Chemometrics (multivariate analysis)

Dr Kate Kemsley
Analytical Sciences Unit
Institute of Food Research
kate.kemsley@ifr.ac.uk
In this talk:

1. **What is multivariate data?**
   - Representing data as matrices and vectors

2. **Introducing Chemometrics**
   - Data compression methods
     - Principal Component Analysis (PCA)
     - Partial Least Squares (PLS) Regression

3. **“Overfitting”**
   - Model validation in multivariate analysis
1. What is a “multivariate” dataset?

- Experimental data that contains measurements of multiple properties or attributes of a collection of objects (people, samples, etc)
1. What is a “multivariate” dataset?

<table>
<thead>
<tr>
<th>Child ('object')</th>
<th>Height (cm) ('predictor')</th>
<th>Age (years) ('predictor')</th>
<th>Weight (kg) ('dependent')</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alison</td>
<td>146</td>
<td>9</td>
<td>29</td>
</tr>
<tr>
<td>Bethany</td>
<td>138</td>
<td>10</td>
<td>35</td>
</tr>
<tr>
<td>Chloe</td>
<td>143</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>Denise</td>
<td>127</td>
<td>8</td>
<td>28</td>
</tr>
<tr>
<td>Emma</td>
<td>141</td>
<td>9</td>
<td>26</td>
</tr>
<tr>
<td>Frances</td>
<td>132</td>
<td>10</td>
<td>37</td>
</tr>
<tr>
<td>Grace</td>
<td>155</td>
<td>12</td>
<td>34</td>
</tr>
<tr>
<td>Helen</td>
<td>128</td>
<td>7</td>
<td>26</td>
</tr>
</tbody>
</table>

- Measurements of 3 different “attributes” of 8 “objects” (= individual children)
- Suppose we want to investigate what factors influence a child’s weight – then this is considered the “dependent” variate
- The remaining variates are called “predictors”
How could we analyse this data?

- Traditional statistical methods – one predictor at a time
  - A well-known way is linear least-squares (LS) regression

\[ y = mx + c \]

\[ \text{WEIGHT} = \text{slope.HEIGHT} + \text{intercept} \]

\[ y = mx + c \]

\[ \text{WEIGHT} = \text{slope.AGE} + \text{intercept} \]

\( m \) and \( c \) are estimated by LS regression of the dependent variate \( y \) onto the predictor variate \( x \)
What about taking a multivariate approach?

- Make use of **both predictor variates** in modelling the dependent $y$:

  \[ \text{WEIGHT} = m_1 \cdot \text{HEIGHT} + m_2 \cdot \text{AGE} + c \]

- Use **multiple linear regression** to estimate the $m$ and $c$ values
  - Until the last couple of decades (i.e. before readily available computers in the 1980s) this was slow and laborious, even for datasets of this size

- The regression model defines a **plane in a 3-d coordinate system**
\[ y = m_1 \cdot x_1 + m_2 \cdot x_2 + c \]
defines a plane in a 3-d coordinate system.
An example from chemistry...

- Concentration of oleic acid (C18:1) in edible oils
- Measured by 60MHz proton NMR
- Spectra from a selection of oils:
An example from chemistry...

- Concentration of oleic acid (C18:1) in edible oils
  - Plotted here versus “olefinic” and “bis-allylic” peak areas separately.

![Graph showing the relationship between concentration of oleic acid and olefinic peak area.](image1)

![Graph showing the relationship between concentration of oleic acid and bis-allylic peak area.](image2)
An example from chemistry...

- Concentration of oleic acid (C18:1) in edible oils
  - Plotted here versus “olefinic” and “bis–allylic” peak areas \textbf{simultaneously}:
Now consider three predictor variates....

<table>
<thead>
<tr>
<th>Child ('object')</th>
<th>Height (cm) ('predictor')</th>
<th>Age (years) ('predictor')</th>
<th>Wrist circumf. (cm) ('predictor')</th>
<th>Weight (kg) ('dependent')</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alison</td>
<td>146</td>
<td>9</td>
<td>13.2</td>
<td>29</td>
</tr>
<tr>
<td>Bethany</td>
<td>138</td>
<td>10</td>
<td>13.9</td>
<td>35</td>
</tr>
<tr>
<td>Chloe</td>
<td>143</td>
<td>9</td>
<td>14.1</td>
<td>31</td>
</tr>
<tr>
<td>Denise</td>
<td>127</td>
<td>8</td>
<td>12.8</td>
<td>28</td>
</tr>
<tr>
<td>Emma</td>
<td>141</td>
<td>9</td>
<td>12.9</td>
<td>26</td>
</tr>
<tr>
<td>Frances</td>
<td>132</td>
<td>10</td>
<td>13.6</td>
<td>37</td>
</tr>
<tr>
<td>Grace</td>
<td>155</td>
<td>12</td>
<td>14.3</td>
<td>34</td>
</tr>
<tr>
<td>Helen</td>
<td>128</td>
<td>7</td>
<td>12.3</td>
<td>26</td>
</tr>
</tbody>
</table>

- **Total of four** variates to deal with (three predictors, one dependent)
- Can’t easily plot on axes – need to leave behind “2-d thinking”
- Need different ways of representing the data graphically
General representation of large, multivariate dataset

Dependent variate

Object 1  \( y_1 \)
Object 2  \( y_2 \)
Object 3  \( y_3 \)
. . .
Object \( n \)  \( y_n \)

Predictor variates

Object 1  \( x_{11} \)
Object 2  \( x_{21} \)
Object 3  \( x_{31} \)
. . .
Object \( n \)  \( x_{n1} \)

A set of \( d \) attributes of the objects, e.g.
- A “continuous” variate – e.g. concentration of some chemical component measured by a reference technique
- A “category” variate – e.g. species, variety, cultivar, genotype, etc.

This could be something like:

- A “continuous” variate – e.g. concentration of some chemical component measured by a reference technique
- A “category” variate – e.g. species, variety, cultivar, genotype, etc.
General representation of large, multivariate dataset

Dependent variate

Sample 1
Sample 2
Sample 3
...
Sample n-1
Sample n

- type ‘A’ coffee
- type ‘A’ coffee
- type ‘A’ coffee
...
- type ‘B’ coffee
- type ‘B’ coffee

Predictor variates (>hundreds!)

Spectrum 1
Spectrum 2
Spectrum 3
Spectrum 4
...
Spectrum n-1
Spectrum n

This could be something like:
- A “continuous” variate – e.g. concentration of some chemical component measured by a reference technique
- A “category” variate – e.g. species, variety, cultivar, genotype, etc

A set of $d$ attributes of the objects, e.g.
- infrared absorbance values at $d$ different wavelengths (each row in the data matrix = a spectrum)
General representation of large, multivariate dataset

Dependent variate

Sample 1
Sample 2
Sample 3
    .
    .
Sample n–1
Sample n

Predictor variates (>hundreds!)

Spectrum 1
Spectrum 2
Spectrum 3
    .
    .
Spectrum n–1
Spectrum n

This could be something like:

- A “continuous” variate – e.g. concentration of some chemical component measured by a reference technique
- A “category” variate – e.g. species, variety, cultivar, genotype, etc

A set of $d$ attributes of the objects, e.g.
- infrared absorbance values at $d$ different wavelengths (each row in the data matrix = a spectrum)
Linear regression using matrix algebra

\[ y = m_1.x_1 + m_2.x_2 + \ldots + c \]

Using matrix algebra this becomes:

\[ y = Xm \]

(to remove the need for \( c \), \( X \) is mean-centered)

Least-squares solution for \( m \) is:

\[ m = (X^T X)^{-1} X^T y \]

(T = matrix transpose, -1 = matrix inverse)

Practically impossible to do this before computers
The problems in analysing such datasets

Large number of objects (people, samples, etc...) and huge numbers of variates (“high-dimensional”) mean that it is...

- Difficult to examine the data set graphically
- Usually “under–determined” (*number of data values per object exceeds number of objects*)
  - $X$ has fewer rows than columns
  - Most columns of $X$ are inter–correlated, sometimes to a large extent
  - $(X^T X)$ is singular (degenerate) and not invertible
  - This leads to mathematical problems in applying many statistical modelling methods directly
2. Chemometric methods

- “Chemometrics” – a family of methods for high-dimensional data analysis

- Principal component analysis (PCA)

- Partial least squares (PLS)

- Also… artificial neural networks, genetic algorithms, discriminant analysis,…and many more… (+ lots of synonyms!)
Chemometrics – history

- Originated in the 1980’s when computers first started to become connected to analytical instruments (*especially near-infrared*)

- Some methods were proposed several decades earlier, e.g. Principal Component Analysis, but were impossible to carry out in practice

- Same methods have spread throughout the sciences – e.g. psychology, sensometrics, econometrics, meteorology, bioinformatics,...
What can we do with chemometric analysis?

- **Data exploration ("Unsupervised analysis")**
  - Examine a matrix of experimental data (e.g. a large collection of spectra) looking for patterns, groups, etc

Probably the most widely used chemometric approach used for data exploration is Principal Component Analysis.
What can we do with chemometric analysis?

- (“Supervised analysis”)

- “Classification” type problems
  - E.g. model the differences between groups of data from different sample types

- “Calibration” type problems
  - e.g. relate spectral data to sensory data, concentration data, etc.

A method commonly used for these modelling approaches is Partial Least Squares (PLS) Regression.
Unsupervised vs. Supervised methods

- **Unsupervised**
  - Matrix $X$ containing the experimental measurements
  - Typically we want to examine data graphically, look for patterns

- **Supervised** – there’s a second set of information
  - Vector $y$ (or matrix $Y$) of dependent variates e.g. reference concentrations, group information
  - Typically we want to explore the relationship between $X$ and $y$
  - Here we need to use MODELLING methods
Rearranges the variance (= information) in the data set to make it easier to deal with (visualise, display, analyse further, etc)

- Definition using matrix algebra:
  \[ Z = X \cdot P \]
- The data in X are post-multiplied by loadings P (“weights”) to give a set of scores Z
- This can be considered as a data rotation
- Loadings are the eigenvectors of data covariance matrix:
  \[ X^T X / (n-1) \]
Partial Least Squares regression

- A commonly used supervised method, for performing multiple linear regression on high-dimensional data

- Regression model: $y = X \, m$ (but can’t solve this directly if $X$ has $d > n$)

- PLSR writes:
  - $y = Z \, b = X \, P \, b$ (where $Z = X \, P$, the data compression step)
  - $b = (Z^T Z)^{-1} Z^T Y$ ($Z^T Z$ is invertible, whereas $X^T X$ is not)
  - $m = P \, b$ (allows us to get a solution for $m$)
PCA and PLS – Both are data compression methods

\[
\begin{align*}
X & \quad \text{Data matrix} \\
Z & \quad \text{Matrix of Scores} \\
PT & \quad \text{Matrix of Loadings} \\
r & \quad \text{data points} \\
d & \quad \text{data points}
\end{align*}
\]

Scores:
- \( Z \) is smaller than \( X \) – easier to explore (e.g. by plotting graphically)
- “redundancy” removed
- may reveal patterns or groups which were not clear in original data

\( r \ll d \)
PCA and PLS – Both are data compression methods

The $k$th spectrum times the 1st and 2nd loading…

…gives the scores along the 1st and 2nd axes for that spectrum

Scores plots:
- Each point represents an individual spectrum

Loadings plots:
- Show relative importance of each variate
- Information on same scale as original spectra
Data compression is useful because...

- After data compression, scores have successively maximized...
  
  - “Information content”  
    \( (= \text{variance, in PCA}) \)
  
  - “Relevant information content”  
    \( (= \text{covariance with the dependent variate, in PLS}) \)

- Scores are uncorrelated, and there are fewer of them than the original variates
  - This facilitates further statistical analyses, using methods that would not otherwise be possible
Data compression: a simple example of its usefulness

Raw data: 60 spectra of edible oils

- Hazelnut
- Extra virgin Olive

- X matrix – [60 x 250] (oils x absorbance values)
- y vector – a category “dummy” variate indicating whether a spectrum was from an olive or hazelnut oil
Scores plots reveal grouping not apparent in raw data.

Loading indicates regions of the spectrum associated with each group type.

Scores on first axis are enough to almost entirely distinguish the groups.
3. “Overfitting”

- Any mathematical model can be “overfit”, by inadvertently including irrelevant information into the model.

- Other potential problems with modelling include:
  - Lack of generalization ability – i.e., model is required to extrapolate to new data, and is unable to do so.
  - Incorrect assumption about nature of model (e.g. is linear modelling appropriate).
"When Elvis Presley died, there were 48 professional Elvis impersonators. Today, there are 7328. If that growth is projected, by the year 2012, one person in four on the face of the globe will be an Elvis impersonator."

Financial Times, March 1995

The “analyst” had...

Assumed log-linear model

Extrapolated

Incorporated noise into the model

n = 2, not enough data!
Examples of overfitting & other modelling problems...

[30 x 200] Matrix of “data” (= noise), assigned to 3 (meaningless) groups

Partial Least Squares Regression attempts to model the difference between “groups”

- n = 30 here, but overfitting occurs easily because n << d
Overfitting in multivariate analysis

- Any multivariate model can be overfit, but this is especially likely when $d >> n$
  - This is the case for virtually all spectroscopy and spectrometry techniques, which measure large numbers of properties (absorbances values, particle counts, etc) on each sample

- The only way to know that overfitting is not a problem is to perform rigorous model validation
  - Using totally independent test samples
  - Using various cross-validation schemes
Modern validation methods make extensive use of resampling (bootstrapping) as computing power has become so readily available (last ~5 years).

However, many standard software tools for multivariate analysis still provide only limited validation approaches, allowing multivariate methods to be misused.
- Particularly true for the software supplied as standard with instrumentation.

Solution: use of bespoke software (coding in statistical and matrix languages).
Detecting overfitting in a real experimental dataset

Raman Spectra from Apple skins

PLS is repeated with a randomly scrambled y-vector

This shows that any random group assignment would have produced the same outcome – so the finding is NOT significant
An example validation strategy: Resampled test sets in a 2-group case

DATA MATRIX

1. 
2. 
3. 
4. 
5.

observation

n

random permutation of observations in group 1

training (grp 1)

test (grp 1)

training (grp 2)

test (grp 2)

random permutation of observations in group 2

Model development

Model validation

Iterate
A more rigorous validation strategy in a calibration–type application: Full double–cross–validation

The X and y data are partitioned into m blocks of observations

“INNER” Cross-validation cycle
Uses m - 1 of the CV blocks to “train” and “tune”: each of the m-1 blocks acts as a “tuning” segment in turn, to produce a set of cross-validation results from which an optimal model is chosen.

“OUTER” Cross-validation cycle.
Each of the m blocks acts as the independent test segment in turn. The results reported are those from these “outer” segments, which give an unbiased estimate of the true calibration ability.
Key points

- Multivariate = more than one predictor variate
- Data from modern analytical techniques is usually highly multivariate, with large \( d \) (number of properties) and relatively smaller \( n \) (number of samples)
- 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 is to use an advanced cross-validation technique that exploits modern computing capability
- Software languages exist that are specifically designed to deal with this type of analysis
Further resources

Software:
Matlab Student License/Trial license, & “Getting Started” Manual
http://www.mathworks.co.uk/academia/student_version/

Books:
“Principles of Multivariate Analysis: A User's Perspective” by Wojtek J. Krzanowski
“Multivariate Calibration” by Harald Martens & Tormod Næs