A Guide to the Practical Use of Multivariate Analysis in SIMS

J LS Lee, I S Gilmore
National Physical Laboratory, Teddington, UK

A copy of these slides can be downloaded from http://www.npl.co.uk/nanoanalysis/chemometrics.html
Why are we here?

Number of Publications

- **PCA**
- **MCR**
- **PLS**
- **DFA**
- **ANNs**

Year of Publication

- 1990
- 2000
- 2010
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
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?

- 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
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!

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

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

Each row (spectra) is represented by a vector
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

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

Samples

Chemical spectra

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

Variables [mass]
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. 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!

\[
\begin{bmatrix}
9 & 32 & 10 & 1 & 21 \\
18 & 20 & 22 & 4 & 12 \\
24 & 12 & 30 & 6 & 6 \\
\end{bmatrix}
= 
\begin{bmatrix}
5 & 1 \\
2 & 4 \\
0 & 6 \\
\end{bmatrix}
\times 
\begin{bmatrix}
1 & 6 & 1 & 0 & 4 \\
4 & 2 & 5 & 1 & 1 \\
\end{bmatrix}
\]
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*!

<table>
<thead>
<tr>
<th>Data matrix</th>
<th>=</th>
<th>“Scores”</th>
<th>×</th>
<th>“Loadings”</th>
</tr>
</thead>
<tbody>
<tr>
<td>[9 32 10 1 21]</td>
<td>=</td>
<td>[3.6 1.9]</td>
<td>×</td>
<td>[4.7 5.7 5.7 1.0 3.5]</td>
</tr>
<tr>
<td>[18 20 22 4 12]</td>
<td>=</td>
<td>[3.7 -0.2]</td>
<td>×</td>
<td>[-3.9 5.9 -5.3 -1.3 4.4]</td>
</tr>
</tbody>
</table>
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

- 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?
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 underlying dimension that contributes to summarising or accounting for the original data set</td>
<td>Principal Component</td>
<td>Pure Component</td>
<td>Latent Vectors, Latent Variables</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)

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

\(I = \text{no. of samples} \)  \(K = \text{no. of mass units} \)  \(D = \text{dimensionality of data} \)
Principal component analysis (PCA)

- PCA extract 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
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!

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

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

- PCA reproduced data matrix
- Scores matrix
- Loadings matrix

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

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
Perform scree test using the log eigenvalue plot
First PCA factor (PC1):
Second PCA factor (PC2):
Third PCA factor (PC3):
PCA walkthrough

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
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

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

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

1\textsuperscript{st} factor goes from origin to centre of gravity of data
1\textsuperscript{st} factor goes from origin and accounts for the highest variance
Normalisation

- 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

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

Poisson and binomial scaling

- 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.

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

• 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

PCA image analysis

- ‘Datacube’ contains a raster of \( I \times J \) pixels and \( K \) mass peaks
- The datacube is rearranged into a 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
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 results after Poisson scaling and mean centering

1st factor distinguishes PVC and PC phases

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

PCA image example (2)

InterTek
Image courtesy of
Dr Ian Fletcher
Intertek MSG

72 ion images (out of > 400!)

Total Spectra

Mass, u

50μm
Hair fibre with multi-component pretreatment

PCA scores

PCA loadings

Factor 1

Factor 2

Factor 3

Factor 4

Factor 5

mass (arb. scale)

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

PCA summary

\[ X = TP' + E \]

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

- 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. 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?

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

\[
\text{Sample composition} = \begin{bmatrix}
5 & 1 \\
2 & 4 \\
0 & 6 \\
\end{bmatrix}
\]

\[
\text{Chemical spectra} = \begin{bmatrix}
1 & 6 & 1 & 0 & 4 \\
4 & 2 & 5 & 1 & 1 \\
\end{bmatrix}
\]

- Try multivariate curve resolution (MCR)!
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)

\[ X = TP' + E \]

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

Data matrix

Projection of samples onto factors (scores matrix)

Projection of factors onto variables (loadings matrix)

Residuals (noise)

\( I = \text{no. of samples} \)

\( K = \text{no. of mass units} \)

\( N = \text{no. of factors} \)

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

\[ X = TP' + E \]

**Raw Data**

**Data Matrix** \( X \)

**PCA**

**Number of Factors**

**MCR alternating-least-squares optimisation**

**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

**Reproduced Data Matrix** \( \hat{X} \)
- Noise filtered data
- Ensures MCR solution is robust

**Constraints**
- Non-negativity
- Equality

**Convergence criterion**
- Non-negativity
- Equality

**MCR Scores** \( T \)

**MCR Loadings** \( P \)
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
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
MCR image example (1)

Simple PVC / PC polymer blend

- MCR extracts two distinctive factors, corresponding to PVC and PC respectively.
- Straightforward interpretation.
MCR image example (2)

Image courtesy of Dr Ian Fletcher
Intertek MSG

72 ion images (out of > 400!)

Total Spectra

Mass, u
MCR loadings resemble SIMS spectra (characteristic peaks A-E) and fragments, and scores directly reveal spatial distributions!

MCR scores
- Factor 1
- Factor 2

MCR loadings
- Factor 3
- Factor 4
- Factor 5

Mass (arb. scale)

MCR image example (3)

- 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 summary

\[ X = TP' + E \]

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

- MCR describes the original data using factors, consisting of loadings and scores 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.

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

PC-DFA, PLS-DA

Classification

Which group does it belong to?
What are the differences between groups?

PLS

Calibration / Quantification

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

SIMS

Classification / differences

Can we known properties?
Is it related to Quantification?
Conclusion

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
- Tutorial slides
- Multivariate analysis terminology defined in ISO 18115-1

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

Surface and Interface Analysis
Multivariate Analysis special issues (Volume 41 Issue 2 & 8, Feb/Aug 2009)

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

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