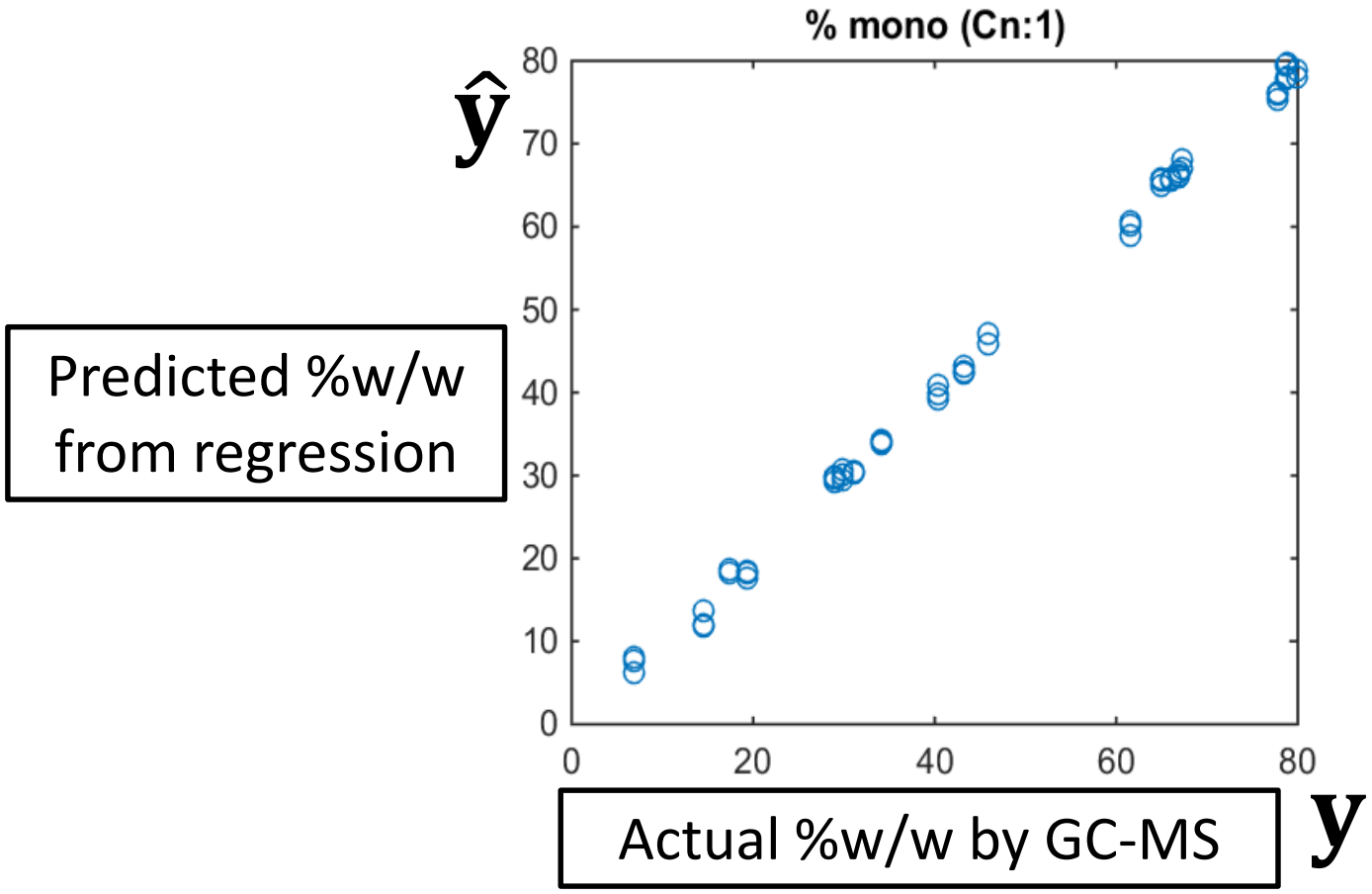
$$\hat{\mathbf{y}} = \mathbf{X}\ \hat{\mathbf{m}}$$

Values of y predicted by regression

Strategies for handling multivariate data

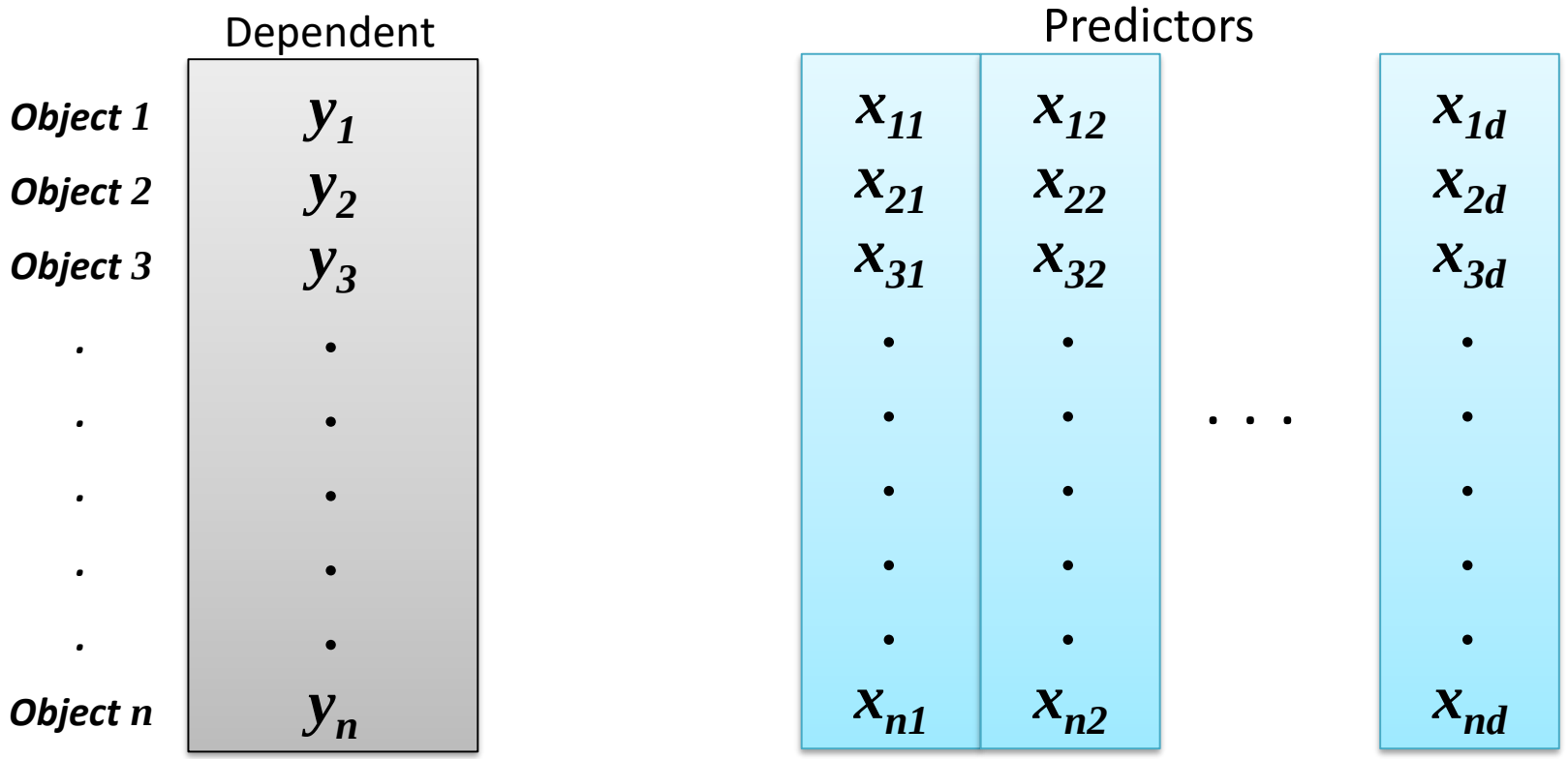
Focus on outcome: 'Actual' versus 'Predicted by regression'

- Multivariate regression of GC-MS %w/w C18:1 onto [olefinic, glyceride, bis-allylic, omega] NMR peak areas:



Strategies for handling multivariate data

Viewing a large, multivariate dataset as a matrix:



This could be something like:

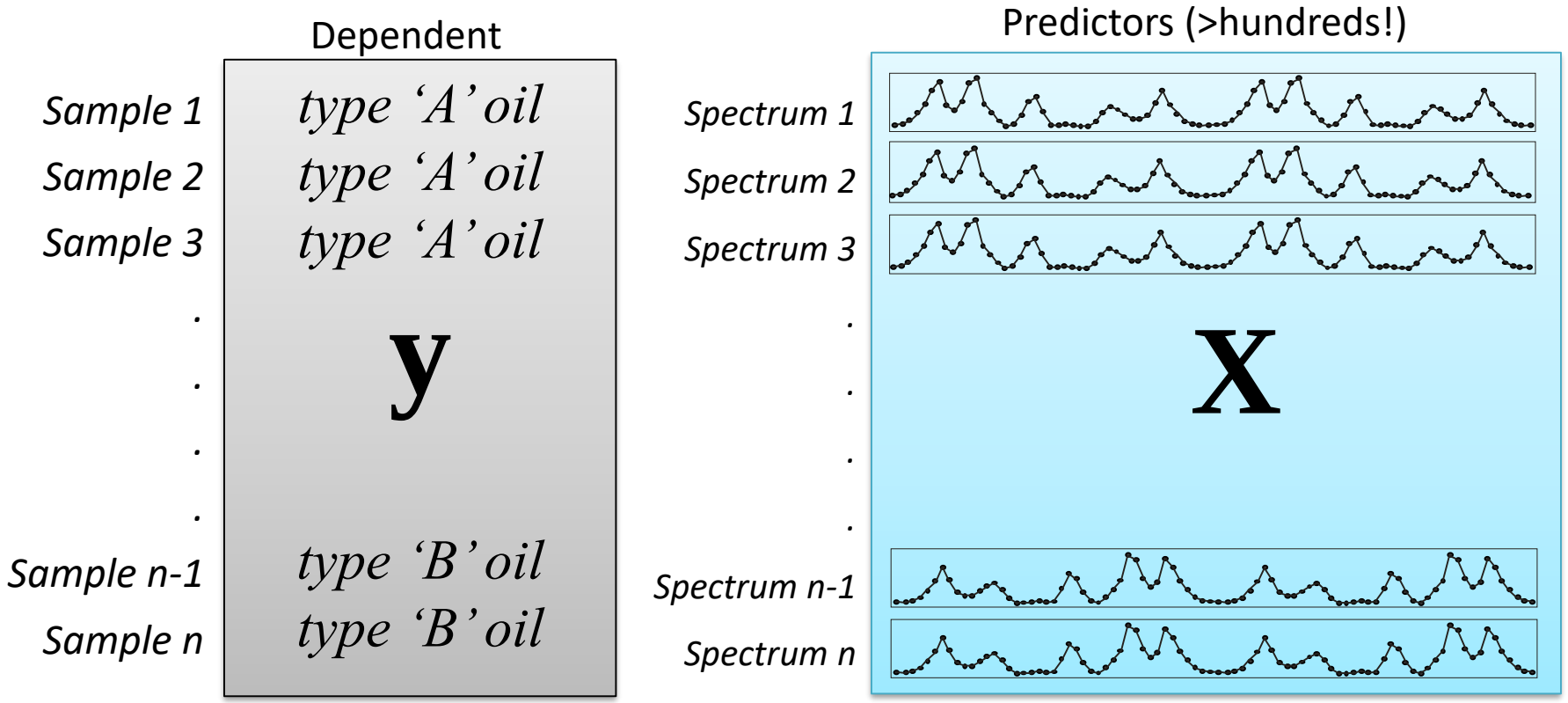
- Concentration of some chemical component
- A category e.g. species, variety,,,,

A set of *d* attributes of the objects, e.g.

- Spectral intensity values (each row in the data matrix = a spectrum)

Strategies for handling multivariate data

Viewing a large, multivariate dataset as a matrix:



This could be something like:

- Concentration of some chemical component
- A category e.g. species, variety,,,,

A set of *d* attributes of the objects, e.g.

- Spectral intensity values (each row in the data matrix = a spectrum)

Strategies for handling multivariate data

Regression with very large data matrices

$$\mathbf{y} = \mathbf{X}\,\mathbf{m}$$

(n x 1) *(n x d)(d x 1)*

d may be very large!

$$\hat{\mathbf{m}} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{y}$$

- If d > n (more *attributes per object than total number of objects*), then ($\mathbf{X}^T\,\mathbf{X}$) is "singular" and not invertible
- This leads to mathematical problems in calculating $\hat{\mathbf{m}}$ directly
- **"Chemometrics"** methods were developed to overcome this problem

Introducing Chemometrics

- A family of multivariate statistical methods for treating the large datasets of modern analytical chemistry
- Originated in the 1980's when computers first started to become connected to analytical instruments (especially infrared)
- Some chemometric methods were proposed theoretically several decades earlier, but were impossible to carry out

Nicolet 7000 Series FTIR Spectrometer, 1984

Introducing Chemometrics

- Some well-known chemometrics methods:
  - **Principal component analysis (PCA)**
  - **Partial least squares (PLS)**
  - Support Vector Machines
  - Linear Discriminant Analysis
  - ...and many more... (+ lots of synonyms!)

- Same methods have spread throughout the sciences
  - Psychometrics
  - Econometrics
  - Meteorology
  - Bioinformatics,...

Principal Component Analysis

- **Rearranges the information in the data set to make it easier to deal with** (visualise, display, analyse further, etc)

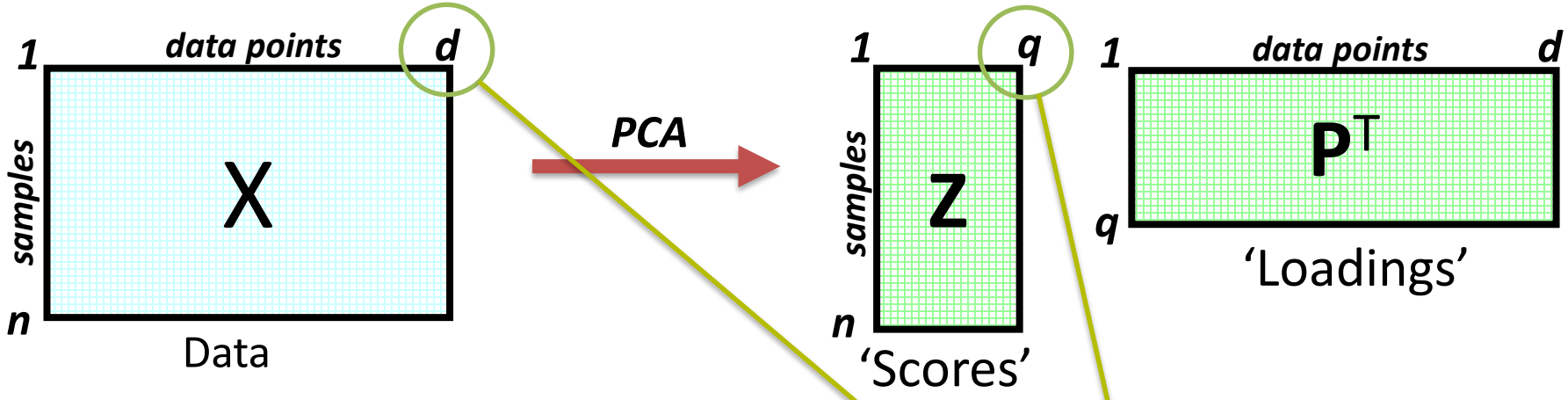
It is a **matrix decomposition method:**

$$\mathbf{X} = \mathbf{Z}.\ \mathbf{P}^{\mathrm{T}}$$

…where $\mathbf{P}$ is a matrix of 'loadings' and $\mathbf{Z}$ a matrix of 'scores'

The loadings are the **eigenvectors** of the data covariance matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X} / (n\text{-}1)$

Principal Component Analysis



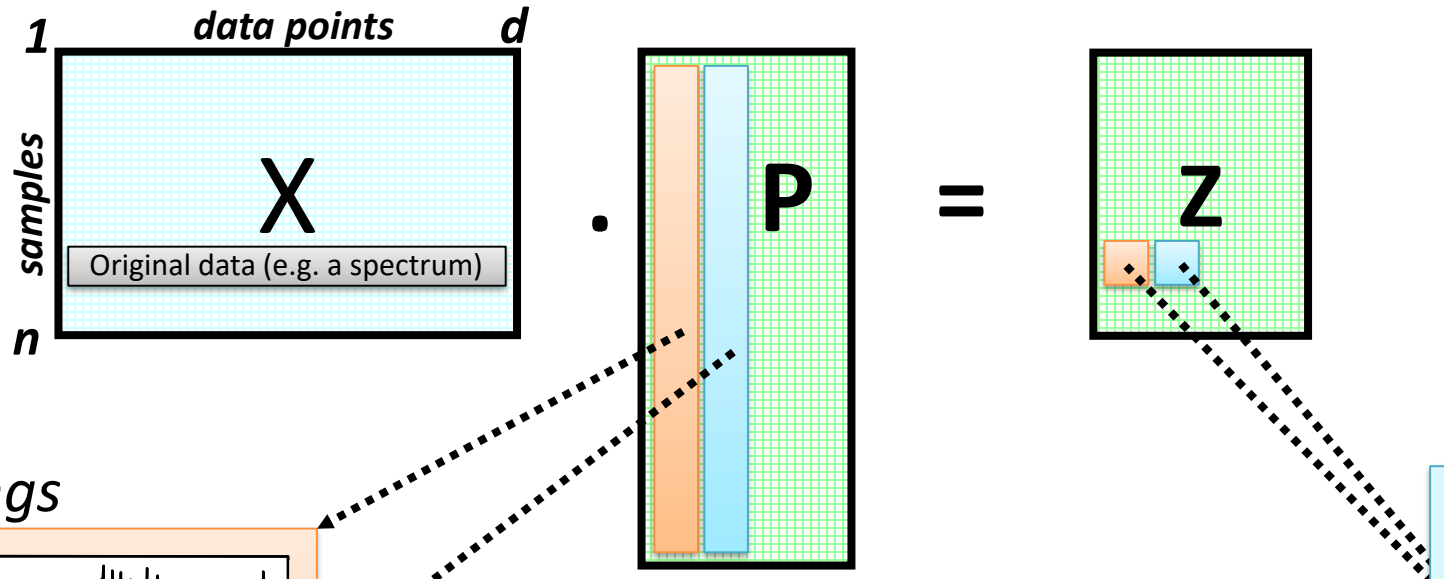
Properties of the Scores:

- *The 'scores' are a set of transformed data*
- *q is smaller than d, so Z is smaller than X – the "redundancy" in the data has been removed*
- *Z is easier to explore than X (e.g by plotting graphically)*

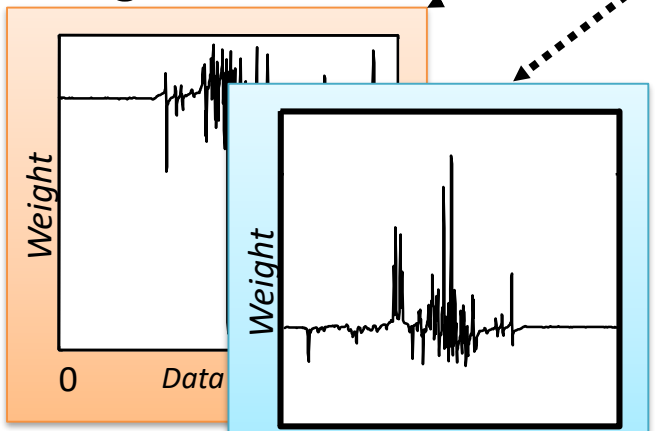
$q << d$

There are q 'principal components'

Principal Component Analysis



Loadings

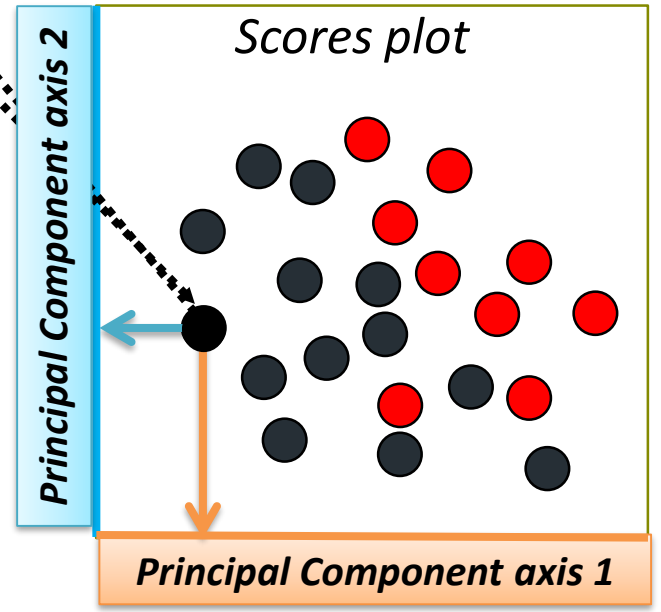


Loadings plots:

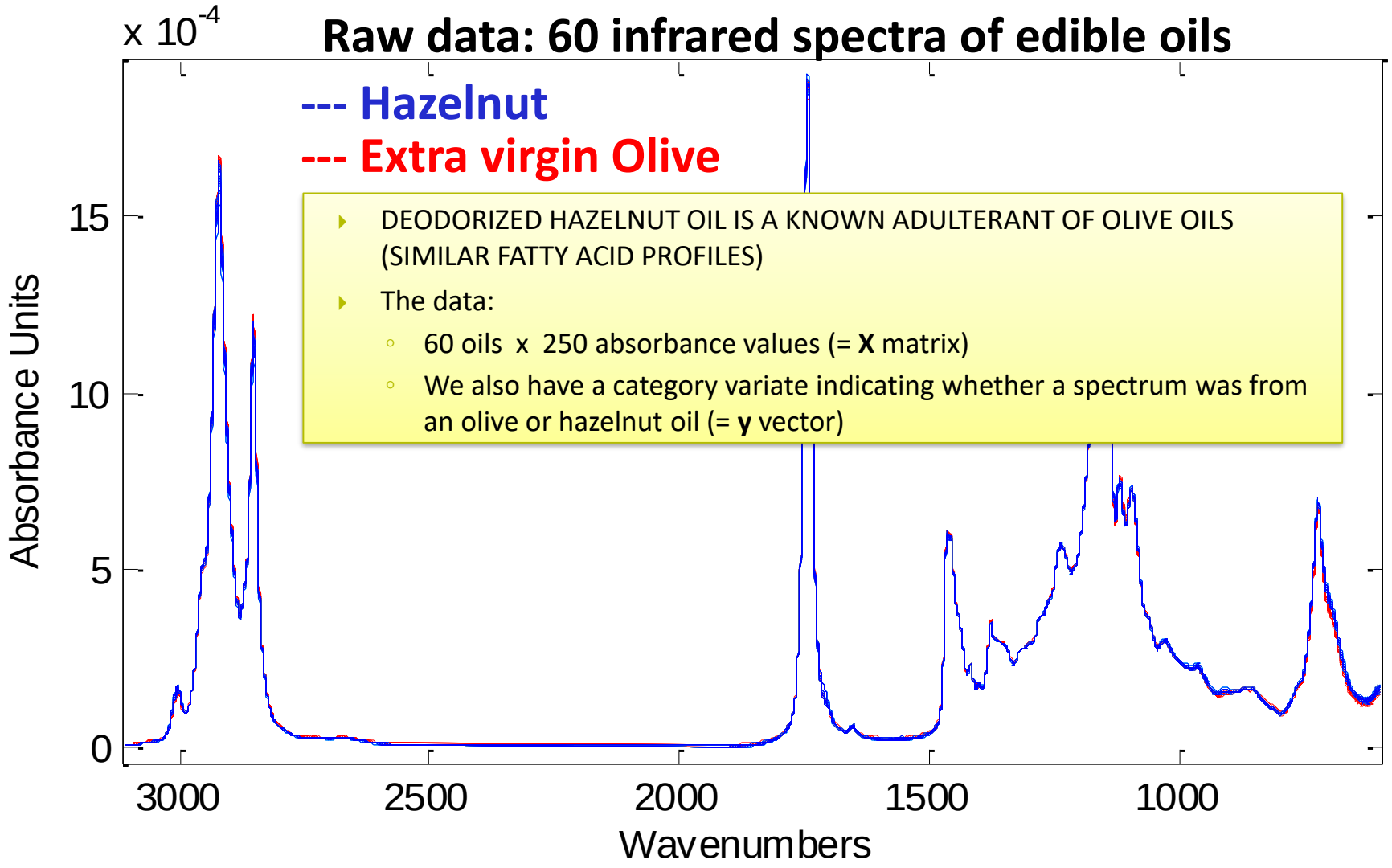
- *Show relative importance of each variate*
- *Information on same scale as original spectra*

Scores plots:

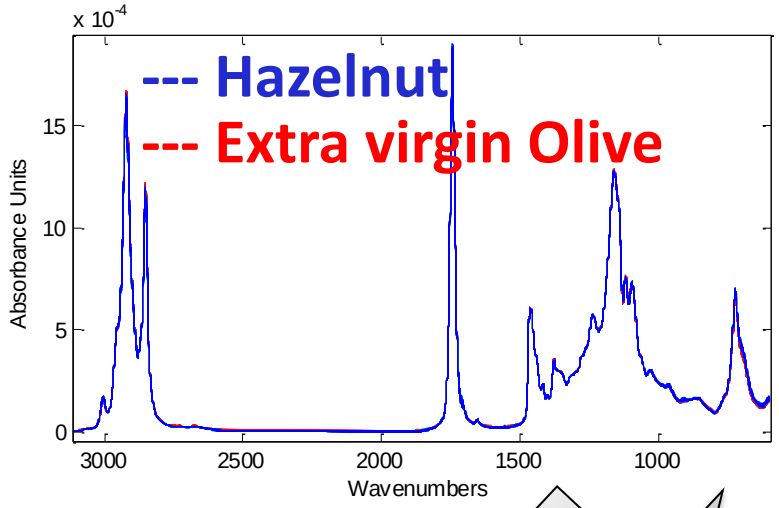
- *Each point represents an individual spectrum*



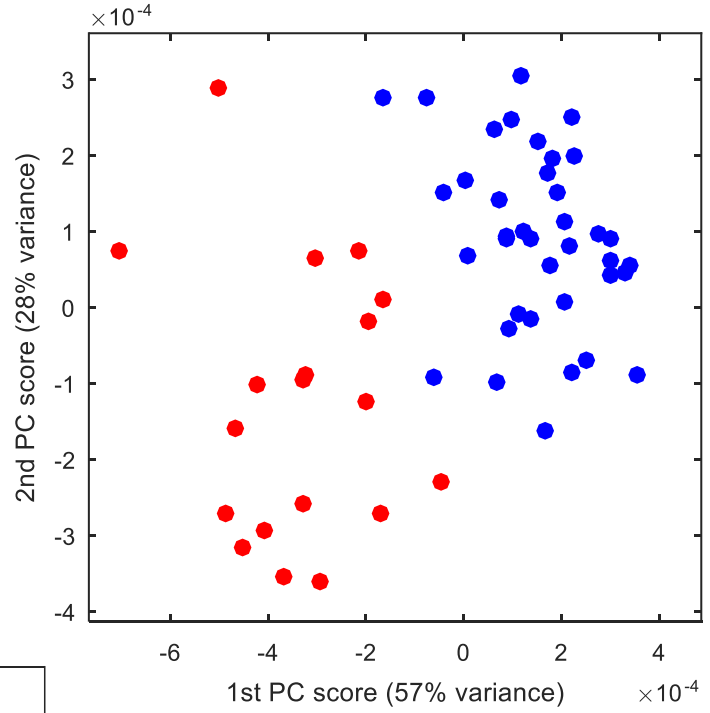
PCA in action: a simple example



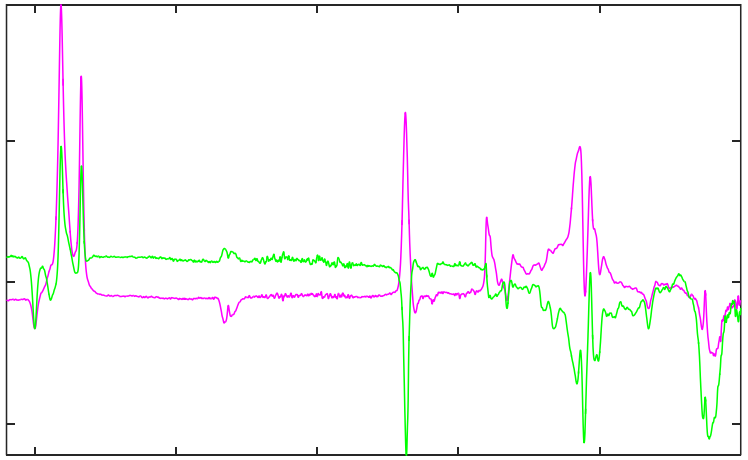
PCA in action: a simple example



PCA Scores



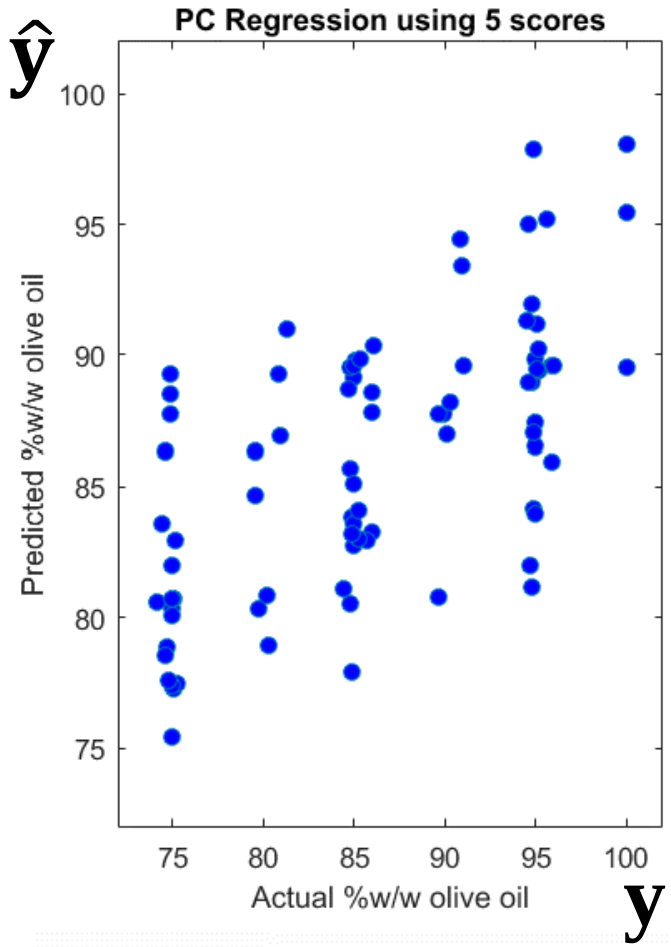
PCA Loadings



- Scores plots reveal grouping not apparent in raw data
- Loadings indicate features responsible for patterns in the scores w.r.t. each axis

PC Regression

- 164 infrared spectra of mixtures of olive oils with hazelnut oils
- Half used in Principal Component regression (half retained as a test set)



Instead of the original data, we regress onto the PC scores

$$\mathbf{y} = \mathbf{Z} \cdot \mathbf{m}$$

(n x 1) *(n x q)(q x 1)*

$$\hat{\mathbf{m}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\,\mathbf{Z}^T\,\mathbf{y}$$ *(regression coefficients)*

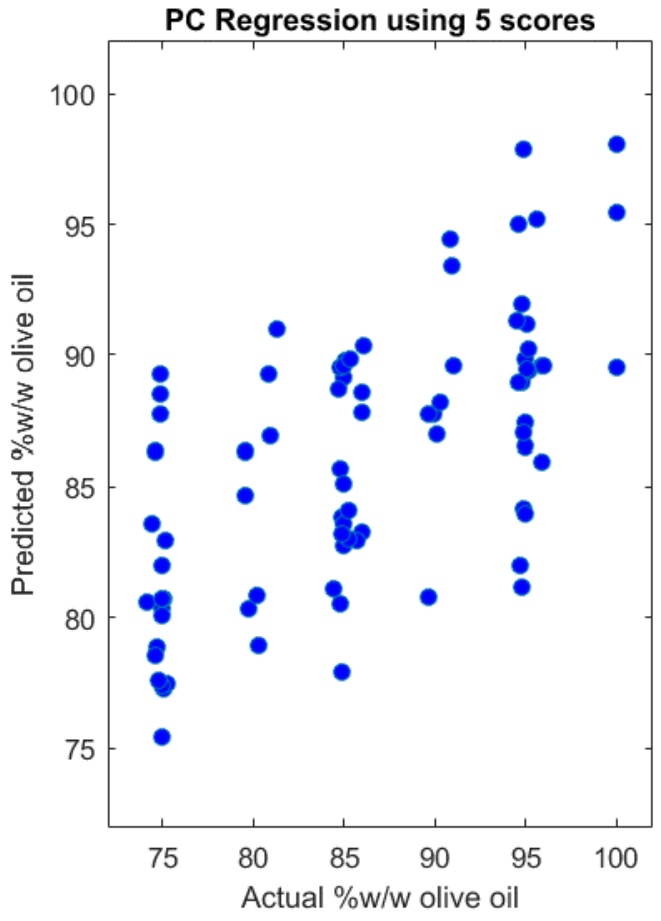
$$\hat{\mathbf{y}} = \mathbf{Z}\,\hat{\mathbf{m}}$$ *(predicted y values)*

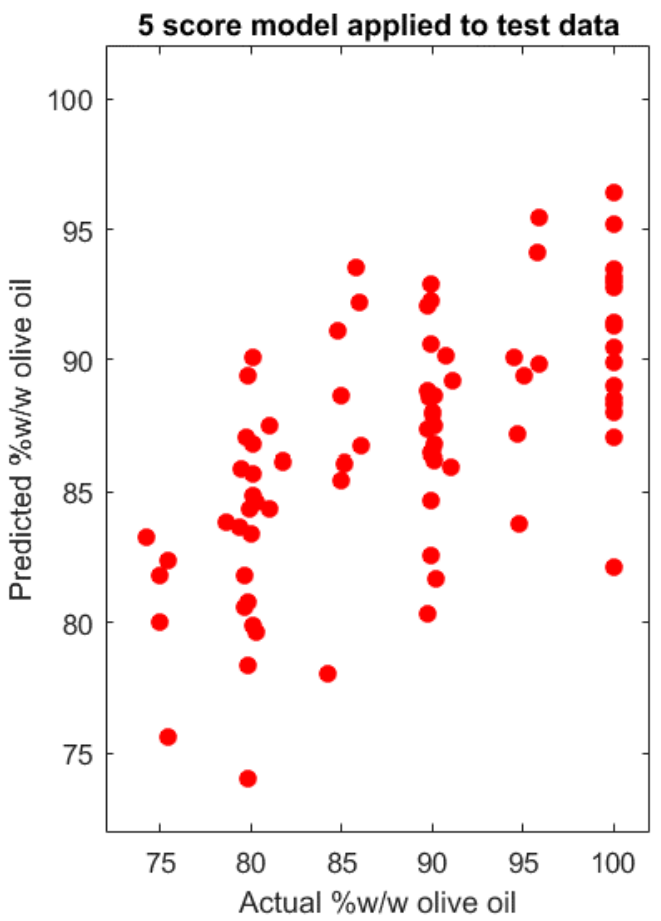
Equivalent to:

$$\hat{\mathbf{y}} = \mathbf{X}\,\mathbf{P}\,\hat{\mathbf{m}}$$

PC Regression

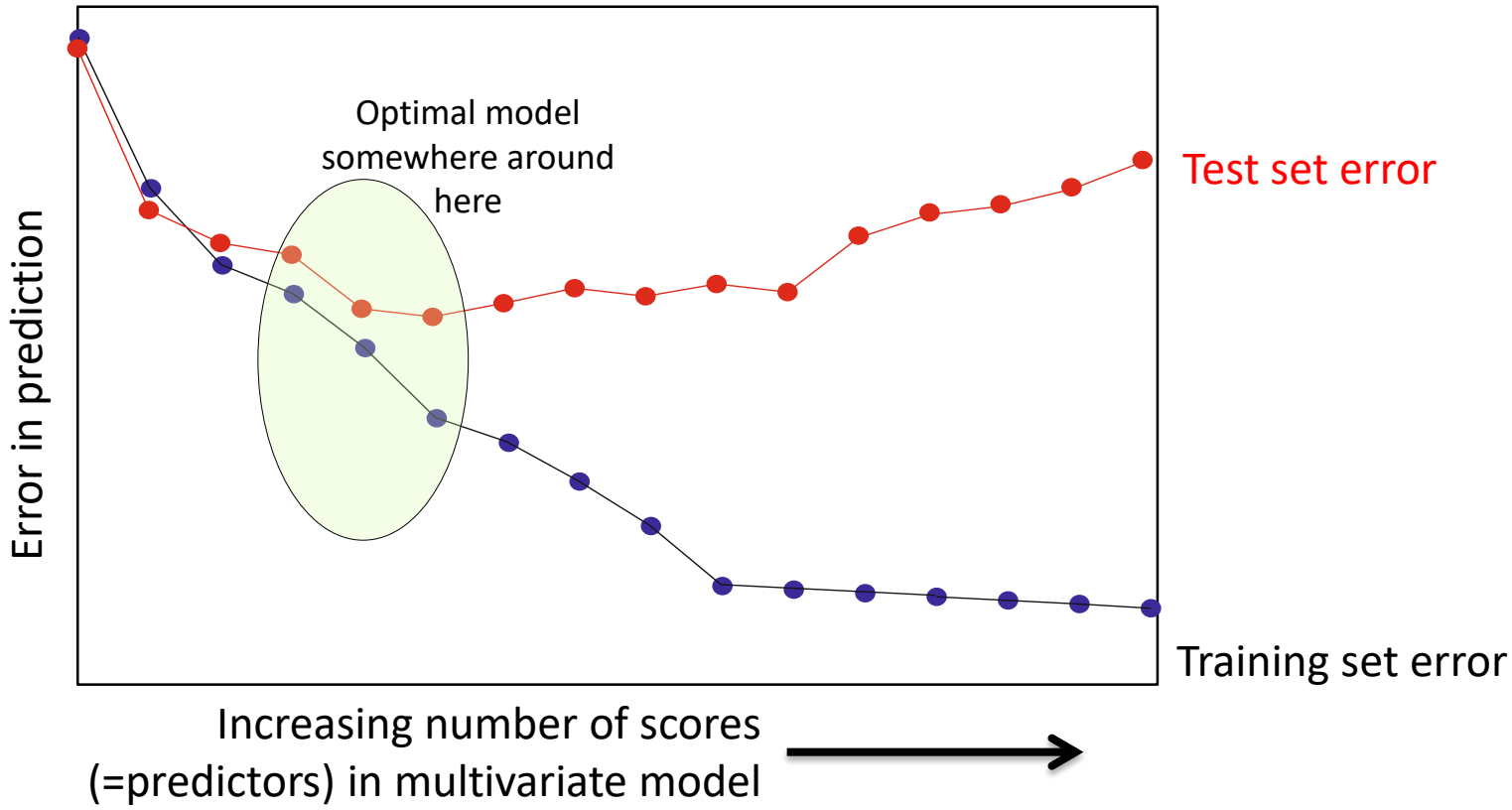
- Using too many scores leads to '**overfitting**' (including irrelevant noise in the model)
- Choosing the right number of scores is hard - this is the **'curse of dimensionality'**





PC Regression

- Choosing the right number of scores is hard - using too many leads to '**overfitting**'
- This is the **'curse of dimensionality'**



Partial Least Squares (PLS) Regression

- Like PCA, PLS is a matrix decomposition method that rearranges and compresses the data into scores:

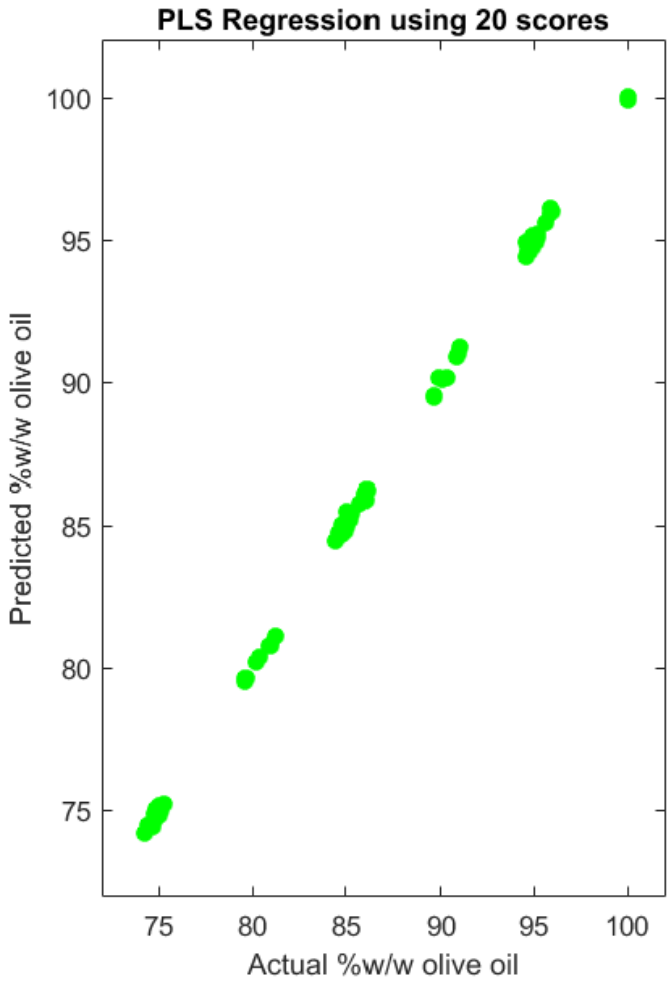
$$\mathbf{X} = \mathbf{Z}.\ \mathbf{P}^{\mathrm{T}}$$

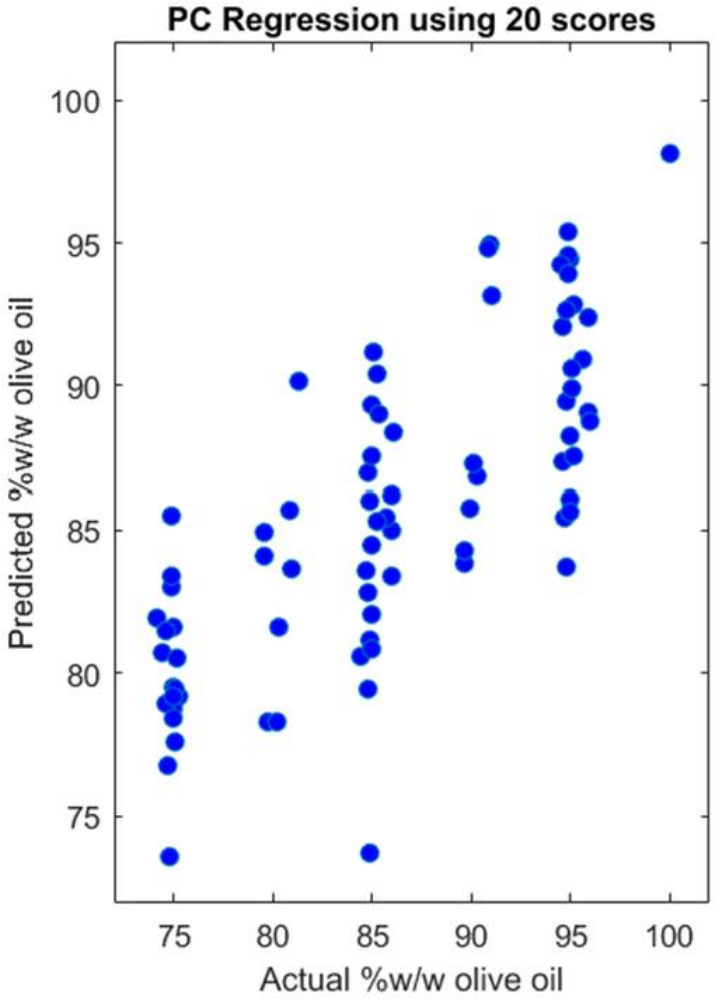
- Again, the scores are **uncorrelated**, and there are **fewer** of them than the original variates

- However, PLS also makes use of the **y** data in the matrix decomposition

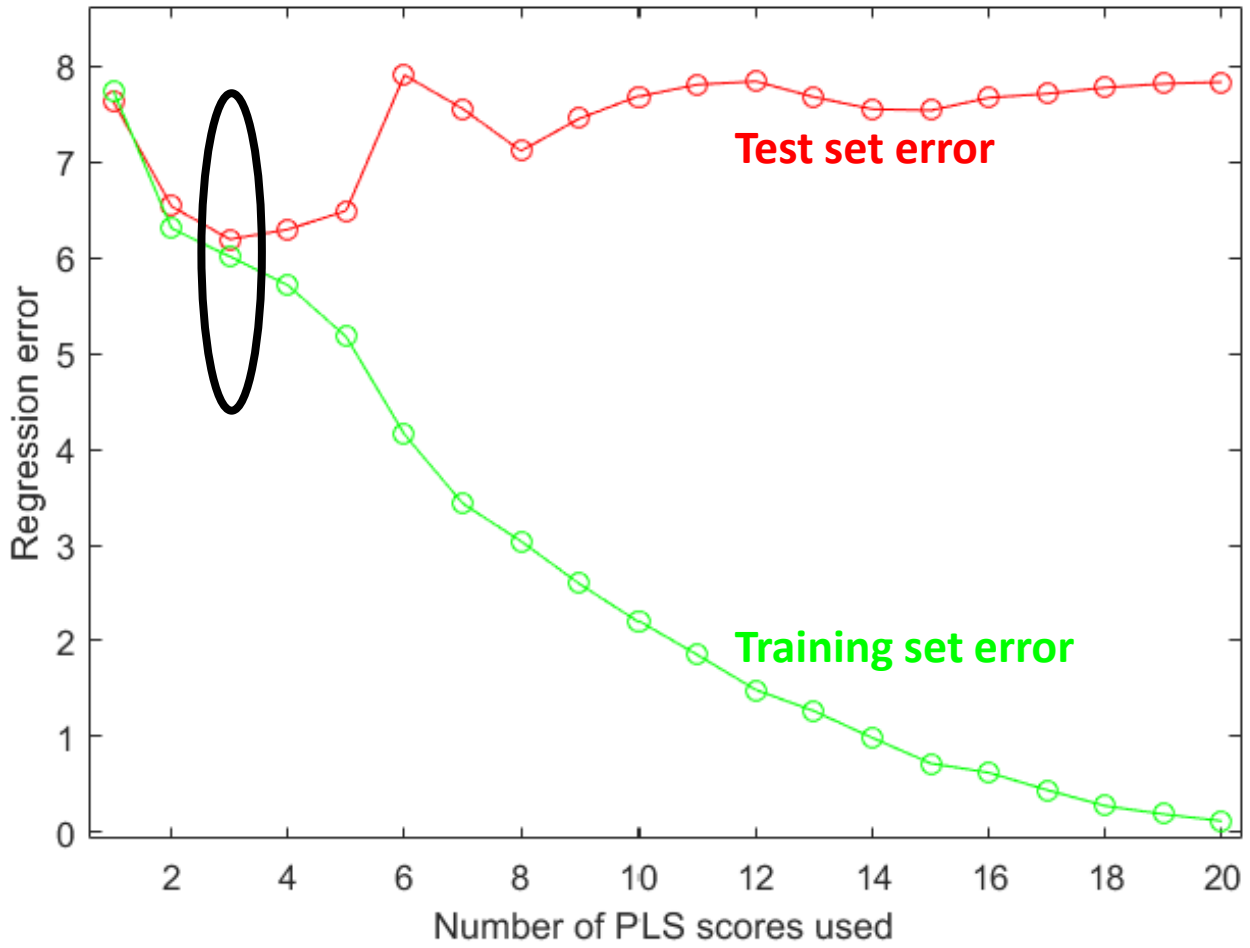
- This means that PLS is more efficient than PCA, because the scores have maximized **relevant** information content

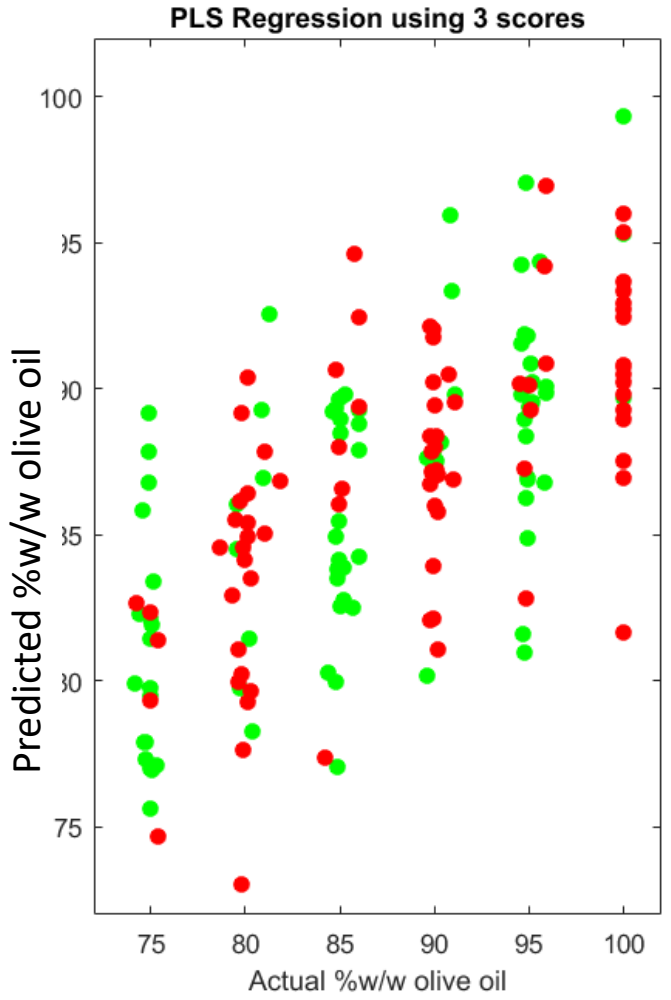
Partial Least Squares (PLS) Regression





Partial Least Squares (PLS) Regression

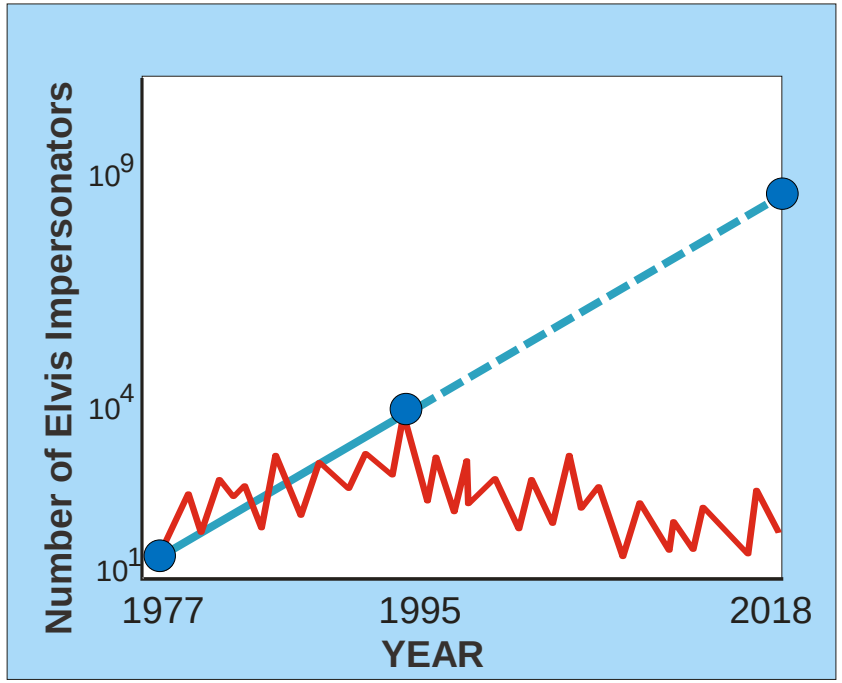




More on overfitting

"When Elvis Presley died, there were 48 professional Elvis impersonators. Today, there are 7328. If that growth is projected, within twenty years, one person in four on the face of the globe will be an Elvis impersonator."

….from the Financial Times

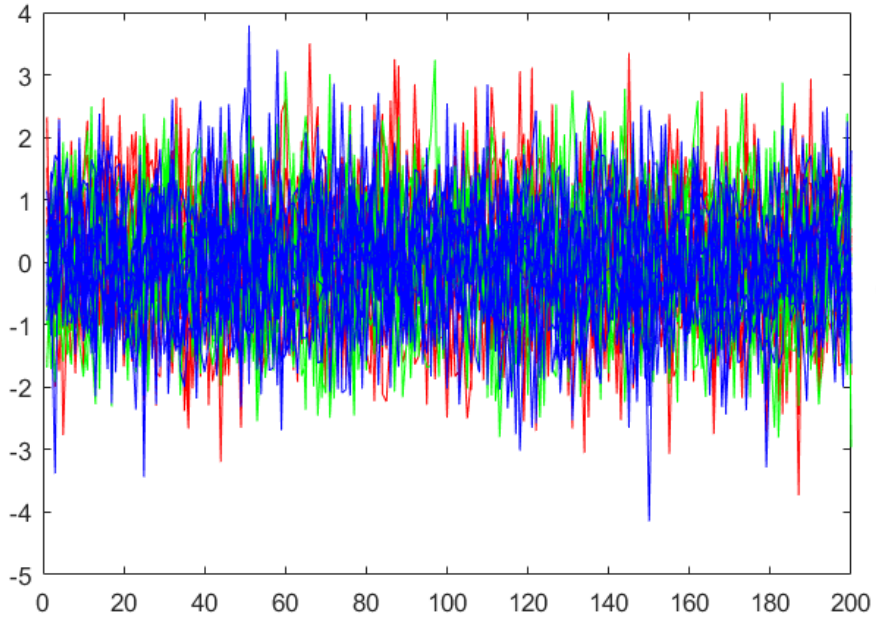


What did the "analyst" do wrong?

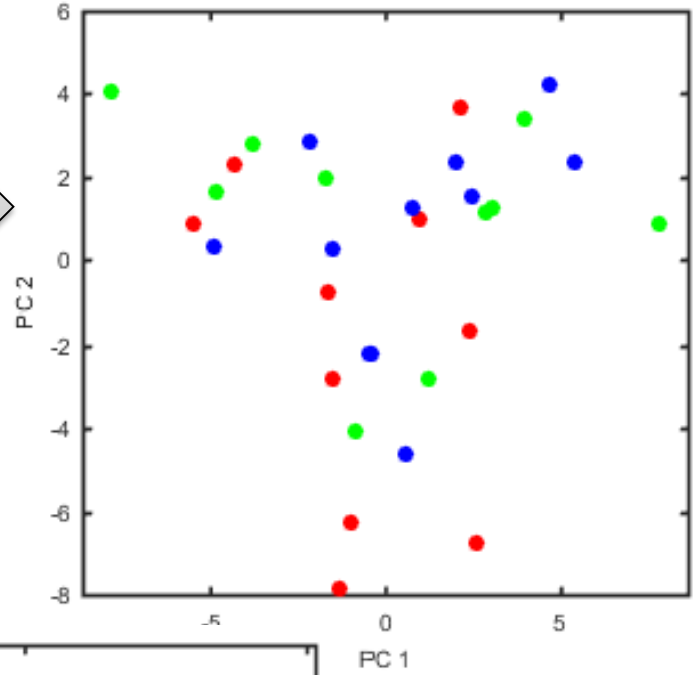
- **Only had n = 2 data points - never enough!**
- **Noise was incorporated into model (overfitting)**
- **Assumed log-linear model for no good reason**
- **Extrapolation fail**

More on overfitting

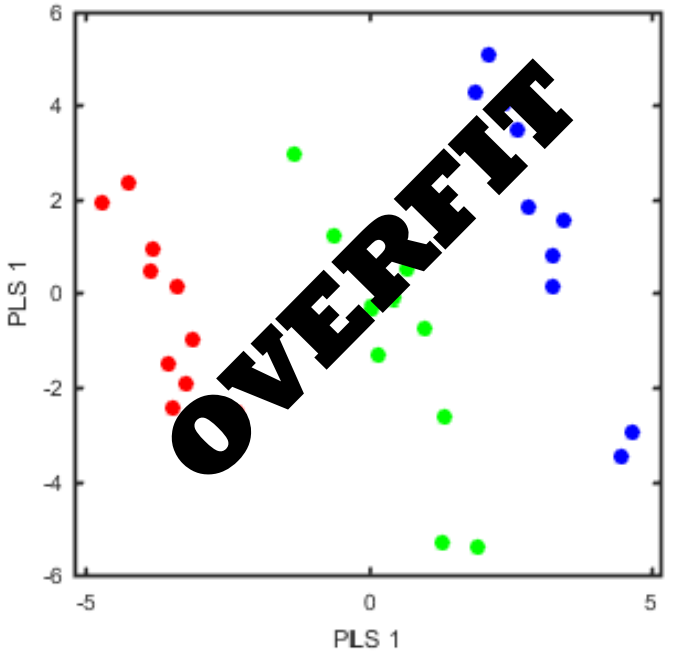
'Group 1' data
'Group 2' data
'Group 3' data



PCA



PLS



- PLS is so efficient, it can even extract some systematic information from noise
- This is, of course, entirely useless

Validation in multivariate analysis

The **only** way to be confident about the performance of a multivariate model is to perform some kind of model validation.

By applying a model to an **independent test set**, we can check for:

- Lack of generalization ability (unable to extrapolate successfully to new data)
- Incorrect assumptions about nature of model (e.g. linear , log-linear, etc)
- **Overfitting** (incorporating irrelevant information e.g. noise into the model)
  - Any multivariate model can be overfit, but this is especially likely when the data are 'high-dimensional' (as in spectroscopy/spectrometry)
  - **People also use various forms of cross-validation to guard against overfitting (permutation tests, bootstrapping etc).**

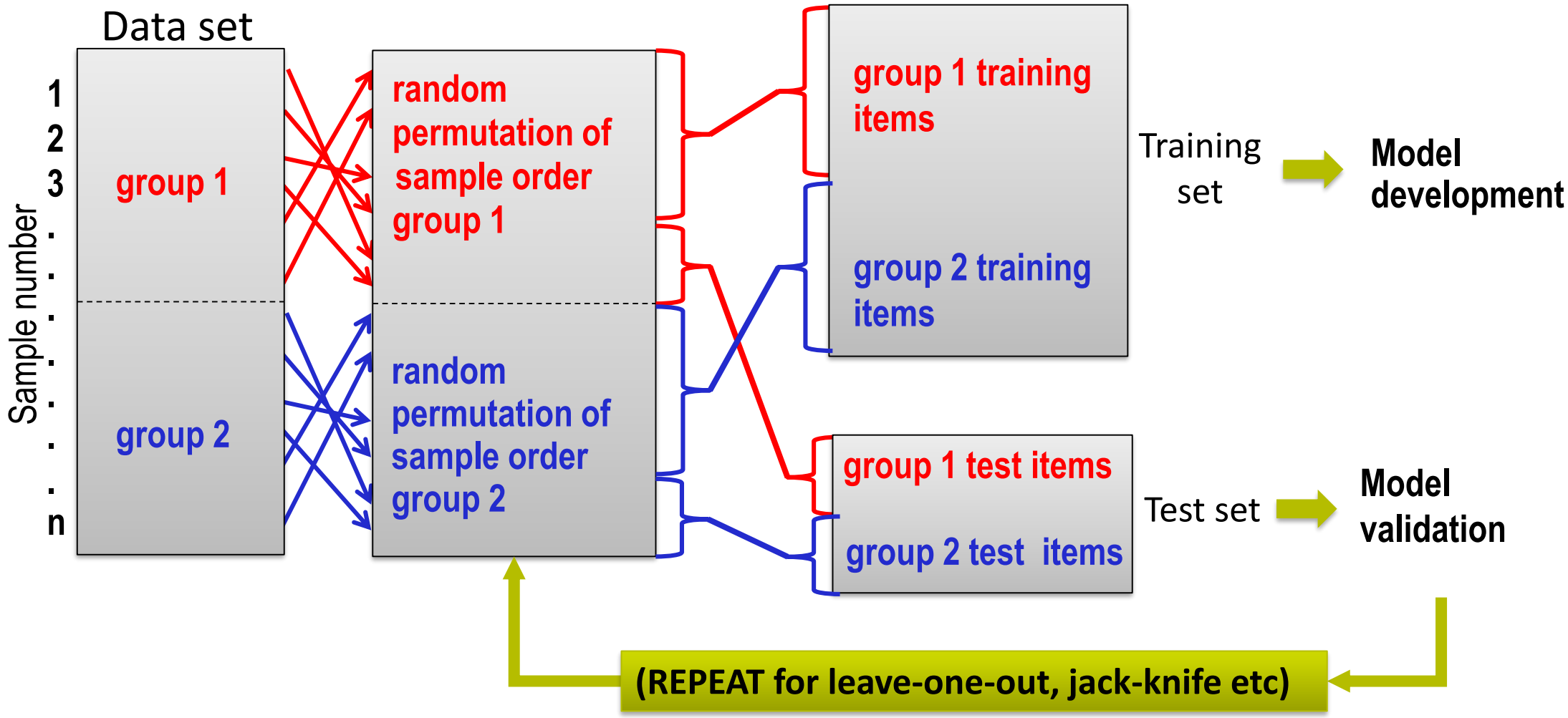
Validation in multivariate analysis

- Fast, modern computers make possible all kinds of bespoke, rigorous validation methods
  - This field of research is still evolving

- Validation not always provided in ‘standard’ data analysis software
- This can allow multivariate methods to be misused
  - Particularly true for the software supplied as standard with analytical instrumentation

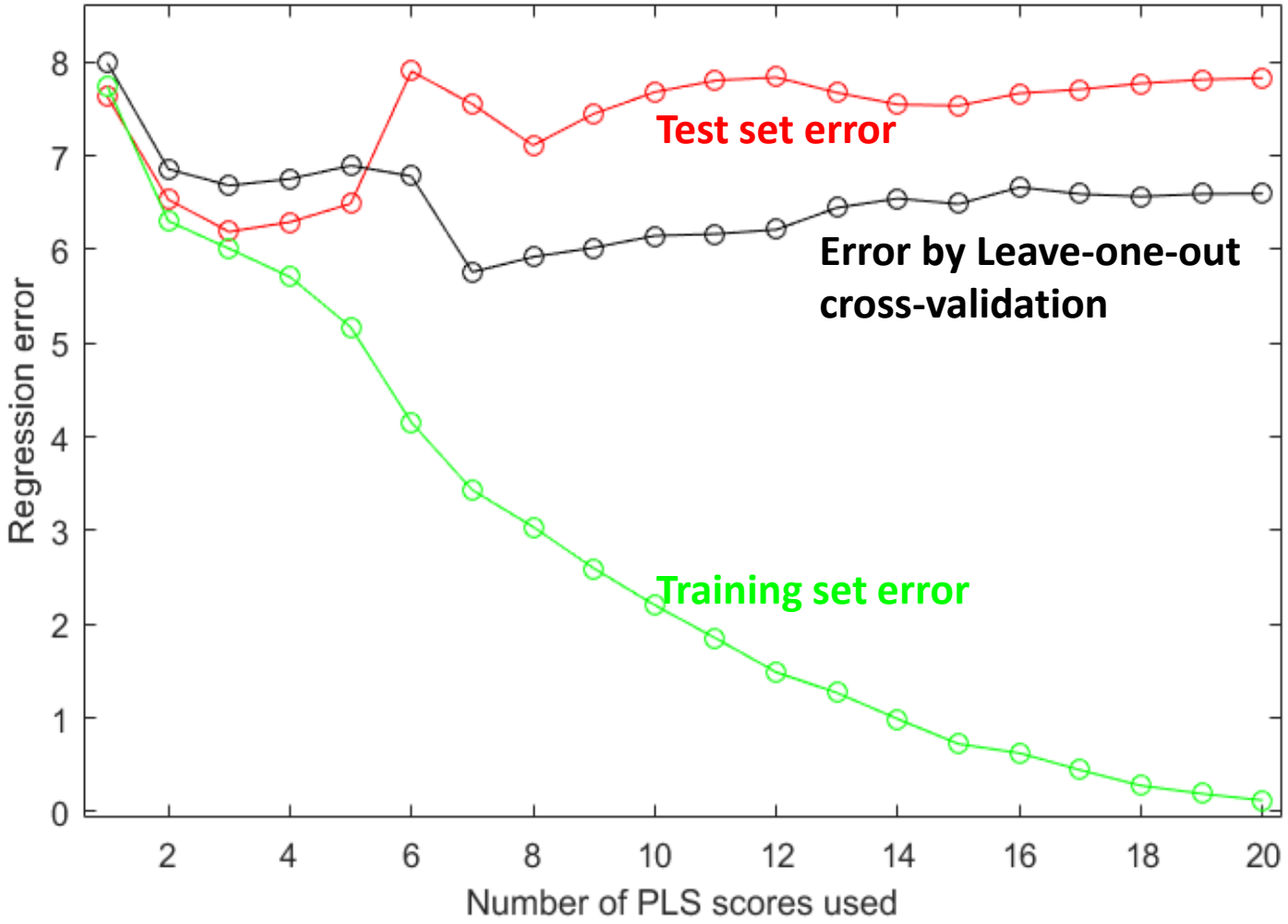
Cross-validation, bootstrapping, permutation, etc…

Example with a two-group classification problem:



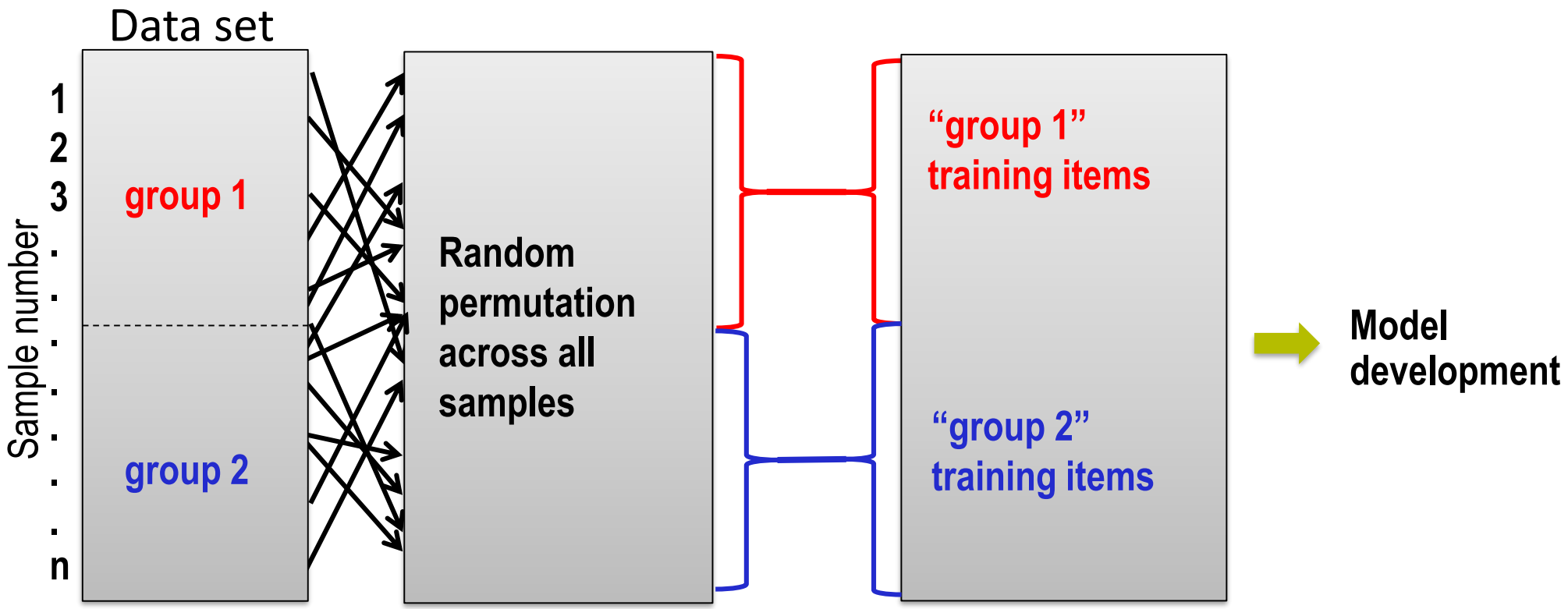
Leave-one-out cross-validation in PLS Regression

PLSR of %w/w olive oil in olive/hazelnut mixtures



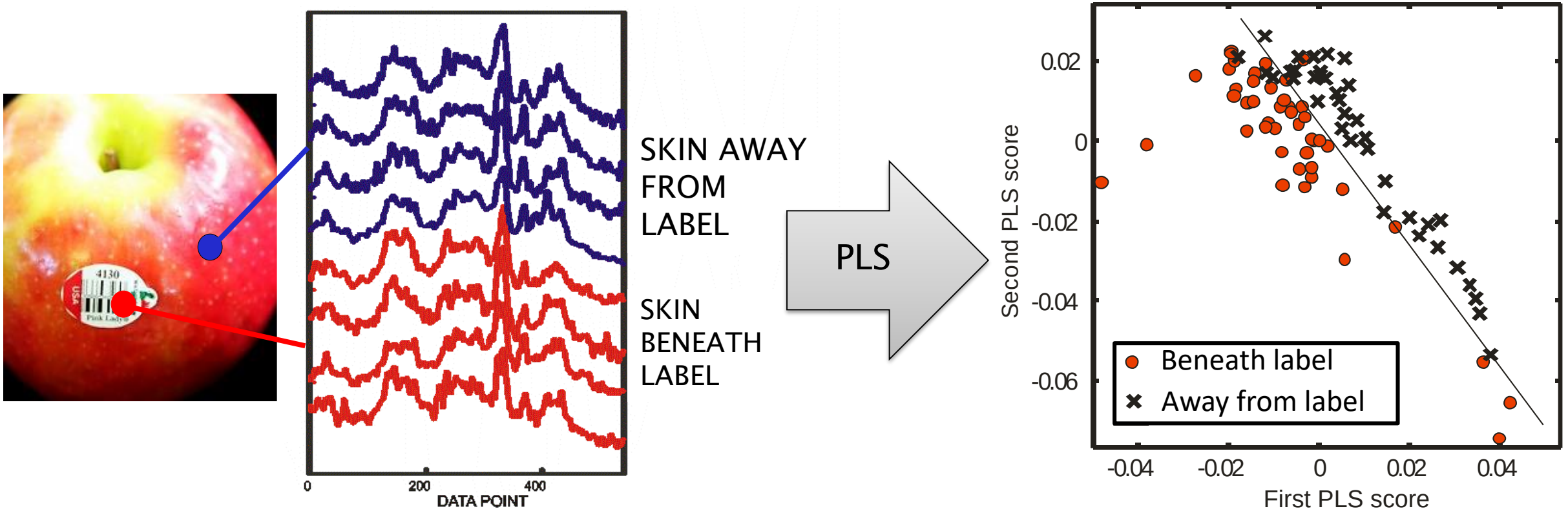
Cross-validation, bootstrapping, permutation, etc…

Permutation test:



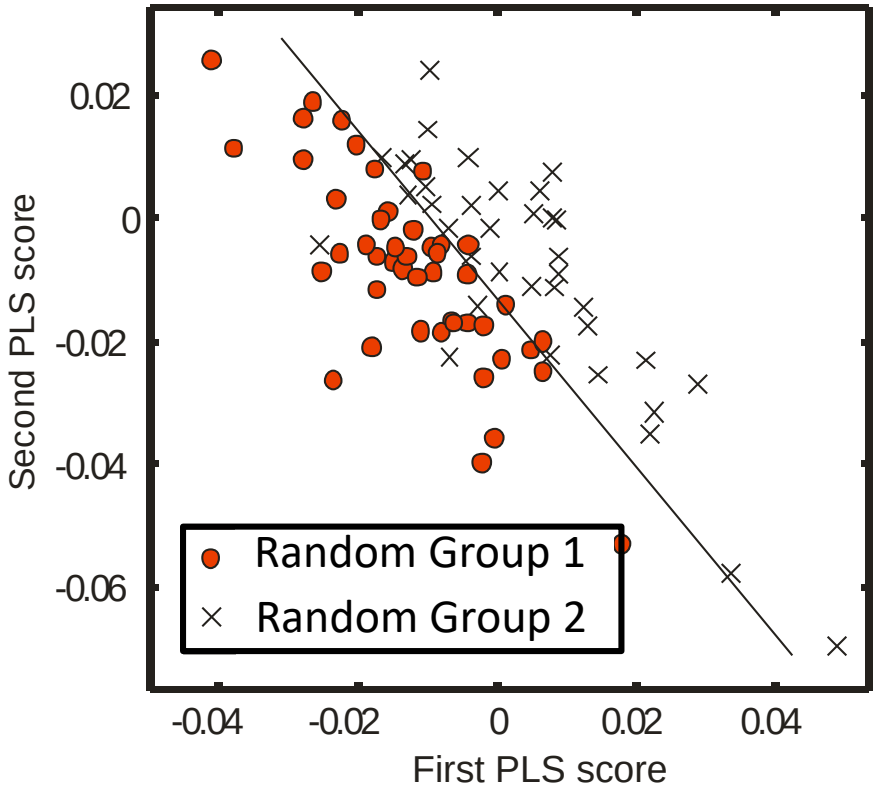
Permutation resampling to validate PLS

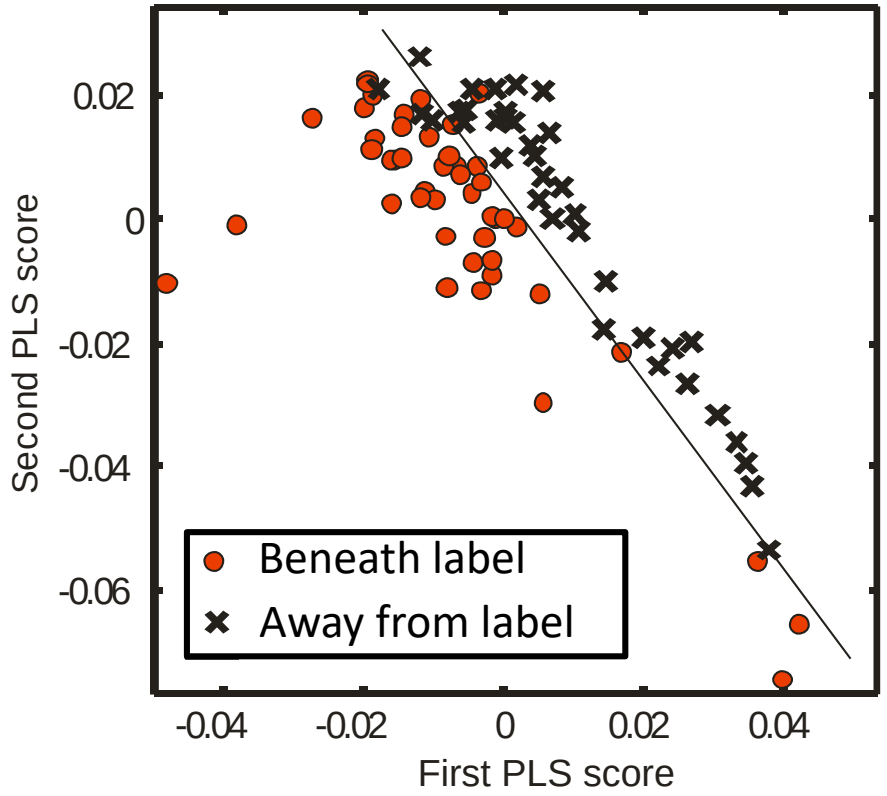
- Raman spectroscopy was used to collect spectra from apples, to look for evidence of residue from labels



Permutation resampling to validate PLS

Now repeat the PLS, but use a **randomly scrambled y-vector** when carrying out the matrix decomposition

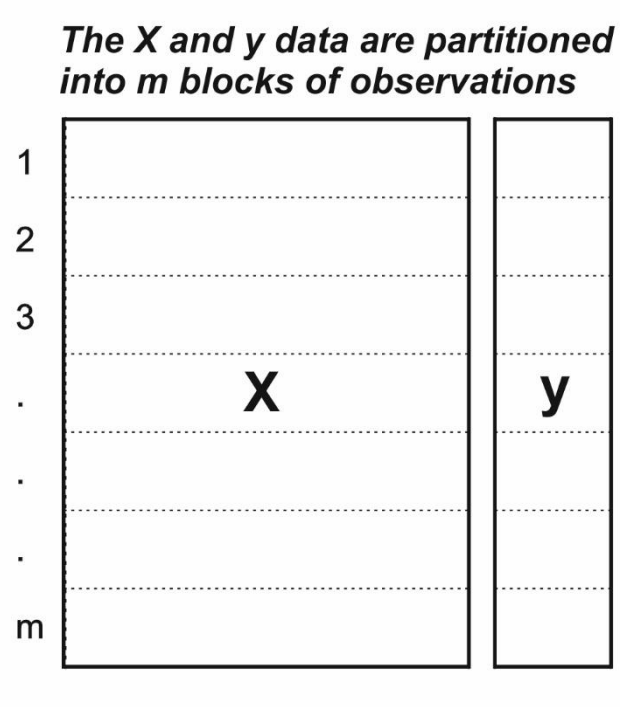


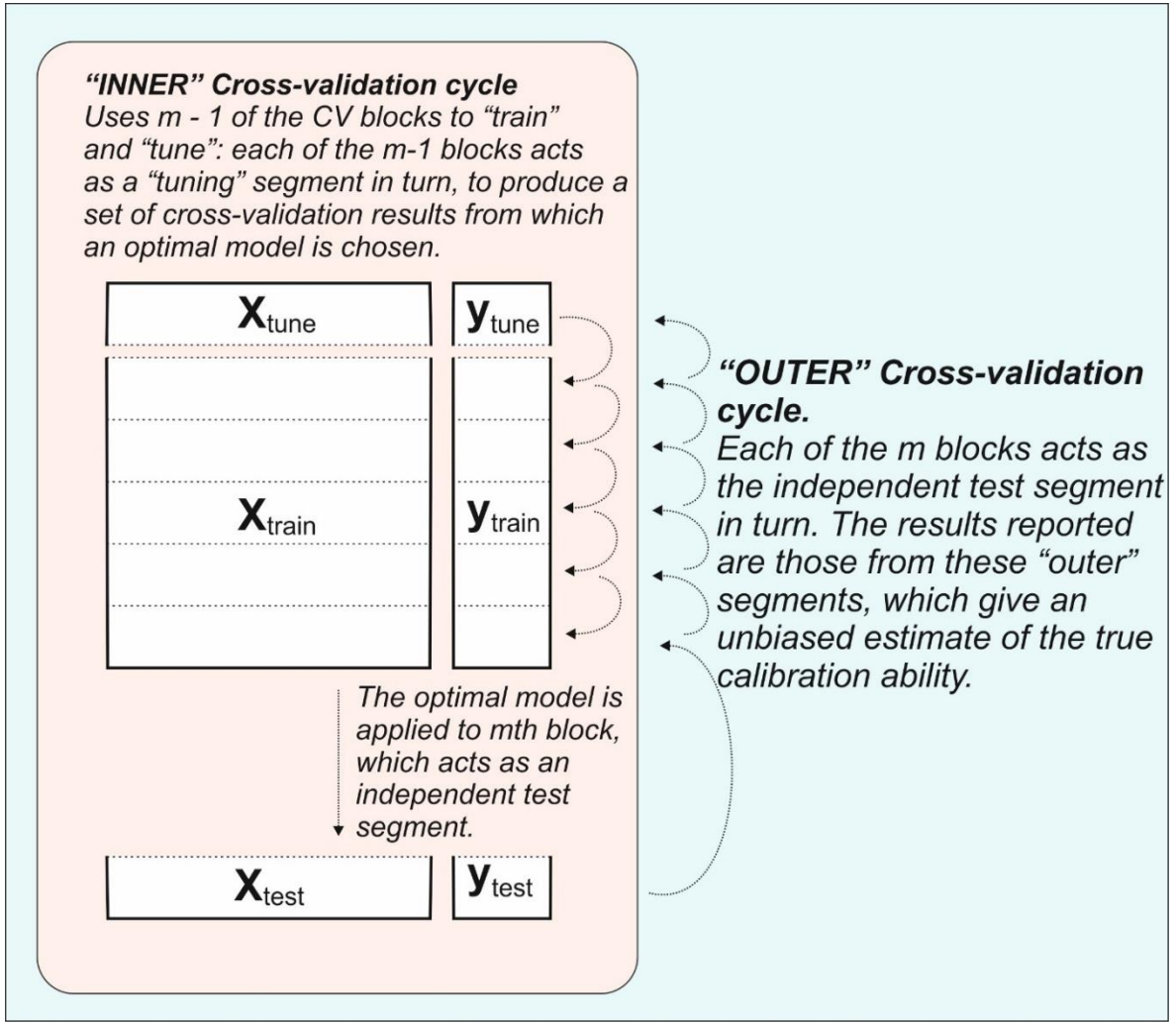


This shows that any random group assignment to the data would have produced a similar outcome – **so the finding is NOT significant**

Even more complicated schemes…

e.g. Leave-m-out double cross-validation





Summary

- Multivariate = more than one predictor variate
  - Data from modern analytical techniques is usually highly multivariate, with large number of attributes and relatively smaller number of samples
- ‘Chemometric’ data compression methods like PCA and PLS are very useful, especially for graphical representation
- Overfitting is a real possibility - naive use of multivariate statistics can lead to misleading results
- Best practice requires the use of a suitable model validation technique
- Matrix language software packages are the best platforms for chemometric analysis (e.g. Matlab, R, Python,….)
- **There is a real shortage of these skills out there in the real world!**