National Physical Laboratory

Proposed terminology for Multivariate Analysis in Surface Chemical Analysis – Vocabulary – Part 1: General Terms and Terms for the Spectroscopies

J L S Lee, I S Gilmore and M P Seah (Email: joanna.lee@npl.co.uk)

National Physical Laboratory, Teddington, Middlesex, UK
08 September 2008

Introduction

As a result of its history and scope, multivariate analysis is laden with confusing terminology, with different names given to similar or equivalent terms depending on the technique and the field of application. A well-defined terminology is essential for ideas and practices to be communicated clearly and accurately. In surface chemical analysis many terms have been defined in ISO standards. ISO 18115:2001 [1, 2], ISO 18115Amd1:2006 [3, 4] and ISO 18115Amd2:2007 [5] currently cover the surface analytical spectroscopies and scanning probe microscopies with some 775 terms. These are currently being restructured into a Part 1 for the surface analytical spectroscopies [6] and a Part 2 for the scanning probe microscopies [7]. In Part 1, are currently the following 23 terms for multivariate analysis, these are listed in Table 1 in three broad categories. Each definition in the vocabulary is followed by a list of terms that are closely associated and also by relevant 'Notes' that provide helpful information. Terms used in the definition and defined elsewhere in the Vocabulary are indicated in bold. These definitions have been developed in close consultation with international experts, and the vocabulary continues to be refined with new terms added. Experts in the user community are greatly encouraged to become involved in the development of the vocabulary and can do so by contacting the authors.

References

[1] ISO 18115:2001, Surface chemical analysis – Vocabulary, ISO, Geneva
[2] M. P. Seah, Surf. Interface Anal. 2001; 31, 1048–1049
[3] ISO 18115 Amd 1:2006, Surface chemical analysis – Vocabulary Amendment 1, ISO, Geneva
[4] M. P. Seah, Surf. Interface Anal. 2007; 39, 367–369.
[5] ISO 18115 Amd 2:2007, Surface chemical analysis – Vocabulary Amendment 2, ISO, Geneva.
[6] ISO 18115:Part 1:2007, Surface chemical analysis – Vocabulary – Part 1: General terms and terms for the spectroscopies, in draft.
[7] ISO 18115:Part 2:2007, Surface chemical analysis – Vocabulary – Part 2: Terms for scanned probe microscopies, in draft.
[8] J. L. S. Lee, B. J. Tyler, M. S. Wagner, I. S. Gilmore and M. P. Seah, Surf. Interface Anal. in press

Table 1 - List of new multivariate analysis terms

	General Term		Multivariate method		Preprocessing
1	multivariate analysis	10	factor analysis	16	data preprocessing
2	samples	11	principal component analysis	17	centering
3	variables	12	multivariate curve resolution	18	scaling
4	data matrix	13	maximum autocorrelation factors	19	transformation
5	reproduced data matrix	14	partial least squares	20	normalization
6	residual matrix	15	discriminant analysis	21	variance scaling
7	factor			22	auto scaling
8	loadings			23	Poisson scaling
9	scores				

1
multivariate analysis
MVA
analysis involving a simultaneous statistical procedure for two or more dependent **variables**

NOTE 1 An essential of aspect of multivariate analysis is the dependence between different **variables**, which may involve their covariance. Multivariate analysis simplifies the interpretation of complex data sets involving a large number of dependent variables by summarizing the data using a smaller number of statistical variables.

NOTE 2 Multivariate analysis methods fall into two broad categories: unsupervised or exploratory methods and supervised methods. Unsupervised methods are used to identify trends in a data set, key differences between **samples** and key co-variances between spectral features. These methods include **factor analysis, PCA**, and **MCR**. Supervised methods are used for prediction, modeling, calibration and classification. These methods include PCR, **PLS**, and **DFA**.

2
samples
<multivariate analysis> a series of individual measurements made on one or more experimental systems
c.f. **variables**

NOTE 1 Data from each sample occupies a row in the **data matrix**

NOTE 2 The term ‘sample’ in **multivariate analysis** is not to be confused with the conventional use of the word in practical analysis, meaning a physical entity that is under measurement. In multivariate analysis, each ‘sample’ simply denotes an independent measurement. This could be repeat measurements of the same physical sample, measurements of different physical samples, or a combination of both.

3
variables
<multivariate analysis> a series of channels or parameters over which experimental measurements are made on the **samples**
c.f. **samples**

NOTE 1 Data from each variable occupies a column in the **data matrix**

NOTE 2 In **SIMS**, the variables refer to the mass or time-of-flight of secondary ions, and in **XPS** the variables refer to the binding energies of photoelectrons detected.

4
data matrix
table of numbers with *I* rows and *K* columns, containing experimental data obtained for *I* **samples** over *K* values of one or more **variables**, where *I* and *K* are integers

NOTE 1 'S**ample** denote any individual measurements made on a system and **variable** denote the channels over which the measurements are made. For example, in **SIMS**, the variables refer to the mass or time-of-flight of secondary ions, and in **XPS** the variables refer to the binding energies of photoelectrons detected.

NOTE 2 For a multivariate image with dimensions of *I* pixels x *J* pixels x *K* variables, the data is often 'unfolded' prior to **multivariate analysis** to form a data matrix with dimensions *IJ* x *K*. On completion of the analysis, the results can be 'folded' to restore the original image dimensions.

5
data matrix, reproduced
<factor analysis> the product of the **scores** matrix and the transpose of the **loadings** matrix in a **factor analysis** model

NOTE 1 The reproduced data matrix is the difference between the **data matrix** and the **residuals matrix** for a given **factor analysis** model

NOTE 2 The reproduced data matrix is often considered to be the noise-filtered approximation of the **data matrix**. This is true if the **residual matrix** is assumed to contain noise only.

6
residual matrix
<factor analysis> the difference between the **data matrix** and the **reproduced data matrix** for a given **factor analysis** model

NOTE 1 The residual matrix contains data that are not described by the **factor analysis** model, and is usually assumed to contain noise.

7
factor
component (deprecated)
pure component (deprecated)
principal component (deprecated)
<factor analysis> axis in the data space of a **factor analysis** model, representing an underlying dimension that contributes to summarising or accounting for the original data set

NOTE 1 In **PCA** each factor is called a "principal component". The first PCA factor is called "PC1". This is deprecated where **PCA** is used along other **factor analysis** techniques such as **MCR** when it becomes clearer to refer to "PCA factor 1" and "MCR factor 1".

NOTE 2 In **MCR** each factor is called a "pure component". The term "component" and "pure component" is deprecated as it may be confused with real chemical components of the system.

NOTE 3 Each factor is associated with a set of **loadings** and **scores**, which occupies a column in the **loadings** and **scores** matrices respectively.

8
loadings
principal component spectrum (deprecated)
pure component spectrum (deprecated)
<factor analysis> projection of the **variables** onto the **factors**, reflecting the covariance relationship between **variables**
c.f. **scores**

NOTE 1 “loadings” (plural) refers to a whole column in the loadings matrix that relates to a particular **factor**. “loading” (singular) is the particular contribution of a **variable** in the original space to the **factor**.

NOTE 2 In **PCA**, the loadings are also the cosine angles between the **variables** and a particular **factor**.

NOTE 3 In **MCR**, the term "pure component spectrum" is interchangeable with the term "loading" and is therefore deprecated. The term, in spectroscopy, may be confused with the spectrum for a pure material.

NOTE 4 When analyzing multivariate spectral data such as those obtained from **SIMS** or **XPS**, the **loadings** for a **factor** can be interpreted as a “pseudo-spectra” and can be used to develop a chemical or physical interpretation to that **factor**. Since misinterpretation of these pseudo-spectra is a common caveat, it is important to verify any interpretation with the original data.

9
scores
projections (deprecated)
pure component concentration (deprecated)
<factor analysis> projection of the **samples** onto the **factors**, reflecting the relationship between **samples**
c.f. **loadings**

NOTE 1 “scores” (plural) refers to a whole column in the scores matrix that relates to a particular **factor**. “score” (singular) is the projection of a particular **sample** onto the **factor**.

NOTE 2 In **MCR**, the term "pure component concentration" is interchangeable with the term "MCR score" and is therefore deprecated. The term, in spectroscopy, may be confused with the concentration for a pure material.

NOTE 3 When analyzing multivariate spectral data such as those obtained from **SIMS** or **XPS**, the **scores** for a **factor** can be interpreted as a “pseudo-contribution” for the chemical or physical phenomena associated with that factor. There is not necessarily a simple linear relationship between the scores and real physical and chemical properties such as concentration. Calibration standards are essential between attempting to use the scores quantitatively, and any patterns observed in the scores must be tested for statistical significance by the proper use of replicates, cross validation and other statistical tests.

10
factor analysis
matrix decomposition of the **data matrix** (X) into the product of the **scores** matrix (T) and the transpose of the **loadings** matrix (P′), together with a **residual matrix** (E), with the aims of describing the underlying structure of the data set using **factors** in order to reduce the dimensionality of the data

NOTE 1 Hence X = TP′ + E.

NOTE 2 Factor analysis methods include **PCA**,**MCR** and **MAF**.

NOTE 3 The number of **factors** selected in factor analysis is smaller than the rank of the **data matrix**.

NOTE 4 Factor analysis is equivalent to a rotation in data space where the **factors** form the new axes. This is not necessarily rotation that that maintains orthogonality except in the case of **PCA**.

NOTE 5 The residual matrix contains data that are not described by the factor analysis model, and is usually assumed to contain noise.

11
principal component analysis
PCA
principal components analysis (deprecated)
factor analysis involving the extraction of orthogonal **factors** that successively capture the largest amount of variance in the data set
c.f. **MAF**

NOTE 1 PCA **factors** are eigenvectors of the matrix Z, where Z is the matrix transpose of the **data matrix** multiplied by the **data matrix** itself. The **data matrix** may be with or without **data preprocessing**. PCA factors are sorted by their associated eigenvalues in descending order. Eigenvalues are the amount of variance described by their associated factor.

NOTE 2 PCA has found extensive use in exploring differences in a series of **SIMS** spectra. It is useful, for example, in identifying trends and clusters, discriminating similar materials and detecting small variations, identifying spectral components associated with selected chemical functional groups, and analysis of spectral changes within a depth profile.

NOTE 3 PCA is useful for analysis of individual **SIMS** images and can aid in identifying and enhancing contrast between chemically different regions in an image, and identifying the spectral components associated with image features.

12
multivariate curve resolution
MCR
self modelling mixture analysis (SMMA) (deprecated)
self modelling curve resolution (SMCR) (deprecated)
factor analysis for the decomposition of multi-component mixtures into a linear sum of chemical components and contributions, when little or no prior information about the composition is available

NOTE 1 MCR **factors** are extracted by the iterative minimisation of the **residual matrix** using alternating least squares (ALS), while applying suitable constraints, such as non-negativity, to the **loadings** and **scores**. MCR may be performed on the **data matrix** with or without **data preprocessing**.

NOTE 2 For each **data matrix**, MCR factors are not unique but are dependent on initial estimates, the number of factors to be resolved, constraints applied and convergence criterion.

NOTE 3 MCR with non-negativity constraints is used in **SIMS** and **XPS** to obtain **loadings** and **scores** that resemble physically meaningful chemical component spectra and contributions, which must have positive values. However, the assumption of linearity is only a first approximation, and neglects non-linear effects which may be important in practical analysis, such as matrix effects, topography and detector saturation.

13
maximum autocorrelation factors
MAF
factor analysis for multivariate images involving the extraction of **factors** that successively capture the largest amount of variance across the entire image while minimizing the variation between neighboring pixels
c.f. **PCA**

NOTE 1 MAF **factors** are the eigenvectors of matrix B, where B is the matrix transpose of the **data matrix** multiplied by the **data matrix** itself, all pre-multiplied by the inverse of the covariance matrix of the shift images. The shift images are obtained by subtraction the **data matrix** by a copy of itself that has been shifted by one pixel.

NOTE 2 MAF is useful for analysis of individual **SIMS** images and can aid in identifying and enhancing contrast between chemically different regions in an image, and identifying the spectral components associated with image features.

NOTE 3 **Loadings** obtained from MAF is independent of data **scaling**.

NOTE 4 MAF can be extended to the analysis of three-dimensional images obtained from **SIMS** imaging depth profile.

14
partial least squares
partial least squares regression
PLS
PLSR (deprecated)
a linear multivariate regression method for assessing relationships among two or more sets of **variables** measured on the same entities.

NOTE 1 PLS finds factors (latent variables) in the observable variables that explain the maximum variance in the predicted variables, using the simultaneous decomposition of the two. It removes redundant information from the regression, i.e. factors describing large amounts of variance in the observed data that does not correlates with the predictions.

NOTE 2 PLS can be used for calibration and quantification, provided that an independent validation data set is used assess the accuracy of the prediction and guard against the over-fitting of the model to the calibration data. In the absence of an independent validation set, cross validation can be useful into determining the number of PLS factors to retain in the model. However, any predictions made on independent samples must then be treated with caution.

15
discriminant analysis
discriminant function analysis
DA
DFA
a supervised multivariate technique for classifying **samples** into predefined groups using discriminant functions

NOTE 1 Discriminant functions are **factors** that maximises the variance between different groups while minimising the variance within each group. **Loadings** on DFA **factors** can be used to provide information on the combination of **variables** is best for predicting group membership.

NOTE 2 DFA is often applied after **PCA** for a multivariate data set. This removes collinearity from the multivariate data and ensures that the new predictor variables, which are PCA **scores**, are distributed normally. This method is referred to as principal component-discriminant function analysis (PC-DFA).

NOTE 3 DFA can be used for calibration and prediction, provided that an independent validation data set is used assess the accuracy of the prediction and guard against the over-fitting of the model to the calibration data. In the absence of an independent validation set, cross validation can be useful. However, any predictions made on independent samples must then be treated with caution.

16
data preprocessing
data pretreatment (deprecated)
manipulation of raw data prior to a specified data analysis treatment

NOTE 1 The terms "preprocessing" and "pretreatment" are often used interchangeably but the latter is deprecated to reduce confusion with sample preparation / treatment prior to experimental analysis.

NOTE 2 Aside from the three main categories of data preprocessing methods (**centering**, **scaling** and **transformation**), data preprocessing can refer to any other procedures carried out on the raw data, including mass binning and peak selection. In the case of multivariate images, this can also include region-of-interest selection and image filtering or binning.

NOTE 3 All data preprocessing methods imply some assumptions about the nature of the variance in the data set. It is important that these assumptions are understood and appropriate for the data set involved.

NOTE 4 More than one data preprocessing method can be applied to the same data set. The order of data preprocessing is important and can affect assumptions made on the nature of variance in the data set.

17
centering
mean centering
centring (deprecated)
mean centring (deprecated)
<data preprocessing> a **data preprocessing** procedure in which each **variable** in the **data matrix** is centered by the subtraction of its mean value across all **samples**
c.f. **scaling**, **transformation**

NOTE 1 Mean centering emphasises the differences between **samples** rather than differences between the **samples** and the origin.

NOTE 2 Mean centering is generally recommended for **PCA**, PLS and discriminant analysis of **SIMS** and **XPS** data, where relative intensities of peaks across the **samples** are more important than their absolute deviation from zero intensities. Mean centering is not compatible with non-negativity constraints in **MCR** for the resolution of physically meaningful component spectra and contributions, which must have positive values.

NOTE 3 Mean centering is generally applied after other **data preprocessing** methods, including data selection and **scaling**.

NOTE 4 All **data preprocessing** methods imply some assumptions about the nature of the variance in the data set. It is important that these assumptions are understood and appropriate for the data set involved.

18
scaling
weighting
<data preprocessing> a **data preprocessing** procedure in which the **data matrix** is divided elementwise by a scaling matrix.
c.f. **centering**, **transformation**

NOTE 1 Common methods of data scaling are **normalization**, **variance scaling**, **auto scaling**, and in the case of **SIMS** data, **Poisson scaling**.

NOTE 2 Data scaling can affect both the total variance of the data and the relative variance contained in each **variable**, and may introduce bias in the analysis of data.

NOTE 3 Data scaling is generally applied after appropriate data selection, prior to the **centering** of the data.

NOTE 4 All **data preprocessing** methods imply some assumptions about the nature of the variance in the data set. It is important that these assumptions are understood and appropriate for the data set involved.

19
transformation
<data preprocessing> a **data preprocessing** procedure in which each element in the **data matrix** is transformed by a defined function
c.f. **centering**, **scaling**

NOTE 1 Examples of defined functions are logarithm and square root.

NOTE 2 Transformation by a linear function is equivalent to **scaling** and **centering**.

NOTE 3 All **data preprocessing** methods imply some assumptions about the nature of the variance in the data set. It is important that these assumptions are understood and appropriate for the data set involved.

20
normalization
<data preprocessing> a **scaling** method used in **data preprocessing** in which the scaling matrix consists of a constant for each **sample**
c.f. **variance scaling**, **auto scaling**, **Poisson scaling**

NOTE 1 The scaling constant could be the value of a specific **variable**, the sum of selected variables or the sum of all variables for the **sample**.

NOTE 2 Normalization is commonly used in **SIMS** to compensate for differences in **total ion yield** that arise from instrumental conditions, since the relative intensities within a **SIMS** spectrum are more repeatable than absolute intensities.

NOTE 3 All **data preprocessing** methods imply some assumptions about the nature of the variance in the data set. It is important that these assumptions are understood and appropriate for the data set involved.

21
scaling, variance
<data preprocessing> a **scaling** method used in **data preprocessing** in which the scaling matrix consists of the standard deivation of each **variable** across the **samples**
c.f. **auto scaling**, **normalization**, **Poisson scaling**

NOTE 1 Variance scaling is referred to as **auto scaling** when combined with **mean centering**

NOTE 2 Variance scaling equalizes the importance of each **variable** in **multivariate analysis**, and is commonly applied in conjunction with peak selection in **SIMS**.

NOTE 3 All **data preprocessing** methods imply some assumptions about the nature of the variance in the data set. It is important that these assumptions are understood and appropriate for the data set involved.

22
scaling, auto
<data preprocessing> a **data preprocessing** method involving the application of **variance scaling** followed by **mean centering**.
c.f. **variance scaling**, **normalization**, **Poisson scaling**

NOTE 1 Auto scaling equalizes the importance of each **variable** in **multivariate analysis**, and is commonly applied in conjunction with peak selection in **SIMS**.

NOTE 2 All **data preprocessing** methods imply some assumptions about the nature of the variance in the data set. It is important that these assumptions are understood and appropriate for the data set involved.

23
Poisson scaling
Keenan scaling (deprecated)
Optimal scaling (deprecated)
<data preprocessing> a **scaling** method used in the **data preprocessing** of data based on Poisson statistics, in which the scaling matrix consists of the outer product of two vectors, containing the square root of the mean sample intensity and the square root of the mean spectrum, respectively
c.f. **normalization**, **variance scaling**

NOTE 1 Poisson scaling is only valid for **SIMS** and **XPS** raw data where the detector is operating within linearity, and cannot be applied in conjunction with other data **scaling** methods.

NOTE 2 Poisson scaling is shown to improve the results obtained in many **multivariate analysis** of **SIMS** data, including **PCA** and **MCR**, by scaling the data such that each element of the **data matrix** has the same experimental uncertainty. In **SIMS**, this experimental uncertainty can be dominated by the Poisson counting statistics of the detector, such that high intensity peaks have a higher absolute measurement uncertainty than low intensity peaks. Poisson scaling weights each peak in each spectrum (i.e. each element of the **data matrix**) by this uncertainty, which is estimated from the raw data, using the fact that uncertainty arising from Poisson statistics is equal to the average counted intensity.

NOTE 3 Poisson scaling is especially valuable for **ToF-SIMS** images, which has low counts per pixel and can therefore be dominated by Poisson counting noise.

NOTE 4 In Poisson scaling it is customary to transform the results obtained by **multivariate analysis**, such as **loadings** and **scores** from **PCA** and **MCR**, from the Poisson scaled space back to the original physical space, by the multiplication of the original scaling vectors.