A Guide to the Practical Use of Multivariate Analysis in SIMS

Joanna Lee, Ian Gilmore
National Physical Laboratory, Teddington, UK

Email: joanna.lee@npl.co.uk
Web: http://www.npl.co.uk/nanoanalysis

© Crown Copyright 2010
1. Introduction
   • What is multivariate analysis?
   • Some matrix algebra…
2. Identification
   • Principal component analysis (PCA)
   • Multivariate curve resolution (MCR)
3. Quantification and prediction
   • Partial least squares regression (PLS)
4. Classification
   • PCA classification
   • Principal Component Discriminant Function Analysis (PC-DFA)
   • Partial Least Squares Discriminant Analysis (PLS-DA)
5. Conclusion
Why are we here?

Number of Publications

- PCA
- MCR
- PLS
- DFA
- ANNs

Year of Publication

- 1990
- 2000
- 2010

Original Artist, reproduction rights obtainable from www.CartoonStock.com
Data analysis

Identification

What chemicals are on the surface?
Where are they located?

SIMS Dataset

Calibration / Quantification

How is it related to known properties?
Can we predict these properties?

Classification

Which group does it belong to?
What are the differences between groups?
1. Introduction
   • What is multivariate analysis?
   • Some matrix algebra…
2. Identification
3. Quantification and prediction
4. Classification
5. Conclusion
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**

- **Multivariate** = *More than 1 variable*
- **Multivariate analysis** is the statistical study of the dependence (covariance) between different variables

- **Variables** are *numerical values* that we can measure on a *sample*
  
  *Example 1:* A sample of **people**
  Variables: Height, weight, shoe size, days since last haircut…
  
  *Example 2:* A sample of **weather**
  Variables: Temperature, humidity, wind speed, visibility, UV index…
  
  *Example 3:* A sample of **SIMS spectra**
  Variables: Intensity of Si⁺ peak, intensity of O⁺ peak, intensity of C₃H₅⁺ peak…

**Key points**
- Many surface analytical technique, incl. SIMS and XPS, gives data that are *multivariate* in nature
- Finding *correlations* in the data is the key to multivariate analysis!
Why use multivariate analysis?

72 ion images (out of > 400!)

- Modern ToF-SIMS instrument generates huge, multivariate data sets
- Manual analysis involves selecting a sub-set of most interesting features for analysis by eye
- Multivariate analysis involves simultaneous statistical analysis of all the variables
- Multivariate analysis can summarise the data with a large number of variables, using a much smaller number of “factors”
Advantages and disadvantages

• Advantages
  – Fast and efficient on modern computers
  – Uses all information available
  – Improves signal to noise ratio
  – Statistically valid, removes potential bias

• Disadvantages
  – Lots of different methods, procedures, terminologies
  – Can be difficult to understand and interpret
Why use multivariate analysis?

Multivariate methods such as PCA, PLS useful

Counts per peak

Peaks or information bins

Ordinary “x ÷ + -” jolly good here
1. Introduction
   - What is multivariate analysis?
   - Some matrix algebra…
2. Identification
3. Quantification and prediction
4. Classification
5. Conclusion
A matrix is simply a rectangular table of numbers!

\[
\mathbf{X} = \begin{bmatrix}
9 & 32 & 10 & 1 & 21 \\
18 & 20 & 22 & 4 & 12 \\
24 & 12 & 30 & 6 & 6
\end{bmatrix}
\]

\(\mathbf{X}\) has 3 row and 5 columns \(\rightarrow 3 \times 5\) data matrix

Each row (spectra) is represented by a vector.
Matrix algebra

Matrix addition

\[ \mathbf{A} + \mathbf{B} = \mathbf{C} \]

\[(I \times K) + (I \times K) = (I \times K)\]

- \(\mathbf{A}\) and \(\mathbf{B}\) 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

\[ \mathbf{A} \times \mathbf{B} = \mathbf{C} \]

\[(I \times N) \times (N \times K) = (I \times K)\]

- No. of columns of \(\mathbf{A}\) must be equal no. of rows of \(\mathbf{B}\)
- Row \(i\) of \(\mathbf{A}\) times column \(j\) of \(\mathbf{B}\) gives the row \(i\) and column \(j\) of the product matrix \(\mathbf{AB}\)

\[
\begin{bmatrix}
1 & 4 \\
2 & 2 \\
4 & 2
\end{bmatrix} \times
\begin{bmatrix}
1 & 2 \\
3 & 2
\end{bmatrix} =
\begin{bmatrix}
1 \times 1 + 4 \times 3 & 1 \times 2 + 4 \times 2 \\
2 \times 1 + 2 \times 3 & 2 \times 2 + 2 \times 2 \\
4 \times 1 + 2 \times 3 & 4 \times 2 + 2 \times 2
\end{bmatrix}
\]

\[
=
\begin{bmatrix}
13 & 10 \\
8 & 8 \\
10 & 12
\end{bmatrix}
\]
Matrix algebra and SIMS

Data matrix

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

Variables [mass]

Sample composition

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

Chemical spectra

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

Variables [mass]

\[
\begin{bmatrix}
\text{Chemical 1} \\
\text{Chemical 2} \\
\end{bmatrix}
\]

\[
\begin{array}{cc}
\text{Sample 1} & 5 \\
\text{Sample 2} & 2 \\
\text{Sample 3} & 0 \\
\end{array}
\]

\[
\begin{array}{cc}
\text{Sample 1} & 1 \\
\text{Sample 2} & 4 \\
\text{Sample 3} & 6 \\
\end{array}
\]
1. Each spectrum can be represented by a vector
2. Instead of \( x, y, z \) in 3D real space, the axes are \( \text{mass1}, \text{mass2}, \text{mass3} \ldots \) etc in \( \text{variable space} \) (also ‘\( \text{data space} \)’)
3. Assuming the data are a linear combination of chemical spectra, we can write it as a product of two matrices.
4. There are infinite number of possible solutions!

\[
\begin{bmatrix}
9 & 32 & 10 & 1 & 21 \\
18 & 20 & 22 & 4 & 12 \\
24 & 12 & 30 & 6 & 6 \\
\end{bmatrix}
= \begin{bmatrix}
5 & 1 \\
2 & 4 \\
0 & 6 \\
\end{bmatrix}
\begin{bmatrix}
1 & 6 & 1 & 0 & 4 \\
4 & 2 & 5 & 1 & 1 \\
\end{bmatrix}
\]
We can describe the data as a linear combination of spectra, by writing the data matrix as product of two matrices:

One contains the spectra ("loadings")
One containing the contributions ("scores")

This is the basis of **factor analysis**!
1. Introduction
2. **Identification**
   - Principal component analysis (PCA)
   - PCA walkthrough
   - Data preprocessing
   - PCA examples
   - Multivariate curve resolution (MCR)
   - MCR examples
3. Quantification and prediction
4. Classification
5. Conclusion
Data analysis

Identification

**What** chemicals are on the surface?

**Where** are they located?

SIMS Dataset

Classification

Calibration / Quantification

**How** is it related to known properties?

Can we **predict** these properties?

**Which** group does it belong to?

What are the **differences** between groups?
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.

<table>
<thead>
<tr>
<th>Terms Here</th>
<th>Symbol</th>
<th>Definition</th>
<th>PCA</th>
<th>MCR</th>
<th>PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
<td>-</td>
<td>An axis in the data space of a <strong>factor analysis</strong> model, representing an</td>
<td>Principal Component</td>
<td>Pure Component</td>
<td>Latent Vectors, Latent Variables</td>
</tr>
<tr>
<td></td>
<td></td>
<td>underlying dimension that contributes to summarising or accounting for the original data set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loadings</td>
<td>P</td>
<td>Projection of a factor onto the variables</td>
<td>Loadings, Eigenvector</td>
<td>Pure Component Spectrum</td>
<td>Loadings</td>
</tr>
<tr>
<td>Scores</td>
<td>T</td>
<td>Projection of the samples onto the factors</td>
<td>Scores, Projections</td>
<td>Pure Component Concentration</td>
<td>Scores</td>
</tr>
</tbody>
</table>
**Principal component analysis (PCA)**

- Factors are directions in the data space chosen such that they reflect interesting properties of the dataset.
- Equivalent to a rotation in data space – factors are new axes.
- Data described by their projections onto the factors.
Principal component analysis (PCA)

- The projection of the PCA factors onto the original variables \((m_1, m_2)\) are ‘loadings’
- The projection of the samples (stars) onto the PCA factors are ‘scores’
- The data is fully described by \(D\) factors, where \(D\) is the ‘dimensionality of the data’ (number of samples or variables, whichever is smaller)

\[
X = TP' \quad (I \times K) = (I \times D)(D \times K)
\]

\(I = \) no. of samples  
\(K = \) no. of mass units  
\(D = \) dimensionality of data

Data matrix  
Scores matrix  
Loadings matrix
Principal component analysis (PCA)

- PCA extracts orthogonal (uncorrelated) factors that successively capture the largest amount of variance within the data.
- The amount of variance described by each factor is called ‘eigenvalue’.

\[ X = TP' \]

\[ (I \times K) = (I \times D)(D \times K) \]

- \( I \) = no. of samples
- \( K \) = no. of mass units
- \( D \) = dimensionality of data

\( X \) = Data matrix
\( T \) = Scores matrix
\( P' \) = Loadings matrix
Principal component analysis (PCA)

- By removing higher factors (small variance due to noise) we can reduce the dimensionality of data ⇒ ‘factor compression’
- Often hundreds of variables can be described with just a handful of factors!

\[ \bar{X} = TP' \]

\( l = \text{no. of samples} \)
\( K = \text{no. of mass units} \)
\( N = \text{no. of PCA factors} \)
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 $N^{\text{th}}$ PCA factor:
   \[
   \frac{N^{\text{th}} \text{ eigenvalue}}{\text{sum of all eigenvalues}} \times 100\%
   \]
4. Percentage of total variance captured by first $N$ PCA factors:
   \[
   \frac{\text{sum of eigenvalues up to } N}{\text{sum of all eigenvalues}} \times 100\%
   \]
Contents

1. Introduction
2. Identification
   - Principal component analysis (PCA)
   - **PCA walkthrough**
   - Data preprocessing
   - PCA examples
   - Multivariate curve resolution (MCR)
   - MCR examples
3. Quantification and prediction
4. Classification
5. Conclusion
Eight library spectra:

- PS 2480
- PS 3550
- PMMA 2170
- PMMA 2500
- PEG 1470
- PEG 4250
- PPG 425
- PPG 1000

Unit mass binned and mean centered prior to analysis.

Calculation using MATLAB with PLS Toolbox 4.0

---

**PCA walkthrough**

**Data:** modeled (calibration set)
- **Var:** Data
- **Size:** 8 x 300
- **Samp Lbls:**
- **Var Lbls:**

**Model:** calibrated on loaded data
- **Type:** PCA (3 PCs)
- **Preprocessing:** Mean Center
- **Data:** 8 x 300

**Number PCs:** 3

<table>
<thead>
<tr>
<th>Principal Component</th>
<th>Eigenvalue</th>
<th>% Variance</th>
<th>% Variance</th>
<th>% Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.72e-001</td>
<td>54.32</td>
<td>54.32</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4.21e-001</td>
<td>29.60</td>
<td>83.92</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.26e-001</td>
<td>15.87</td>
<td>99.79</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2.18e-003</td>
<td>0.15</td>
<td>99.95</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4.17e-004</td>
<td>0.03</td>
<td>99.98</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.74e-004</td>
<td>0.02</td>
<td>99.99</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7.79e-005</td>
<td>0.01</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
Perform scree test using the log eigenvalue plot
PCA walkthrough

First PCA factor (PC1):

- Variables/Loadings Plot for Data
- Samples/Scores Plot for Data
Second PCA factor (PC2):
Third PCA factor (PC3):
Biplot - PCA factor 1 against PCA factor 2:
Using PCA we have effectively reduced 300 correlated variables (mass units) to 3 independent variables (factors) by which all the samples can be characterised.
1. Introduction
2. Identification
   • Principal component analysis (PCA)
   • PCA walkthrough
   • Data preprocessing
   • PCA examples
   • Multivariate curve resolution (MCR)
   • MCR examples
3. Quantification and prediction
4. Classification
5. Conclusion
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 the nature of variance in data
- Can distort interpretation and quantification

- Includes:
  - Peak selection and binning
  - Centering
    - Mean centering
  - Scaling
    - Normalisation
    - Variance scaling
    - Poisson scaling
    - Binominal scaling
Peak selection and binning

**Manual selection**
Peaks of interest only
Unexpected features lost

**Auto peak search**
All peaks of interest included?
What threshold to use?

**Unit mass binning**
Straightforward to use but
detailed information lost

**0.5 u binning**
Separates organic from inorganics

Peak selection and binning

Manual selection
Peaks of interest only
Unexpected features lost

Auto peak search
All peaks detected
What threshold?

Unit mass binning
Straightforward to use but detailed information lost

0.5 u binning*
Separates organic from inorganics

Important considerations

- **What information are we putting into PCA?**
  What is included? What is omitted?

- **Do we need to apply further processing**
  e.g. dead time correction?

Mean centering

\[ \tilde{X}_{ik} = X_{ik} - mean(X_{ik}) \]

- 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
Normalisation

\[ \tilde{X}_{ik} = \frac{1}{\text{sum}(X_{i:})} \times X_{ik} \]

- Divide each spectrum by a constant for each sample e.g. intensity of a specific ion, total ion intensity
- Assumes chemical variances can be described by relative changes in ion intensities
- Preserves the shape of spectra
- Reduces effects of topography, sample charging, changes in primary ion current
Variance scaling

\[ \tilde{X}_{ik} = \frac{1}{\text{var}(X_{ik})} \times X_{ik} \]

- Divide each variable by its standard deviation 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)

• SIMS data is dominated by Poisson counting noise – statistical uncertainty of a peak is proportional to intensity
• The noise becomes binomial for saturated data with dead time correction
• Divide data by the estimated noise variance of each data point
• Emphasises weak peaks which vary above the expected counting noise, over intense peaks varying solely due to counting statistics
• Provides better noise rejection in PCA
Contents

1. Introduction
2. Identification
   • Principal component analysis (PCA)
   • PCA walkthrough
   • Data preprocessing
   • PCA examples
   • Multivariate curve resolution (MCR)
   • MCR examples
3. Quantification and prediction
4. Classification
5. Conclusion
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
- Loadings consistent with total amino acid composition of various proteins
- 95% confidence limits provide means for identification / classification

‘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.

Prior to $K$, mass peaks

$J$, columns

$I$, rows

unfold
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

PC1 scores

PC1 loadings

$^{35}\text{Cl} + ^{37}\text{Cl}$

$^{1}\text{st}$ factor distinguishes PVC and PC phases

PC2 scores

PC2 loadings

$^{2}\text{nd}$ factor shows detector saturation for intense $^{35}\text{Cl}$ peak

PCA image example (2)

Image courtesy of Dr Ian Fletcher
Intertek MSG

Total Spectra

Mass, u

72 ion images (out of > 400!)

NPL
PCA image example (2)

Hair fibre with multi-component pretreatment

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

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.

\[ X = TP' + E \]

- Data matrix
- Projection of samples onto factors (scores matrix)
- Projection of factors onto variables (loadings matrix)
- Residuals (noise)

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

1. Introduction
2. Identification
   • Principal component analysis (PCA)
   • PCA walkthrough
   • Data preprocessing
   • PCA examples
   • Multivariate curve resolution (MCR)
   • MCR examples
3. Quantification and prediction
4. Classification
5. Conclusion
Multivariate curve resolution (MCR)

- PCA factors are directions that describe variance
  - positive and negative peaks in the loadings
  - can be difficult to interpret
- What if we want to resolve original chemical spectra and reverse the following process?

\[
\text{Data matrix} = \begin{bmatrix}
9 & 32 & 10 & 1 & 21 \\
18 & 20 & 22 & 4 & 12 \\
24 & 12 & 30 & 6 & 6 \\
\end{bmatrix}
\]

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

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

- Try multivariate curve resolution (MCR)!
Multivariate curve resolution (MCR)

\[ X = TP' + E \]
\[ (I \times K) = (I \times N)(N \times K) + (I \times K) \]

Data matrix

Projection of samples onto factors (scores matrix)

Residuals (noise)

Projection of factors onto variables (loadings matrix)

\[ I = \text{no. of samples} \]
\[ K = \text{no. of mass units} \]
\[ N = \text{no. of factors} \]

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 assumes linear combination of chemical spectra (loadings) and contributions (scores) – only an approximation in SIMS.
Multivariate curve resolution (MCR)

\[ \mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} \]

Data matrix

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

Projection of samples onto factors (scores matrix)

Residuals (noise)

Projection of factors onto variables (loadings matrix)

MCR uses an iterative least-squares algorithm to extract solutions, while applying suitable constraints

With non-negativity constraint, MCR factors resemble SIMS spectra and chemical contributions more directly, as these must be positive
Outline of MCR

Raw Data

Data Matrix $X$

$X = TP' + E$

Initial Estimates of $T$ or $P$
- Random initialisation
- PCA loadings or scores
- Varimax rotated PCA loadings or scores
- Pure variable detection algorithm e.g. SIMPLISMA

MCR alternating-least-squares optimisation

Reproduced Data Matrix $\tilde{X}$
- Noise filtered data
- Ensures MCR solution is robust

Constraints
- Non-negativity
- Equality

Number of Factors

MCR Scores $T$

MCR Loadings $P$

Convergence criterion
- Non-negativity
- Equality

Constraints
- Non-negativity
- Equality

PCA
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 (‘selectivity’)

- Good initial estimates, suitable data preprocessing and correct number of factors are essential
1. Introduction
2. Identification
   - Principal component analysis (PCA)
   - PCA walkthrough
   - Data preprocessing
   - PCA examples
   - Multivariate curve resolution (MCR)
   - MCR examples
3. Quantification and prediction
4. Classification
5. Conclusion
MCR calculations using Matlab with MCR-ALS toolbox, freely available from http://www.mcrals.info/

Simple PVC / PC polymer blend
MCR image example (1)

MCR calculations using Matlab with MCR-ALS toolbox, freely available from http://www.mcrals.info/

MCR scores (‘pure component concentration’) – these will be ‘folded’ to form projection images

MCR loadings (‘pure component spectra’)
**Simple PVC / PC polymer blend**

- MCR extracts two distinctive factors, corresponding to PVC and PC respectively.
- Straightforward interpretation.

Loadings on MCR factor 1

Scores on MCR factor 1

Loadings on MCR factor 2

Scores on MCR factor 2

J L S Lee et al., *Surf. Interface Anal.* **2008**, *40*, 1-14
MCR loadings resemble SIMS spectra (characteristic peaks A-E) and fragments, and scores directly reveal spatial distributions!

• 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)
• 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!
MCR spectra example

- 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 summary

\[ \mathbf{X} = \mathbf{TP}' + \mathbf{E} \]

* Data matrix
  * Projection of samples onto factors (scores matrix)
  * Projection of variables onto factors (loadings matrix)
  * Residuals (noise)

- MCR describes the original data using factors, consisting of loadings and scores which resemble 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.
## Identification summary

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Manual analysis</th>
<th>PCA</th>
<th>MCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ease of interpretation</td>
<td>Easy – Single ion images</td>
<td>Medium / Difficult – Abstract, orthogonal factors</td>
<td>Easy – Non-negative scores and loadings</td>
</tr>
<tr>
<td>Chemical identification</td>
<td>Difficult – Characteristic peaks only</td>
<td>Medium – Important peaks and correlation</td>
<td>Easy – Full spectra obtained</td>
</tr>
<tr>
<td>Detection of minor components</td>
<td>Difficult – Only if substance is known</td>
<td>Easy – Higher factors capture small variance</td>
<td>Difficult ? – Possibly depend on system studied</td>
</tr>
<tr>
<td>Most suitable for</td>
<td>Simple dataset with good prior knowledge</td>
<td>Discrimination of similar chemical phases</td>
<td>Identification for unknown mixtures</td>
</tr>
</tbody>
</table>
1. Introduction
2. Identification
3. Quantification and prediction
   • Partial least squares regression (PLS)
   • Calibration, validation and prediction
   • PLS examples
4. Classification
5. Conclusion
Data analysis

Identification

What chemicals are on the surface?
Where are they located?

SIMS Dataset

Classification

Which group does it belong to?

What are the differences between groups?

Calibration / Quantification

How is it related to known properties?
Can we predict these properties?
We use regression analysis to find a predictive relationship between two variables.

\[ y = b^*x + e \]

- \( y \): Response variable
- \( b^* \): Regression coefficient
- \( x \): Predictor variable
- \( e \): Error term

Graph showing a linear relationship with the equation \( y = 0.6944^*x \).
Can we predict the properties of similar materials from their SIMS spectra?

\[ y = f(x) + e \]
\[ y = b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_m x_m + e \]

'Same' variable

- i.e. measured property

Regression coefficient

- 'Predictor' variable
- i.e. intensity at mass \( m \)
Multivariate regression

Extending to \( I \) samples and \( M \) response variables

\[
Y = XB + E
\]

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

1. We can calculate \( B \) to gain an understanding of the covariance relationship between \( X \) and \( Y \)
   e.g. relating SIMS spectra with sample preparation parameters

2. \( B \) can be applied to future samples in order to predict \( Y \) using only measurements of \( X \)
   e.g. quantifying surface composition or coverage of samples using only their SIMS spectra
Partial least squares regression (PLS)

\[ Y = XB + E \]
\[ (I \times M) = (I \times K)(K \times M) + (I \times M) \]

- Partial least squares regression (PLS) is a multivariate regression method for data \( X \) containing a large number of strongly correlated variables.

- PLS finds factors (called ‘latent variables’) that successively accounts for the largest covariance between \( X \) and \( Y \)
  - Removes redundant information from the regression i.e. information describing \( X \) that has no correlation with \( Y \)
  - Higher PLS factors that describe little covariance between \( X \) and \( Y \) can be discarded.

- The regression vectors \( B \) are a linear combination of PLS loadings that best predict \( Y \) from \( X \)

**Important to determine the number of factors to include in PLS!**
Calibration
Fit a PLS model to a calibration data set with known $\mathbf{X}$ and $\mathbf{Y}$

Validation
Apply model to an independent validation data set with known $\mathbf{X}$ and $\mathbf{Y}$, and calculate error between predicted $\mathbf{Y}$ and measured $\mathbf{Y}$

Prediction
Apply model to future samples to predict $\mathbf{Y}$ using only measurements of $\mathbf{X}$

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

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

Use cross-validation to determine number of factors

Use validation to determine prediction error
Number of factors – cross validation

- PLS can be used to build predictive models (calibration)
- Cross-validation can be used to determine the number of factors and guard against over-fitting

Good predictive model

Data is overfitted!
Number of factors – cross validation

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)

- ‘Leave one out’ cross validation most popular
  - Calculate PLS model excluding sample \( i \), for \( N \) PLS factors
  - Use model to predict sample \( i \) and calculate error
  - Repeat for all different samples
  - Calculate root mean square error of cross validation (RMSECV)
  - Repeat for different number of PLS factors
Validation and prediction

- Validation data should be statistically independent from calibration data. e.g. data taken on a different batch of samples, on a different day.
- Calculate RMSEP (Root Mean Square Error of Prediction)
- Independent validation set is essential if we want to use model to predict new samples!
1. Introduction
2. Identification
3. Quantification and prediction
   • Partial least squares regression (PLS)
   • Calibration, validation and prediction
   • PLS examples
4. Classification
5. Conclusion
• 13 SIMS spectra of thin films of Irganox were compared with their thicknesses measured with XPS
• Both $\mathbf{X}$ and $\mathbf{Y}$ are mean centered before analysis, and peaks below 50 $\mu$ were removed
• Two PLS factors are retained, explaining 99.8% of the variance in $\mathbf{X}$ (SIMS data) and 98.8% of the variance in $\mathbf{Y}$ (thicknesses)

Calculation using MATLAB with PLS Toolbox 4.0
• Leave-one-out cross validation shows 2 PLS factors to include in model

• Scores and loadings on PLS factor 1 and factor 2 are calculated
- PLS model able to predict thicknesses for $t < 6$ nm
- PLS regression vector shows Irganox characteristic peaks most correlated with thickness
- Irganox dewets on the surface so initial thickness is proportional to surface coverage!
• 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 $X$

Hydrocarbons
e.g. $C_nH_m^+$
hydrophobic

Polar species
e.g. $C_nH_mO^+$ & CN$^-$
hydrophilic

• PLS is a multivariate linear regression technique
• PLS find factors that best describe the structure of covariance between X and 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 spectra
1. Introduction
2. Identification
3. Quantification and prediction
4. Classification
   - PCA classification
   - Principal Component Discriminant Function Analysis (PC-DFA)
   - Partial Least Squares Discriminant Analysis (PLS-DA)
5. Conclusion
Data analysis

Identification
- What chemicals are on the surface?
- Where are they located?

SIMS Dataset

Calibration / Quantification
- How is it related to known properties?
- Can we predict these properties?

Classification
- Which group does it belong to?
- What are the differences between groups?
PCA classification

- 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

**Discriminant analysis**

**PCA**
Describes the spread of the data in order of importance:

**Discriminant analysis**
Describes the difference between the smiley and the heart:
Example 1 – PC-DFA

- 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

Example 2 – PLS-DA

Example: Hyperaccumulator plant which stores high levels of Ni in epidermal cells. What is the difference between epidermal and other cells?

PCA

PC1, 2 and 3 overlay

PC10 scores and loadings

PC 10 describes differences between epidermal cells and other areas – but this is not efficient!
Example 2 – PLS-DA

Step 1 – Define epidermal cells

Step 2 – PLS-DA factors: The largest **co-variance** between the data and the group assignments (0 or 1)

![Images of LV1, LV2, LV3](image1)

PLS-DA prediction

Regression vector

Step 3 – Regression vector: Combination of peaks which best predicts the differences between epidermal and other cells

[Image courtesy of Dr Kat Smart and Prof Chris Grovenor at the University of Oxford](image2)
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!
Contents

1. Introduction
2. Identification
3. Quantification and prediction
4. Classification
5. Conclusion
Data analysis

PCA, MCR
Identification

What chemicals are on the surface?
Where are they located?

SIMS Dataset

PLS
Calibration / Quantification
How is it related to known properties?
Can we predict these properties?

PC-DFA, PLS-DA
Classification

Which group does it belong to?
What are the differences between groups?
In this lecture we looked at
- Identification, quantification & classification using multivariate analysis
- Importance of validation for predictive models
- Data preprocessing techniques and their effects

http://www.npl.co.uk/nanoanalysis/chemometrics.html
- Copy of tutorial slides
- Draft terminology for ISO18115
- MVA vocabulary

http://mvsa.nb.uw.edu/
Community website with tutorials, links and software
Developed by Dan Graham and hosted by NESAC/BIO

Surface Analysis: The Principal Techniques 2nd edition, Chapter 10 “The application of multivariate data analysis techniques in surface analysis”
General

- S. Wold, Chemometrics; what do we mean with it, and what do we want from it?, Chemom. Intell. Lab. Syst. 30 (1995) 109
- 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


MCR


PLS