IUPAC Recommendations

David B. Hibbert*

Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016)

DOI 10.1515/pac-2015-0605
Received June 20, 2015; accepted March 4, 2016

Abstract: Recommendations are given concerning the terminology relating to chemometrics. Building on ISO definitions of terms for basic concepts in statistics the vocabulary is concerned with mainstream chemometric methods. Where methods are used widely in science, definitions are given that are most useful to chemical applications. Vocabularies are given for general data processing, experimental design, classification, calibration and general multivariate methods.

Keywords: calibration; chemometrics; data analysis; design of experiments; experimental design; IUPAC Analytical Chemistry Division; multivariate analysis; pattern recognition; regression; terminology.

CONTENTS

1 PREFACE 407
2 INTRODUCTION 408
3 DATA, SAMPLING AND DATA PROCESSING 408
4 EXPERIMENTAL DESIGN 413
5 MULTIVARIATE METHODS AND RELATED CONCEPTS 417
6 CLASSIFICATION 426
7 CALIBRATION AND REGRESSION 433
8 INDEX OF TERMS 437
9 INDEX OF ABBREVIATIONS 441
10 MEMBERSHIP OF SPONSORING BODIES 442
REFERENCES 442

1 Preface

The recommendations contained in this document concern the terminology relating to concepts in chemometrics. It recognises the existence of ISO Standards on terms used in statistics and probability [1] and applied statistics [2], and has not attempted to redefine basic concepts in statistics. See ISO 3454 [1, 2] and the IUPAC Green Book [3] for general rules on symbols and terminology in mathematics and statistics.

Generic quantities are denoted by upper-case letters, and individual values ('best estimates' in a mathematical framework) by the corresponding lower-case letter. In a measurement model, Y denotes the measurand, $X_1, \ldots, X_N$ the input quantities, and $y, x_1, \ldots, x_N$ the corresponding best estimates.

Article note: This work was started under project 2008-002-1-500: A glossary of concepts and terms in chemometrics, with membership D Brynn Hibbert, Pentti Minkkinen and Barry Wise. Public input was via an open wiki that was active from 2010 to 2012 [D. B. Hibbert, P. Minkkinen, N. M. Faber, B. M. Wise. *Anal. Chim. Acta* **642**, 3 (2009)].

***Corresponding author: David B. Hibbert,** School of Chemistry, UNSW Australia, Sydney, NSW 2052, Australia, e-mail: b.hibbert@unsw.edu.au

The compilation has drawn on existing standards and literature and has been the subject of consultation with the chemometrics community by the establishment of a wiki in 2010 (closed 2012) [4].

Where a definition from another work is used in its entirety the reference includes the item number (*e.g.* [5] 6.11 refers to entry 6.11 in ISO 18115-1:2010). When a specific item number is absent, the reference indicates the source of the inspiration for the present definition. However basic definitions from statistics given, for example, in ISO 3534 are not reproduced here.

These Recommendations will become part of a chapter in the revised Orange Book (Compendium of Terminology in Analytical Chemistry, 3rd edition), which will include a complete list of definitions, and further elaboration of concepts.

2 Introduction

The term 'chemometrics' was first used by Svante Wold in 1971 and the International Chemometrics Society was formed in 1974 by Svante Wold (Umeå University, Sweden) and Bruce Kowalski (University of Washington, Seattle) [6, 7]. In a now historically-significant paper to the Journal of Chemical Information and Computer Sciences [8], Kowlski reproduced a letter, signed by himself and Wold to a "Prospective Chemometrician". In it chemometrics is defined as "… the application of mathematical and statistical tools to chemistry." The definition given below at 2.1 is the latest refinement, maintaining brevity and highlighting the practical nature of chemometrics (see Note 2 in 2.1).

There has been no complete vocabulary of chemometrics, the nearest being a web site, now defunct, by Vandeginste [9], and some extended glossaries in books. Terms have been defined as the subject evolved, sometimes leading to different terms for the same concept in different fields of chemistry (for example spectroscopy and bioinformatics). The approach taken here is to offer definitions that have gained some acceptance, not favouring any particular section of the chemometrics community.

2.1 chemometrics

The science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods.

Note 1: Data treated by chemometrics are often *multivariate*.

Note 2: Although in some cases the mathematical and statistical techniques used in chemometric applications might be the same as those used in theoretical chemistry, it is important to emphasize that chemometrics should not involve theoretical calculations, but should deal primarily with the extraction of useful chemical information from measured data.

Note 3: Chemometrics is widely applied outside chemistry, *e.g.* in biology, metabolomics, engineering as well as sub-disciplines such as forensics, cultural studies *etc*.

Reference: [10]

3 Data, sampling and data processing

3.1 autocorrelation

Correlation of a variable with itself over a successive time or space intervals (lags).

Note 1: If the mean of autocorrelated data is estimated the standard deviation of the mean depends on sampling mode.

3.2 autoscaling

Variance scaling of mean-centered data.
Note: See *mean centering*

3.3 categorical data

Data, values of which are one of a fixed number of nominal categories.
Note 1: Data in a *contingency table* is categorical.

3.4 contingency table
cross tabulation

Type of table in a matrix format that displays the multivariate frequency distribution of variables.
Note: The entries in the cells of a contingency table can be frequency counts or relative frequencies. See: *categorical data*

3.5 data matrix

Measurement results on a system arranged in a $m \times n$ matrix, with m objects and n variables.
Note 1: By convention a matrix is arranged with m rows and n columns.
Note 2: Objects are also called 'samples', but confusion with physical 'test samples' in analytical chemistry should be avoided.
Note 3: Variables are also called features, or explanatory variables.
Note 4: For *multi-way data* the data is arranged in a hypercube of $m \times n_1 \times n_2 \times \ldots$, where n_1, n_2 … are the kinds of explanatory variable.

3.6 data pre-processing

Manipulation of *raw data* prior to a specified data analysis treatment.
Note 1: The term "pre-processing" is preferred to the term "pre-treatment" to reduce confusion with physical sample preparation or treatment prior to experimental analysis.
Note 2: Aside from the three main categories of data pre-processing methods (*mean centering*, *scaling* and *transformation*), data pre-processing 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 pre-processing methods imply some assumptions about the nature of the variability 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 pre-processing method can be applied to the same data set. The order of data pre-processing is important and can affect assumptions made on the nature of variance in the data set.

Reference: [11] 6.3

3.7 dynamic time warping

Process of synchronizing a data matrix so that it represents the same time shifts.
Note 1: The method is used in chromatography to align peaks by their retention times.

3.8 evaluation data

validation data
deprecated test data
deprecated test set
deprecated prediction data
deprecated prediction set

Data used to validate a *model*.
Note 1: Evaluation data should be independent of the data used to calibrate or train a model. See *cross validation*, *training data*.
Note 2: 'Test data' is also used for data from an unknown sample, and should not be used for 'evaluation data'.

3.9 explanatory variable

Variable that influences the value of a response variable, and is used to build models of the response variable.

3.10 exploratory data analysis

initial data analysis
EDA

Summary of the main characteristics of data, often using graphical methods.
Note 1: Exploratory data analysis is recommended before deciding an approach to chemometric modelling.

Reference: [12]

3.11 mean centering

centering

Data pre-processing in which the mean value of a variable is subtracted from data across all objects.

$$a^{*}_{i,j} = a_{i,j} - \frac{1}{n}\sum_{i=1}^{n} a_{i,j}$$

Note 1: Mean centering emphasises the differences between samples rather than differences between the samples and the variable's origin (zero).
Note 2: Mean centering is generally recommended for *principal component analysis*, *partial least squares* and *discriminant analysis* of data, where relative values across the samples are more important than their absolute deviation from zero. Mean centering is not compatible with non-negativity constraints in, for example, *multivariate curve resolution*.
Note 3: Mean centering is generally applied with other *data pre-processing* methods. See *scaling*.

Reference: [11]

3.12 multiplicative scatter correction

MSC
multiplicative signal correction

Data pre-processing in which a constant and a multiple of a reference data set is subtracted from data.
Note 1: MSC is typically used in near-infrared spectrometry to remove effects of non-homogeneous particle size [13].

3.13 multivariate data

Data having two or more variables per object.
Note 1: The measurements results are often of the same *kind of quantity*.
Example 1: Absorbances measured at 101 wavelengths in the range 200 nm atond 400 nm.
Example 2: Mass fractions of 10 elements measured by ICP-MS.

3.14 multi-way data

N-way data

Multivariate data having two or more kinds of *explanatory variable* per object.
Note 1: For two groups of explanatory variables the data is termed 'three-way'.
Note 2: Models that decompose multi-way data include *PARAFAC*, *Tucker3 model*.
Example 1: Fluorescence intensities with excitation wavelength and emission wavelength representing the two variable axes, and the objects making the third direction.

3.15 noise

Response that gives no information.

3.16 normalization (in data pre-processing)

row scaling

Scaling method in which the scaling matrix consists of a single value for each object.
Note 1: The scaling value could be the value of a reference variable, the sum of selected variables or the sum of all variables for the sample.
Note 2: 'Normalization' has many meanings in statistics (see https://en.wikipedia.org/wiki/Normalization_(statistics)). To resolve any ambiguity the nature of the scaling constant should be explained.

Reference: *variance scaling*, *autoscaling* [11]

3.17 random sampling

Sampling in which the sample locations are selected randomly from the whole population.
Note 1: The population defines the kind of quantity of the location. For example, in a time series, the quantity giving the location is time, in *QSAR* the location is a point in design space.

3.18 raw data

primary data

Data not yet subjected to analysis.
Note 1: Raw data can be an *indication* obtained from a measuring instrument or measuring system (VIM 4.1)

3.19 sampling

Selection of a subset of individuals from within a population to estimate characteristics of the whole population.

3.20 sampling error

The difference between an estimate of a parameter obtained from a sample and the population value.
Note 1: Unless the population values have been measured, sampling error cannot be directly estimated.

3.21 sampling unit

A defined quantity of material having a boundary which may be physical or temporal.
Note 1: Examples of physical boundaries are capsules, containers, and bottles.
Note 2: A number of sampling units may be gathered together, for example in a package or box.

3.22 scaling

weighting

Element-wise division of a *data matrix* by a scaling matrix.
Note: See *variance scaling, autoscaling, normalization*

3.23 smoothing

Transformation using an approximating function to capture important patterns in data, while removing *noise* or other fine-scale structures.
Note 1: Examples of smoothing functions are moving average, and Savitzky-Golay.

3.24 systematic sampling

Sampling in which individual samples are taken at equal intervals in location.
Note 1: In a time series, the quantity giving the location is time.
Note 2: The starting point may be assigned randomly within the first stratum.

3.25 take-it-or-leave-it data

TILI

Happenstance data that must be processed, or not processed, as is.
Reference: [14]

3.26 training data
training set

Data used for creating a model in *supervised classification*.
Note: See *evaluation data*.

3.27 transformation

Application of a deterministic mathematical function to each point in a set of data.
Note 1: Mathematically each data point z_i is replaced with the transformed value $y_i = f(z_i)$, where $f(.)$ is a mathematical function.
Note 2: Transforms may be applied so that the data appear to more closely meet the assumptions of a statistical inference procedure that is to be applied, or to improve the interpretability or appearance of graphs.
Note 3: *Smoothing* is an example of a transformation.
Example: $f(x) = \log(x)$.

3.28 variance scaling

Scaling in which the scaling matrix is the standard deviation of each variable across the objects.

$$a^*_{i,j} = \frac{a_{i,\mathrm{j}}}{s_j}$$

Note 1: A variable occupies a column of the data matrix
Note 2: Variance scaling equalizes the importance of each variable in *multivariate data*.
Note 3: When used with *mean centering* variance scaling is known as *autoscaling*.

Reference: [11] 6.20

4 Experimental design

Experimental design has become an important step in investigating the effects of factors on systems. Traditional approaches to optimisation in which one factor at a time is considered, while maintaining other factors constant, has been shown to be inefficient and, for correlated factors, incapable of producing the optimum [15]. The definitions here mostly differ from, but do not contradict, those in ISO 3534-3 [16]

4.1 alias structure

List of combinations of *effects* that are aliased (confounded).
Note: See *aliased effects*

4.2 aliased effects
confounded effects

In a *fractional-factorial design*, *effects* for which the information obtained are identical.
Note 1: In a two-level design, the product of the coded levels for the aliased effects are equal.
Example 1: If there are four factors in a design: A, B, C, D then the main effect of A can be aliased with the three way effect $B \times C \times D$. So for the run that has B = –1, C = –1, D = +1, then A must = +1.

4.3 coded experimental design

coded design

Matrix of runs by *factor levels* in which each level is denoted by a code that represents the relative magnitude of the level.

Example 1: A two-level design is coded –1 and +1, a three level design is coded –1, 0 and +1.

Example 2: In a *rotatable central composite design* for 3 factors the coded levels are $-\sqrt{2}$, –1, 0, +1, $+\sqrt{2}$.

4.4 effect of a factor

effect

Coefficient of a term in a response *model*.

Note: See *main effect*, *n^{th}-order effect*, *interaction effect*.

4.5 design matrix

Matrix with rows representing individual experimental treatments (possibly transformed according to the assumed model) which can be extended by deduced levels of other functions of *factor levels*.

Reference: [16] 3.2.25

4.6 dummy factor

Factor that is known to have no effect on the *response*, used in an *experimental design*, to estimate repeatability standard deviation.

Example: A factor having levels '+' singing the first verse of the National Anthem at the experiment, and '–' singing the second verse of the National Anthem at the experiment.

4.7 experimental design

design of experiments

DoE

Efficient procedure for planning combinations of values of *factors* in experiments so that the data obtained can be analyzed to yield valid and objective conclusions.

Note 1: Experimental design is applied to determine the set of conditions that are required to obtain a product or process with desirable, often optimal properties. A characteristic of experimental design is that these conditions are determined in a statistically-optimal way.

Note 2: Response surface methodology is considered an important part of experimental design.

Note 3: An 'experimental design' (noun) usually refers to a table giving the levels of each factor for each run. See *coded experimental design*.

Reference: [17, 18]

4.8 factor (experimental design)

Input quantity in a *model*.

Note 1: The term has a different meaning when used in factor analysis.

4.9 factor level

level

Value of a *factor* in an *experimental design*.

Note 1: A design may be designated by the number of levels chosen for each factor, as in "two-level design".

Note 2: When writing an experimental design the levels are usually coded. (See *coded experimental design*).

4.10 fractional-factorial design

deprecated: incomplete-factorial design

Experimental design obtained from a full factorial design in which experiments are systematically removed to fulfil stated statistical requirements.

Note 1: The aim of a fractional design is to reduce the number of experiments by confounding low-order effects (*e.g.* main effect, two-way interaction) with high order interactions, which are assumed to be small.

Note 2: A design, having L^k (see *full factorial design*) experiments, is fractionated to L^{k-p} experiments where p is an integer $<k$.

Note 3: The choice of design is governed by an *alias structure*.

Note 4: A fractional factorial design is incomplete, but all incomplete designs are not fractional factorial. See *Plackett Burman design*.

4.11 full-factorial design

Experimental design with all possible combinations of *factor levels*.

Note 1: If there are k factors, each at L levels, a full factorial design has L^k runs.

4.12 interaction effect

Effect of a factor where the term is the product of two or more factors.

Example 1: The yield of a synthesis is modelled in terms of the temperature T and concentration of a reactant c
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 T + \hat{\beta}_2 c + \hat{\beta}_{11} T^2 + \hat{\beta}_{22} c^2 + \hat{\beta}_{12} Tc$. The estimated coefficient $\hat{\beta}_{12}$ is the interaction effect of T and c.

Note: See *main effect, n^{th}-order effect*.

4.13 main effect

Effect of a factor where the term is a single factor.

Example 1: The yield of a synthesis is modelled in terms of the temperature T and concentration of a reactant c
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 T + \hat{\beta}_2 c + \hat{\beta}_{11} T^2 + \hat{\beta}_{22} c^2 + \hat{\beta}_{12} Tc$. The estimated coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are the main effects of T and c respectively.

Note: See *n^{th}-order effect, interaction effect*.

4.14 model (experimental design)

Equation describing the *response* as a function of values of the *factors*.

Note 1: The model can be based on knowledge of the chemistry or physics of the system, but usually the model is empirical, being linear or quadratic with interaction terms.

Note 2: To obtain information about the significance of effects, data is usually *mean centered* and assessed against a *coded experimental design*.

Example 1: The yield of a synthesis is modelled in terms of the temperature T and concentration of a reactant c

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 T + \hat{\beta}_2 c + \hat{\beta}_{11} T^2 + \hat{\beta}_{22} c^2 + \hat{\beta}_{12} Tc$$

Note: See *n^{th}-order interaction effect*, *main effect*

4.15 n^{th}-order effect

Effect of a factor where the term is a factor raised to the power n.

Example 1: The yield of a synthesis is modelled in terms of the temperature T and concentration of a reactant c

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 T + \hat{\beta}_2 c + \hat{\beta}_{11} T^2 + \hat{\beta}_{22} c^2 + \hat{\beta}_{12} Tc$. The coefficients $\hat{\beta}_{11}$ and $\hat{\beta}_{22}$ are the second-order effects of T and c respectively.

Note: See *main effect*, *interaction effect*.

4.16 optimization

Minimization or maximization of a real function by systematically choosing the values of real or integer variables from within an allowed set.

4.17 Plackett-Burman design

Incomplete *experimental design* to estimate *main effects* for which each combination of *factor levels* for any pair of *factors* appears the same number of times.

Note 1: Plackett-Burman designs are typically given for two levels, with a number of experiments that is a multiple of 4 but not a power of 2. (The latter case is a *fractional factorial design*).

Note 2: For $4 \times N$ experiments, $4 \times N - 1$ main effects and the mean are estimated.

Note 3: If less than $4 \times N - 1$ factors are being studied, *dummy factors* are inserted which allow estimation of the repeatability standard deviation of the measurements.

Example: The coded experimental design for $4 \times N = 12$, where +1 and –1 represent the two levels of factors $X_1 \ldots X_{11}$ is given below. The order of performing the runs should be randomised.

Run	Pattern	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1	++++++++++	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
2	−+−+++−−−+−	−1	+1	−1	+1	+1	+1	−1	−1	−1	+1	−1
3	−−+−+++−−−+	−1	−1	+1	−1	+1	+1	+1	−1	−1	−1	+1
4	+−−+−+++−−−	+1	−1	−1	+1	−1	+1	+1	+1	−1	−1	−1
5	−+−−+−+++−	−1	+1	−1	−1	+1	−1	+1	+1	+1	−1	−1
6	−+−−+−+++−	−1	−1	+1	−1	−1	+1	−1	+1	+1	+1	−1
7	−−−+−−+−+++	−1	−1	−1	+1	−1	−1	+1	−1	+1	+1	+1
8	+−−−+−−+−++	+1	−1	−1	−1	+1	−1	−1	+1	−1	+1	+1
9	++−−−+−−+−+	+1	+1	−1	−1	−1	+1	−1	−1	+1	−1	+1
10	+++−−−+−−+−	+1	+1	+1	−1	−1	−1	+1	−1	−1	+1	−1
11	−+++−−−+−−+	−1	+1	+1	+1	−1	−1	−1	+1	−1	−1	+1
12	+−+++−−−+−−	+1	−1	+1	+1	+1	−1	−1	−1	+1	−1	−1

4.18 resolution of a design

resolution

One more than the smallest n^{th}-*order interaction effect* that some *main effect* is aliased with.

Note 1: Resolution is used to describe the extent to which *fractional factorial designs* create *aliased effects*.

Note 2: The resolution is written as a Roman numeral.

Note 3: *Full factorial designs* have no effects that are aliased and therefore have infinite resolution.

Example 1: For resolution III designs the main effects are aliased with two-factor interactions. For resolution IV designs no main effects are aliased with two-factor interactions, but two-factor interactions are aliased with each other. For resolution V designs no main effect or two-factor interaction is aliased with any other main effect or two-factor interaction, but two-factor interactions are aliased with three-factor interactions.

Reference: [19]

4.19 response

Measured or observed quantity in an *experimental design*.

4.20 response surface methodology

Experimental design in which the *response* is modelled in terms of one or more *factor levels*.

Note 1: Response surface methodology is usually associated with *optimization*. The model used is typically a quadratic function leading to a maximum or minimum response in the factor space.

Note 2: The term 'surface' implies two factors and a single response, when a plot of the modelled response as a function of values of the factors leads to a surface in the three dimensional space. This can be generalized to any number of factors.

Reference: [20]

5 Multivariate methods and related concepts

5.1 alternating least squares regression

alternating regression

ALS

Solution to the *multivariate* decomposition of a data matrix in which iteratively a solution of one output matrix is used to compute the second matrix, after the application of constraints.

Note: ALS is used to decompose multiple spectra ($\boldsymbol{X}$) into concentration ($\boldsymbol{C}$) and component spectra ($\boldsymbol{S}$). $\boldsymbol{X} = \boldsymbol{C}\,\boldsymbol{S}^{\mathrm{T}}$. Because spectra and concentrations cannot be negative at each iteration negative values are set to zero.

5.2 autoregression

A stochastic process in which future values are estimated based on a weighted sum of past values.

Note: A process called AR(1) is a first order process, meaning that the current value is based on the immediately preceding value. An AR(2) process has the current value based on the previous two values.

5.3 biplot

Combination plot of a *scores plot* as points and *loadings plot* as vectors for common *factors*.
Note 1: The plots are scaled to facilitate interpretation.
Note 2: Points in the scores plot (objects) that fall on a *loadings* vector are considered to be characterised by the variable associated with the vector.

5.4 bootstrapping

Estimation of parameters by multiple re-sampling from measured data to approximate its distribution.
Note 1: Multiple resamples of the original data allow calculation of the distribution of a parameter of interest, and therefore its standard error.
Example 1: The standard error of an estimate of parameter θ
$\hat{s}_E = \frac{1}{B}\sum_{i=1}^{B}(\theta_i^* - \overline{\theta}^*)^2$ where B is the number of bootstrap samples, θ_i^* the i-th bootstrap estimate, and $\overline{\theta}^*$ the mean value of the bootstrap estimates.
Note 2: Random sampling with replacement is used when the data are assumed to be from an independent and identically-distributed population.
Note 3: Bootstrapping is an alternative to *cross validation* in model validation.

Reference: [21, 22]

5.5 canonical variables

Linear combinations of data with the greatest correlation.
Note: See: *canonical variate analysis*.

5.6 canonical variate analysis

canonical analysis

Multivariate technique which finds linear combinations of two sets of data that are most highly correlated.
Note 1: The combinations with the greatest correlation, denoted U1 and V1 are known as the "first canonical variables".
Note 2: The relationship between the canonical variables is known as the canonical function.
Note 3: The next canonical functions, U2 and V2 are then restricted so that they are uncorrelated with U1 and V1. Everything is scaled so that the variance equals 1.

5.7 common factor analysis

exploratory factor analysis
factor analysis

Factor analysis in which *latent variables* are calculated that maximise the correlation with observed variables.
Note 1: The common factors are not unique. Typically factors are rotated so that the factors are more easily interpreted in terms of the original variables.

5.8 core consistency diagnostic (CONCORDIA)

Method to assess the appropriateness of a *PARAFAC* model.

Note 1: An appropriate PARAFAC model is a model where the components primarily reflect low-rank, trilinear variation in the data.
Note 2: The principle of the method is to assess the degree of superdiagonality of the model.

Reference: [23]

5.9 correspondence factor analysis

correspondence analysis

Factor analysis applied to categorical data in which orthogonal factors are obtained from a contingency table.

5.10 cross validation

A re-sampling procedure that predicts the class or property of objects from a classification or regression model that is obtained without those observations.
Note 1: When a single object is removed, the procedure is known as *leave-one-out cross validation*. When n/G objects are deleted, the procedure is known as G-*fold cross validation*.
Note 2: The procedure is iterated leaving out all the objects in turn.
Note 3: The model is assessed by calculation of the *root mean square error of prediction* for continuous variables, and by the *misclassification probability* for classification.
Note 4: Use of independent *evaluation data* is preferred to cross validation, when there is concern about the independence of the objects in the data set.
Note 5: Cross-validation can be used with *bootstrapping*, one to optimize a model (*e.g.* how many PCs are appropriate) and the other for validation.

5.11 evolving factor analysis (EFA)

Factor analysis that follows the change or evolution of the rank of the data matrix as a function of an ordered variable.
Note 1: The ordering variable may be time. (see [24])
Note 2: The changing rank is calculated by *principal-component analysis* on an increasing data matrix.

Reference: [25]

5.12 factor (factor analysis)

deprecated component
deprecated pure component

Axis in the data space of a *factor analysis* model, representing an underlying dimension that contributes to summarizing or accounting for the original data set.
Note 1: In *principal component analysis* each factor is called a *principal component*. It is deprecated when used outside this context. To avoid confusion "principal component factor" is recommend by ISO.
Note 2: In *multivariate curve resolution* each factor is called a "pure component". The terms "component" and "pure component" are deprecated as they may be confused with 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.

5.13 factor analysis

Matrix decomposition of a *data matrix* ($\boldsymbol{X}$) into the product of a *scores* matrix ($\boldsymbol{T}$) and the transpose of the *loadings* matrix ($\boldsymbol{P}^{\mathrm{T}}$).

Note 1: Hence $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E}$, where $\boldsymbol{E}$ is a residual matrix.

Note 2: Factor analysis methods include *common factor analysis* (also called 'factor analysis') *principal component analysis,* and *multivariate curve resolution*.

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 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*.

Reference: [11] 6.5

5.14 *G*-fold cross validation

Cross validation of a data set of N objects in which N/G objects are removed at each iteration of the procedure.

Note 1: Objects 1 to N/G are removed on the first iteration, then objects $N/G+1$ to $2N/G$ after replacement of the first N/G objects, and so on.

Note 2: Because the perturbation of the model is larger than in leave-one-out cross validation, the prediction ability of the G-fold cross validation is less optimistic than obtained with *leave-one-out cross validation*.

5.15 latent variable

latent construct
hidden variable

Variable that is inferred through a mathematical model from other variables that are observed.

Note 1: The *factors* obtain from *common factor analysis* are termed latent variables.

Note 2: A distinction can be made between 'hidden variable', which is considered to be an actual variable that is buried in the effects of other variables and *noise*, and a 'latent variable' that is entirely hypothetical.

5.16 leave-one-out cross validation (LOOCV)

Cross validation in which one object is removed in each iteration of the procedure.

5.17 loadings

deprecated principal component spectrum
deprecated pure component spectrum

Projection of a *factor* onto the *variables*.

Note 1: 'Loadings' (plural) refers to a 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: The loadings on a factor reflect the relationships between the variables on that factor. (See *score*)

Note 3: In *principal component analysis* the loadings are also the cosine angles between the variables and a particular factor.
Note 4: In *multivariate curve resolution* 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.

Reference: [11] 6.7

5.18 loadings plot

Plot of one *loading* against variable number, or two or three loadings against each other.
Note 1: Usually the loadings associated with the early *factors* (1, 2, 3) are plotted to reveal relationships among the variables.
Note 2: See: *loadings plot, biplot*

5.19 maximum likelihood principal component analysis (MLPCA)

Principal component analysis that incorporates information about measurement uncertainty to develop models that are optimal in a maximum likelihood sense.
Reference: [26].

5.20 mean squared error of prediction (MSEP)

mean squared error of estimation (MSEE)

In *multivariate calibration* the average of the squared deviation of estimated values from the values of *evaluation data*.
For *N evaluation data* where c_i is an observed value and $\hat{c}_i$ is the predicted value

$$MSEP = \frac{\sum_{i=1}^{i=N} (\hat{c}_i - c_i)^2}{N}$$

Note 1: mean squared error of prediction is the square of root mean squared error of prediction.

5.21 multivariate curve resolution (MCR)

deprecated self-modelling curve resolution (SMCR)
deprecated self-modelling mixture analysis (SMMA)

Factor analysis for the decomposition of multicomponent data into a linear sum of chemically-meaningful components when little or no prior information about the composition is available.
Note 1: MCR factors are extracted by the iterative minimization of the residual matrix using an *alternating least squares* approach, while applying suitable constraints, such as non-negativity, to the loadings and scores. MCR can be performed on the data matrix with or without data pre-processing.
Note 2: MCR factors are not unique but are dependent on initial estimates, the number of factors to be resolved, constraints applied and convergence criteria. MCR factors are not required to be orthogonal.

5.22 non-linear iterative partial least squares (NIPALS)

Iterative decomposition of a *data matrix* to give *principal components*.

Note 1: Writing the model as $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E}$, the first principal component is computed from a data matrix. The data explained by this PC are then subtracted from $\boldsymbol{X}$ and the algorithm applied again to residual data. The procedure is repeated until sufficient principal components are obtained.

Note 2: The algorithm is very fast if only a few principal components are required, because the covariance matrix is not computed.

5.23 nonlinear mapping (NLM)

Projection of objects defined in a multivariate space onto two- or three-dimensional space so that the distances between objects are preserved as well as possible.

Note 1: An often applied criterion for the mapping error (E) is the relative squared error between the true distance d_{ij} and mapped distance δ_{ij}.

Note 2: Several iterative optimization procedures can be applied to minimize E, such as steepest descent.

5.24 parallel factors analysis (PARAFAC)

canonical decomposition (CANDECOMP)

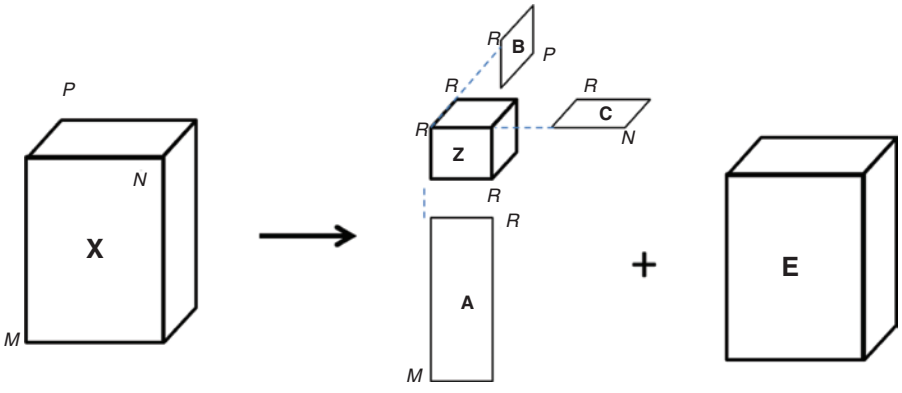
Decomposition of a three-way data matrix into the sum of sets of two-way loadings matrices.

Note 1: The PARAFAC model is also known as Canonical Decomposition (CANDECOMP).

Note 2: A representation of PARAFAC is $x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk}$, where x_{ijk} is i,j,k –th element of the data matrix, and a_{ir}, b_{jr}, c_{kr} are the components of the loadings matrices. e_{ijk} is the i,j,k –th element of residual matrix.

Note 3: PARAFAC is a special case of the Tucker3 model (see *Tucker tri-linear analysis*) where the core matrix is the identity matrix, and $r = s = t = R$

Note 4: A schematic representation of the PARAFAC model is



Reference: [27]

5.25 prediction error sum of squares (PRESS)

sum of squared errors of prediction (SSEP)
residual sum of squares (RSS)
sum of squared residuals (SSR)

In *multivariate calibration* for a prediction set of N data where c_i is an observed value and $\hat{c}_i$ is the predicted value.

$$PRESS = \sum_{i=1}^{i=N} (\hat{c}_i - c_i)^2$$

Note: See *root mean squared error of prediction*

5.26 principal component– discriminant analysis (PC-DA)

Discriminant analysis on a multivariate data set that has been subject to *principal component analysis*.

Note 1: This procedure removes collinearity from the multivariate data and ensures that the new predictor variables, which are PCA scores, are distributed normally.

5.27 principal-component analysis (PCA)

Factor analysis in which *factors* are calculated that successively capture the greatest variance in the data set.

Note 1: The factors are orthogonal and are known as *principal component factors*.

Note 2: The factorization is written $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E}$, where $\boldsymbol{T}$ is the *scores* matrix, $\boldsymbol{P}$ is the *loadings* matrix and $\boldsymbol{E}$ is a residual matrix. See *non-linear iterative partial least squares*.

Reference: [11] 6.15

5.28 principal-component factor

principal component (PC)

Orthogonal *factors* obtained in a *principal-component analysis*.

Note 1: The successive factors explain reducing fractions of the variance of the data set, and are written PC1, PC2 …

Note 2: ISO recommends use of the term principal-component factor.

5.29 Procrustes analysis

Comparison of shapes of multi-dimensional objects by a series of geometrical transformations to minimise the sum of squared distances between the transformed and target structures while maintaining the internal structure of the objects.

Note 1: For two objects defined by $\boldsymbol{X}$ and $\boldsymbol{Y}$ the manipulation orthogonal rotation/reflection matrix $\boldsymbol{R}$ Procrustes analysis minimises: $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{R}\|^2$ subject to $\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = 1$

Note 2: Ordinary, or classical, Procrustes analysis is when an object is compared to one other object, which may be a reference shape. Generalized Procrustes analysis compares three or more shapes to an optimally-determined mean shape.

Reference: [28] p 310

5.30 Quantitative structure-activity relationship (QSAR)

QSPR

QSA/PR

Relationships between chemical structure, or structural-related properties, and target property of studied compounds.

Note 1: Typical target property is biological (or therapeutic) activity of a drug.
Note 2: Typical structural-related properties are Hammett electronic parameter, lipophilicity parameter, boiling and melting points, molecular weight and molar refractivity.
Note 3: Relationships are established by *multivariate calibration*.

Reference:[28] Ch 37.

5.31 root mean square error of cross validation (RMSECV)

Root mean square error of prediction when the predicted data is obtained by *cross validation*.

5.32 root mean squared error of prediction (RMSEP)

root mean squared error of estimation (RMSPE)
standard error of prediction (SEP)
standard error of estimation

In *multivariate calibration* or *classification* for *N evaluation data* where c_i is an observed value and $\hat{c}_i$ is the predicted value

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{i=N} (\hat{c}_i - c_i)^2}{N}}$$

Note 1: RMSEP is related to the prediction error sum of squares (PRESS) by $RMSEP = \sqrt{\frac{PRESS}{N}}$
Note 2: For completely independent, normally distributed evaluation data, RMSEP is a measure of the bias of the calibration.
Note 3: When prediction is by *cross validation* RMSEP may be termed root mean square error of cross validation.

5.33 simple-to-use interactive self-modelling mixture analysis (SIMPLISMA)

Interactive method to obtain concentrations and pure spectra from spectra of mixtures using directly-measured variables.
Note 1: The directly-measured variables are called 'pure variables' in the method.
Note 2: A *data matrix* $\boldsymbol{D} = \boldsymbol{C} \times \boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E}$ where $\boldsymbol{C}$ is a concentration matrix, $\boldsymbol{P}$ pure spectra of mixture components and $\boldsymbol{E}$ an error matrix. Pure spectra are estimated $\hat{\boldsymbol{P}}$ which allows projection of a concentration matrix $\boldsymbol{C}^{\star}$ from which the data matrix can be reconstructed and compared with the measured spectra.

$$\begin{cases} \hat{\boldsymbol{P}} = \boldsymbol{D}^{T}\boldsymbol{C}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C})^{-1} \\ \boldsymbol{C}^{*} = \boldsymbol{D}\hat{\boldsymbol{P}}(\hat{\boldsymbol{P}}^{\mathrm{T}}\hat{\boldsymbol{P}})^{-1} \end{cases} \text{ then } \boldsymbol{D}_{\mathrm{recon}} = \boldsymbol{C}^{*} \times \hat{\boldsymbol{P}}^{\mathrm{T}}$$

Note 3: Second derivatives of spectra can be used for modelling.

Reference: [30]

5.34 score

deprecated projections
deprecated pure-component concentration

Factor analysis projection of an object onto a *factor*.

Note 1: In PCA, the factors are orthogonal and the scores are an orthogonal projection of the objects onto a factor.
Note 2: The scores on a factor reflect the relationships between objects for that factor. (See *loading*).
Note 3: The term scores (plural) refers to a whole column in the scores matrix that relates to a particular factor. The term score (singular) is the projection of a particular object onto the factor.

Reference: [11] 6.21

5.35 scores plot

Plot of one *score* against object number, or two or three scores against each other.
Note 1: Usually the scores associated with the early *factors* (1, 2, 3) are plotted to reveal relationships among the objects.
Note: See: *loadings plot, biplot*

5.36 simulated annealing

Generic probabilistic meta-heuristic to locate a good approximation to the global optimum of a given function in a large search space, in which there is a slow decrease in the probability of accepting worse solutions as the solution space is explored.
Note 1: The function E(s) to be minimized is analogous to the internal energy of the system in that state. The goal is to bring the system, from an arbitrary initial state, to a state with the minimum possible energy. At each step, the heuristic considers some neighbouring state s' of the current state s, and probabilistically decides between moving the system to state s' or staying in state s. These probabilities ultimately lead the system to move to states of lower energy. Typically this step is repeated until the system reaches a state that is good enough for the application, or until a given computation budget has been exhausted.

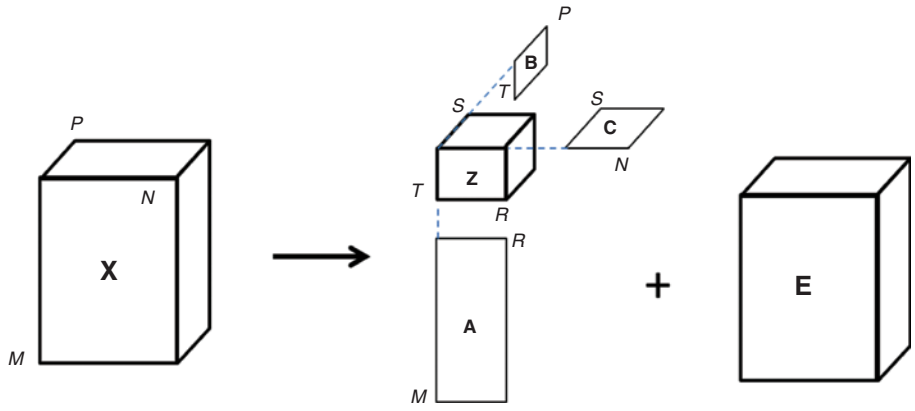
Reference: [31], http://en.wikipedia.org/wiki/Simulated_annealing

5.37 singular value decomposition
SVD

A factorization of an $m \times n$ matrix ($\boldsymbol{M}$) such that $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U}$ is an $m \times m$ matrix, $\boldsymbol{\Sigma}$ is a $m \times n$ matrix and $\boldsymbol{V}^{\mathrm{T}}$ is a $n \times n$ matrix.
Note 1: If $\boldsymbol{M}$ is a data matrix with m objects and n variables, the matrix $\boldsymbol{U}$ is the *scores* matrix, the diagonal of $\boldsymbol{\Sigma}$ contain the square roots of the eigenvalues and $\boldsymbol{V}$ is the *loadings* matrix.

5.38 Tucker tri-linear analysis
Tucker3 model

Decomposition of a three-way data matrix into a three-way core matrix, and three, two-way loadings matrices.
Note 1: A representation of the Tucker3 model is $x_{ijk} = \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{t=1}^{T} a_{ir} b_{js} c_{kt} z_{rst} + e_{ijk}$, where x_{ijk} is the data matrix, a_{ir}, $b_{js,}$ c_{kt} are the loadings matrices, and z_{rst} is the core matrix. e_{ijk} is the residual matrix.
Note 2: A graphical representation of the Tucker3 model is



Note 3: See *parallel factors analysis*

6 Classification

6.1 artificial neural network (ANN)

Computing system made up of a number of simple, highly interconnected elements, which process information by their dynamic state response to external inputs.

Note 1: An ANN is composed of layers of nodes with an input layer accepting data, one or more hidden layers computed from earlier layers, and an output layer giving the results of the classification.

Note 2: Nodes are connected by non-linear functions that calculate the contribution (weight) of an earlier node to a later node.

Reference: Definition adapted from [32] and quoted in [33] (See: http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html)

6.2 backward chaining

back chaining

Inference method from hypothesis to data that supports the hypothesis

Note 1: Backward chaining is used in *backward-propagation* to train an *artificial neural network*.

6.3 backward propagation

back-propagation learning rule
back-propagation of errors

Supervised classification method for an *artificial neural network* in which weights of connections between nodes are calculated from the known output layer back to the input layer.

Reference: [34]

6.4 backward stepwise linear discriminant analysis

Linear discriminant analysis in which variables to build the discriminant functions are removed one at a time to minimise the loss of discrimination, until there is a significant loss.

Note 1: Significant loss is tested by an F-test
Note 2: Interrelationships between variables that have not yet been selected are ignored, and variables already added, which may become largely redundant through subsequent additions, cannot be removed.
Note 3: See *forward stepwise linear discriminant analysis*.

6.5 Bayes classifier

Supervised classification that minimises the *misclassification probability*.
Note: Misclassification probability can be estimated as the frequency of misclassified objects, also known as the misclassification rate.

6.6 city-block distance

Manhattan distance
Hamming distance
taxi distance

Distance ($d_{i,j}$) between two objects (i and j) calculated as the sum of the absolute difference between k variables (x) that describe the objects.

$$d_{i,j} = \sum_{k=1}^{k=N} |x_{i,k} - x_{j,k}|$$

6.7 classification

Assignment of a series of objects to membership of groups.
Note 1: Classification is a particular example of *pattern recognition*. (See https://en.wikipedia.org/wiki/Pattern_recognition).
Note 2: The groups may be pre-defined (see *supervised classification*), or not (see *unsupervised classification*)

Reference: [35]

6.8 cluster

Region of high density of objects in a space based on their characterising data.
Note 1: Measures of density are the distance between objects in variable space, or the (dis)similarity of objects. See *Euclidean distance*, *city block distance*, *Tanimoto similarity index*.
Note 2: Several such regions of high density may exist indicating that the objects form groups with similar properties.
Note 3: Cluster analysis is a synonym for *classification*.

Reference: [28] Chapter 30

6.9 complete linkage

Linkage criterion in *hierarchical clustering* in which the distance between two clusters is the distance between those two objects (one from each cluster) that are farthest apart.
Note 1: Complete linkage tends to find compact clusters of approximately equal size.

6.10 dendrogram

Tree diagram used to illustrate the arrangement of clusters produced by *hierarchical clustering*.

Note 1: The objects are successively grouped (or un-grouped) in each layer of the dendrogram, where the length of the linking lines is proportional to the dissimilarity of the objects.

6.11 discriminant analysis (DA)

discriminant function analysis (DFA)

Supervised classification method in which functions of the observed variables are used to classify observations into designated groups.

Note 1: The classification functions are known as discriminant functions, discriminant criteria, or classification criteria. They maximise the variance between different groups while minimizing the variance within each group. Loadings on DA factors can be used to provide information on the combination of variables is best for predicting group membership.

Note 2: When the distribution within each group is assumed to be multivariate normal, a parametric method can be used to develop a discriminant function. The discriminant function is determined by a generalized squared distance. The classification criterion can be based on either the individual within-group covariance matrices (yielding a quadratic function) or the pooled covariance matrix (yielding a linear function).

Note 3: The model takes into account the prior probabilities of the groups, which can be taken as proportional to the number in each group or equal across all groups.

Note 4: L*inear discriminant analysis*, *quadratic discriminant analysis* and *regularized discriminant analysis* are types of discriminant analysis used in chemometrics.

Reference [11] 5.4.

6.12 disjoint principal component analysis

Principal component analysis independently performed on each class as a step in *classification*.

Note 1: Soft independent modelling of class analogy is an example of the use of disjoint principal component analysis.

Note 2: When a PCA model is obtained using all classes, it is known as conjoint PCA.

Reference: [36]

6.13 Euclidean distance

Distance ($d_{i,j}$) between two objects (i and j) calculated as the square root of the sum of the squared differences between k variables (x) that describe the objects.

$$d_{i,j} = \sqrt{\sum_{k=1}^{k=N} (x_{i,k} - x_{j,k})^2}$$

6.14 forward chaining

Inference method from data to hypothesis.

Note 1: In logic, forward chaining is the application of modus ponens (if P implies Q, given P then Q).

6.15 forward stepwise linear discriminant analysis

Linear discriminant analysis in which variables to build the discriminant functions are introduced one at a time to maximise the discrimination, until there is no significant improvement.

Note 1: Significant improvement is usually tested by an F-test

Note 2: Interrelationships between variables that have not yet been selected are ignored, and variables already added, which may become largely redundant through subsequent additions, cannot be removed.

Note 3: See *backward stepwise linear discriminant analysis*.

6.16 fuzzy clustering

Classification in which membership of an object in each possible class is given a weight between zero and one.

Note 1: In so-called hard cluster algorithms membership of a group can only take values 0 or 1, but in fuzzy clustering any value between 0 and 1 is allowed subject to the sum of all memberships of an object being 1.

Reference: [37].

6.17 hidden layer

Group of *nodes* in an *artificial neural network* between input layer and output layer.

6.18 hierarchical clustering

Pattern recognition in which objects are linked together by use of an appropriate measure of distance between pairs of objects, and a linkage criterion, which specifies the dissimilarity of sets as a function of the pairwise distances between objects.

Note 1: Distance measures include *Euclidean distance*, *Mahalanobis distance* and *city block distance*.

Note 2: Linkage criteria include *single linkage*, *complete linkage*. See also *Ward's minimum variance method*.

6.19 *k*-means clustering

Unsupervised classification method which partitions objects into k groups, in which each object belongs to the group with the nearest mean.

Note 1: Although k-means clustering is called *unsupervised classification* the value of k may be specified.

6.20 *k*-nearest neighbour (kNN)

Non-parametric *supervised classification* method for objects based on the closest training examples in the variable space.

Note 1: An object is classified by a majority vote of its k-nearest neighbours, with the object being assigned to the class most common.

Note 2: k is a small positive integer.

Note 3: When $k = 1$ the method is known as 'nearest neighbour'.

6.21 Kohonen network

Type of *self-organising map* with low dimensional grid.

6.22 linear discriminant analysis (LDA)

Discriminant analysis in which the criterion function is based on the pooled covariance matrix.

6.23 Mahalanobis distance

Distance ($d_{i,j}$) between an object characterised by vector of variables $\boldsymbol{x}_i$ and the centroid of a class $\boldsymbol{\mu}_j$ with covariance matrix $\boldsymbol{S}$ calculated as

$d_{i,j} = \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)}$, where $\boldsymbol{S}$ is the covariance matrix of the variables.

6.24 misclassification rate

misclassification frequency

Fraction of objects incorrectly assigned to a group in *supervised classification*.

Note 1: Misclassification rate may be calculated for an *evaluation data set*, or in *leave-one-out cross validation*.

Note 2: Misclassification rate is an estimate of the misclassification probability.

6.25 one-class classification

unary classification

independent classification

Classification identifying objects of a specific class amongst all objects, by learning from *training data* containing only objects of the specific class.

Note 1: The assignment of an object to a group may be in the form of a probability of membership.

Reference: [38]

6.26 node

Locations in an *artificial neural network* that carry values that are calculated from values of connected nodes and weights by a transfer.

Note 1: Nodes are arranged in layers, input, *hidden* and output.

Note 2: The value of a node is used in the calculation of nodes in subsequent layers.

6.27 partial least squares discriminant analysis (PLS-DA)

Linear *classification* in which the criterion function is obtained by *partial least squares* analysis.

Note 1: PLS-DA can also be useful for *exploratory data analysis*.

Note 2: PLS-DA results depend on *data pre-processing* and choice of parameters and so is believed to be more difficult to implement well than other forms of *discriminant analysis* such as *linear discriminant analysis*, *quadratic discriminant analysis* or *regularized discriminant analysis*.

Reference: [39]

6.28 pattern recognition

Assignment of a label to an object characterised by data.

Note 1: Pattern recognition is a broad term that includes *classification,* regression, sequencing, outlier detection, biomarker identification and parsing. In chemometrics pattern recognition is often used as a synonym for classification.

Note 2: As with classification, pattern recognition may be supervised or unsupervised. See *supervised classification*, *unsupervised classification*.

6.29 quadratic discriminant analysis (QDA)

Discriminant analysis in which the criterion function is based on the individual within-group covariance matrix.

6.30 regularized discriminant analysis

Discriminant analysis in which the criterion function is based on a combination of pooled covariance matrix and the individual within-group covariance matrix.

Note: The covariance matrix ($\Sigma_k(\lambda)$) is related to the class covariance matrix (Σ_k) and the pooled covariance matrix (Σ_{pooled}) by $\Sigma_k(\lambda) = (1-\lambda)\Sigma_k + \lambda\Sigma_{\text{pooled}}$

Reference: [40]

6.31 self-organising map (SOM)

Unsupervised classification algorithm that creates a projection of a set of given data items onto a regular grid, the nodes of which have a minimum distance from the data in some metric.

Note 1: The grid is usually two dimensional when it is also called a *Kohonen network*.

Note 2: A SOM is considered a type of *artificial neural network*.

Note 3: A SOM can be used to obtain a picture of the relationships among objects (analogous to a *scores plot* in *principal component analysis*).

Reference: [41]

6.32 similarity index

similarity distance

Quantity that describes the equivalence of two objects characterised by *multivariate data*.

Note 1: A similarity index may be in the interval [0,1], where 0 is complete dissimilarity and 1 is complete equivalence.

Note 2: When the term 'distance' is used the quantity is some function of the differences of coordinates in the multivariate data space.
Note 3: See *Tanimoto similarity index*, *city-block distance*, *Euclidean distance*, *Mahalanobis distance, Ward's minimum variance method*.

6.33 single linkage

Linkage criterion in *hierarchical clustering* in which the distance between two clusters is the distance between those two objects (one from each cluster) that are closest together.
Note 1: A drawback of this method is the so-called chaining phenomenon, which refers to the gradual growth of a cluster as one element at a time gets added to it. This may lead to impractically heterogeneous clusters and difficulties in defining classes that could usefully subdivide the data.

6.34 soft independent modelling of class analogy (SIMCA)

Supervised classification that performs a *principal-components analysis* on each class, and then an unknown is assigned to the class with which it has the lowest residual variance.
Note 1: The number of *principal-component factors* chosen for each class is determined by *cross validation*.
Note 2: SIMCA is an example of the use of *disjoint principal component analysis*.

6.35 supervised classification

Classification in which, in a first step, a *model* is built using data from objects of known classes, and in a second step, the model is applied to new data to assign a class to unknown objects.
Note 1: Algorithms for supervised classification include *discriminant analysis, support vector machine, artificial neural network, Bayesian classifier*

Reference: [28] Chapter 33

6.36 support vector machine (SVM)

Method of *supervised classification* in which decision boundaries (hyperplanes) are determined that maximise the separation of data in different classes.
Note 1: The principle guiding SVM classification is the mapping of the original data from the input space to a higher dimensional (which can be infinite) feature space such that the classification problem becomes simpler in the feature space.
Note 2: Linear and non-linear classification problems are treated by SVM.

Reference: [42]

6.37 Tanimoto similarity index

Tanimoto index

Fraction of variables that are considered to agree between two objects.
Note 1: The rules for agreement must be defined. For example in an elemental analysis, element concentrations equivalent within 20 %.
Note 2: The Tanimoto index is used in drug discovery and *QSAR* for comparison of structures in large data bases in an efficient way.
Note 3: The Tanimoto index is essentially similar to the Jaccard index. (https://en.wikipedia.org/wiki/Jaccard_index).

6.38 unsupervised classification

Classification in which no prior information about membership of groups is known.
Note 1: The number of groups may be specified.
Note 2: Algorithms for unsupervised classification include *k-means cluster analysis*, *hierarchical cluster analysis*.

6.39 unweighted pair group method with arithmetic mean (UPGMA)
average linkage

Linkage criterion in *hierarchical clustering* in which the distance between two clusters is the average of all distances between pairs of objects.

6.40 Ward's minimum variance method
Ward's method

Linkage criterion in *hierarchical clustering* in which the within-cluster variance is minimized.
Note 1: The initial cluster distances in Ward's minimum variance method are the squared *Euclidean distance*

7 Calibration and regression

In multivariate regression and calibration the problem is usually posed as a relation between the indications $\boldsymbol{X}$ (which are multivariate) and the property to be measured $\boldsymbol{y}$. Note that this is a reversal of the traditional form, x (concentration)/y (indication) of linear calibration. Therefore in this section c is used in preference to 'y' to remind the reader that it represents the quantity of interest (concentration, classifier). $\boldsymbol{X}$ is a vector of observations (indications).

In matrix form:

$$\boldsymbol{Xb} = \boldsymbol{c} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{b}$ are the coefficients of the model and $\boldsymbol{e}$ a vector of errors. In the definitions that follow we use this terminology, and terms used in *factor analysis* to describe the different approaches. We start with the VIM definition of *calibration*, and note that regression represents the first part of calibration (establishing the relation between $\mathbf{c}$ and $\boldsymbol{X}$).

7.1 calibration

Operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication.

VIM 2.39 [43]

7.2 least squares regression (LSR)
LS regression

Regression that minimizes the sum of squared differences between observed values of a variable and the values predicted by a model.

Note 1: Least squares regression is used as a synonym for *ordinary least squares regression*. The full term should be used if there is ambiguity about the kind of regression being performed.

7.3 errors-in-variables regression (EIV)

total least squares regression (TLS)

Least squares regression in which both the response variable and predictor variable have measurement error.

Note 1: The model for EIV regression is $\begin{cases} c_i = \alpha + \beta x_i^* + \varepsilon_i \\ x_i = x_i^* + \eta_i \end{cases}$ where x^* is the true value of the predictor variable, and ε and η are errors.

Note 2: When the errors ε and η have the same variance the method is called orthogonal regression and minimises the perpendicular distance of a point to the regression line.

Note 3: Total least squares regression is performed by *singular value decomposition*.

Reference: [29] page 213, [44], https://en.wikipedia.org/wiki/Errors-in-variables_models

7.4 mean squared error of calibration (MSEC)

mean residual sum of squares (MRS, MRSS)

In *calibration* for N calibration data where c_i is an observed value and $\hat{c}_i$ is the value predicted by the calibration function

$$MSEC = \frac{\sum_{i=1}^{i=N} (\hat{c}_i - c_i)^2}{N}$$

Note 1: Mean squared error of calibration is the square of the *root mean squared error of calibration*.

7.5 multilinear least squares regression (MLSR, MLR)

Multivariate calibration in which the coefficients of the regression are calculated directly from the indications and values of standards.

Note 1: $\boldsymbol{b} = \boldsymbol{X}^{+}\boldsymbol{c}$, where $\boldsymbol{X}^{+}$ is the pseudo-inverse of $\boldsymbol{X}$, calculated as $\boldsymbol{X}^{+} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$

Note 2: It is assumed that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ has full rank, *i.e.* there are more objects than variables, and the variables are independent.

Note 3: If $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ has full rank only because of *noise*, the solution can become unstable.

7.6 multivariate calibration

Calibration in which the indications are *multivariate data*.

Note 1: Regression establishes the coefficients $\boldsymbol{b}$ (see Equation 1) from given indications $\boldsymbol{X}$, or some factorization of $\boldsymbol{X}$, and values $\boldsymbol{c}$, or some factorization of $\boldsymbol{c}$. Given indications $\boldsymbol{X}_{\mathrm{u}}$ from an unknown sample, the quantity value c_{u} can be calculated.

7.7 ordinary least squares regression (OLSR)

classical least squares regression

Least squares regression that minimizes the sum of squared differences between the known values of the dependent variable and the values predicted by a linear model.

Note 1: Assumptions of the model are that error is only in the dependent variable, Normally-distributed, and homoscedastic.
Note 2: OLSR is used for *calibration* and *multivariate calibration*.

7.8 overfitting of a calibration model

overfitting

Condition in which a model describes random error or *noise* instead of the underlying relationship.
Note 1: Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. (See also *underfitting*)
Note 2: A model which has been overfit will generally have good fit of the calibration data, but poor predictive performance. Overfitting can be detected by use of *cross validation* or *evaluation data*.

7.9 partial least squares regression (PLS)

partial least squares
PLSR (deprecated by ISO 18115)

Multivariate calibration which finds factors that maximise covariance between two blocks of data.
Note 1: PLS finds factors (latent variables) in observed variables **X** that explain the maximum variance in the variable(s) **c**, 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 correlate with the predictions.
Note 2: The decompositions are $\boldsymbol{X} = \boldsymbol{T}_k\boldsymbol{P}_k^{\mathrm{T}} + \boldsymbol{E}$ and $\mathbf{C} = \boldsymbol{U}_k\boldsymbol{Q}_k^{\mathrm{T}} + \boldsymbol{F}$, and $\boldsymbol{X}^{+} = \boldsymbol{W}_k(\boldsymbol{P}_k\boldsymbol{W}_k)^{-1}(\boldsymbol{T}_k^{\mathrm{T}}\boldsymbol{T}_k)^{-1}\boldsymbol{T}_k^{\mathrm{T}}$, where $\boldsymbol{W}$ are weights that maintain orthogonal scores.
Note 3: PLS1 refers to PLS for a single 'c' variable. PLS2 is PLS that simultaneously obtains values for two or more 'c' variables. Therefore in the equations of Note 2, PLS1 has vectors $\boldsymbol{c}$ and $\boldsymbol{q}$ and PLS2 has matrices $\boldsymbol{C}$ and $\boldsymbol{Q}$.
Note 4: When used for *multivariate calibration*, *evaluation data* or *cross validation* may be used to choose the number of PLS factors and assess the accuracy of the prediction (although the same data must not be used to do both). This is important to guard against *overfitting*.
Note 5: PLS may also be used in *classification*. See *partial least squares discriminant analysis*.

Reference: [11] 6.12

7.10 principal components regression (PCR)

Multivariate calibration in which a dependent variable is regressed against the *scores* of a chosen number of *factors* obtained from *principal component analysis* of the predictor variable.
Note 1: PCA decomposes the predictor variable data $\boldsymbol{X}$ into k *principal component factors* $\hat{\boldsymbol{X}}_k = \boldsymbol{T}_k\boldsymbol{P}_k^{\mathrm{T}}$, where k may be determined by *cross validation*. The dependent variable $\boldsymbol{c}$ is then regressed against $\hat{\boldsymbol{X}}_k$, $\boldsymbol{c} = \hat{\boldsymbol{X}}_k\hat{\boldsymbol{b}}$.
Note 2: The factorization gives orthogonal factors, but no information about the predicted variable $\boldsymbol{c}$ is used.
Note 3: See also *partial least squares, multilinear regression*

7.11 ridge regression

damped regression analysis

Multivariate calibration in which damping factors are added to the diagonal of the correlation matrix prior to inversion.

Note 1: $$\hat{\beta}_{\text{ridge}} = \arg\min \sum_{i=1}^{n} (c_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$
$$= \arg\min ||c - X\beta||_2^2 + \lambda ||\beta||_2^2$$

where the second term is known as the penalty, and λ is a tuning parameter. When l is zero, the estimate is a linear regression estimate.

Note 2: The new estimates are no longer unbiased, their expected values are not equal to the true values. However, the variance of this new estimate can be lower than that of the least-squares estimator, so that the total expected mean squared error is also less.

7.12 root mean squared error of calibration (RMSEC)

standard error of calibration (SEC)

deprecated: standard estimate of error (SEE)

In *calibration* for N calibration data where c_i is an observed value and $\hat{c}_i$ is the value predicted by the calibration function.

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{i=N} (\hat{c}_i - c_i)^2}{N}}$$

Note 1: *Overfitting* results in a small RMSEC, but poor predictive ability.

7.13 underfitting of a calibration model

underfitting

Condition in which a model fails to describe the data adequately.

Note 1: Underfitting generally occurs when a model is not sufficiently complex, such as having too few parameters. (See also *overfitting*)

Example 1: Fitting a first order equation to data that follows a quadratic polynomial.

7.14 weighted least squares regression (WLSR)

Least squares regression in which a nonnegative constant is associated with each value of the dependent variable.

Note 1: The nonnegative constants are called weights.

Note 2: If the only source of dispersion in the dependent variable is Normally distributed, the weights are the inverse of the variance of each value.

Note 3: It is assumed the weights are known exactly.

Reference: [45]

8 Index of terms

alias structure 4.1
aliased effects 4.2
alternating least squares regression 5.1
alternating regression. See *alternating least squares regression* 5.1
artificial neural network 6.1
autocorrelation 3.1
Autoregression 5.2
autoscaling 3.2
average linkage. See *unweighted pair group method with arithmetic mean* 6.39
back chaining. See *backward chaining* 6.2
back-propagation learning rule. See *backward propagation* 6.3
back-propagation of errors. See *backward propagation* 6.3
backward chaining 6.2
backward propagation 6.3
backward stepwise linear discriminant analysis 6.4
Bayes classifier 6.5
biplot 5.3
bootstrapping 5.4
calibration 7.1
canonical analysis. See *canonical variate analysis* 5.6
canonical decomposition. See *parallel factors analysis*
canonical variables 5.5
canonical variate analysis 5.6
categorical data 3.3
centering. See *mean centering* 3.11
chemometrics 2.1
city-block distance 6.6
classical least squares regression. See *ordinary least squares regression* 7.7
classification 6.7
cluster 6.8
coded design. See *coded experimental design* 4.3
coded experimental design 4.3
common factor analysis 5.7
complete linkage 6.9
confounded effects. See *aliased effects* 4.2
contingency table 3.4
core consistency diagnostic 5.8
correspondence analysis. See *correspondence factor analysis* 5.9
correspondence factor analysis 5.9
cross tabulation. See *contingency table* 3.4
cross validation 5.10
damped regression analysis. See *ridge regression* 7.11
data matrix 3.5
data pre-processing 3.6
dendrogram 6.10
design matrix 4.5
design of experiments. See *experimental design* 4.7
discriminant analysis 6.11

discriminant function analysis. See *discriminant analysis* 6.11
disjoint principal component analysis 6.12
dummy factor 4.6
dynamic time warping 3.7
effect of a factor 4.4
effect. See *effect of a factor* 4.4
errors-in-variables regression 7.3
Euclidean distance 6.13
evaluation data 3.8
evolving factor analysis 5.11
experimental design 4.7
explanatory variable 3.9
exploratory data analysis 3.10
exploratory factor analysis. See *common factor analysis* 5.7
factor (experimental design) 4.8
factor (factor analysis) 5.12
factor analysis 5.13
factor analysis. See *common factor analysis* 5.7
factor level 4.9
forward chaining 6.14
forward stepwise linear discriminant analysis 6.15
fractional-factorial design 4.10
full-factorial design 4.11
fuzzy clustering 6.16
G-fold cross validation 5.14
Hamming distance. See *city-block distance* 6.6
hidden layer 6.17
hidden variable. See *latent variable* 5.15
hierarchical clustering 6.18
independent classification. See *one-class classification* 6.25
initial data analysis. See *explanatory data analysis* 3.10
interaction effect 4.12
k-means clustering 6.19
k-nearest neighbour 6.20
Kohonen network 6.21
latent construct. See *latent variable* 5.15
latent variable 5.15
least squares regression 7.2
leave-one-out cross validation 5.16
level. See *factor level* 4.9
linear discriminant analysis 6.22
loadings 5.17
loadings plot 5.18
LS regression. See *least squares regression* 7.2
Mahalanobis distance 6.23
main effect 4.13
Manhattan distance. See *city-block distance* 6.6
maximum likelihood principal component analysis 5.19
mean centering 3.11
mean residual sum of squares. See *mean squared error of calibration* 7.4

mean squared error of calibration 7.4
mean squared error of estimation. See *mean squared error of prediction* 5.20
mean squared error of prediction 5.20
misclassification frequency. See *misclassification rate* 6.24
misclassification rate 6.24
model (experimental design) 4.14
multilinear least squares regression 7.5
multiplicative scatter correction 3.12
multiplicative signal correction. See *multiplicative scatter correction* 3.12
multivariate calibration 7.6
multivariate curve resolution 5.21
multivariate data 3.13
multi-way data 3.14
node 6.26
noise 3.15
non-linear iterative partial least squares 5.22
nonlinear mapping 5.23
normalization (in data pre-processing) 3.16
n^{th}-order effect 4.15
N-way data. See *multi-way data* 3.14
one-class classification 6.25
optimization 4.16
ordinary least squares regression 7.7
overfitting of a calibration model 7.8
overfitting. See *overfitting of a calibration model* 7.8
parallel factors analysis 5.24
partial least squares discriminant analysis 6.27
partial least squares regression 7.9
partial least squares. See *partial least squares regression* 7.9
pattern recognition 6.28
Plackett-Burman design 4.17
prediction error sum of squares 5.25
primary data. See *raw data* 3.18
principal component– discriminant analysis 5.26
principal component. See *principal-component factor* 5.28
principal components regression 7.10
principal-component analysis 5.27
principal-component factor 5.28
Procrustes analysis 5.29
quadratic discriminant analysis 6.29
Quantitative Structure-Activity Relationship 5.30
random sampling 3.17
raw data 3.18
regularized discriminant analysis 6.30
residual sum of squares. See *prediction error sum of squares* 5.25
resolution of a design 4.18
resolution. See *resolution of a design* 4.18
response 4.19
response surface methodology 4.20
ridge regression 7.11

root mean square error of cross validation 5.31
root mean squared error of calibration 7.12
root mean squared error of estimation. See *root mean squared error of prediction* 5.32
root mean squared error of prediction 5.32
row scaling. See *normalization (in data pre-processing)* 3.16
sampling 3.19
sampling error 3.20
sampling unit 3.21
scaling 3.22
score 5.34
scores plot 5.35
self organising map 6.31
similarity distance. See *similarity index* 6.32
similarity index 6.32
simple-to-use interactive self-modelling mixture analysis 5.33
simulated annealing 5.36
single linkage 6.33
singular value decomposition 5.37
smoothing 3.23
soft independent modelling of class analogy 6.34
standard error of calibration. See *root mean squared error of calibration* 7.12
standard error of estimation. See *root mean squared error of prediction* 5.32
standard error of prediction. See *root mean squared error of prediction* 5.32
sum of squared errors of prediction. See *prediction error sum of squares* 5.25
sum of squared residuals. See *prediction error sum of squares* 5.25
supervised classification 6.35
support vector machine 6.36
systematic sampling 3.24
take-it-or-leave-it data 3.25
Tanimoto similarity index 6.37
taxi distance. See *city-block distance* 6.6
total least squares regression 7.3
training data 3.26
training set. See *training data* 3.26
transformation 3.27
Tucker tri-linear analysis 5.38
Tucker3 model. See *Tucker tri-linear analysis* 5.38
unary classification. See *one-class classification* 6.25
underfitting of a calibration model 7.13
underfitting. See *underfitting of a calibration model* 7.13
unsupervised classification 6.38
unweighted pair group method with arithmetic mean 6.39
validation data. See *evaluation data* 3.8
variance scaling 3.28
Ward's method. See *Ward's minimum variance method* 6.40
Ward's minimum variance method 6.40
weighted least squares regression 7.14
weighting. See *scaling* 3.22

9 Index of abbreviations

ALS. See *alternating least squares regression* 5.1
ANN. See *artificial neural network* 6.1
CANDECOMP. See *parallel factors analysis* 5.24
CONCORDIA. See *core consistency diagnostic* 5.8
DA. See *discriminant analysis* 6.11
DFA. See *discriminant analysis* 6.11
DoE. See *experimental design* 4.7
EDA. See *explanatory data analysis* 3.10
EFA. See *evolving factor analysis* 5.11
EIV. See *errors-in-variables regression* 7.3
kNN. See k-*nearest neighbour* 6.20
LDA. See *linear discriminant analysis* 6.22
LOOCV. See *leave-one-out cross validation* 5.16
LSR. See *least squares regression* 7.2
MCR. See *multivariate curve resolution* 5.21
MLPCA. See *maximum likelihood principal component analysis* 5.19
MLR. See *multilinear least squares regression* 7.5
MLSR. See *multilinear least squares regression* 7.5
MRS. See *mean squared error of calibration* 7.4
MRSS. See *mean squared error of calibration* 7.4
MSC. See *multiplicative scatter correction* 3.12
MSEC. See *mean squared error of calibration* 7.4
MSEE. See *mean squared error of prediction* 5.20
MSEP. See *mean squared error of prediction* 5.20
NIPALS. See *non-linear iterative partial least squares* 5.22
NLM. See *nonlinear mapping* 5.23
OLSR. See *ordinary least squares regression* 7.7
PARAFAC. See *parallel factors analysis* 5.24
PC. See *principal-component factor* 5.28
PCA. See *principal-component analysis* 5.27
PC-DA. See *principal component– discriminant analysis* 5.26
PCR. See *principal components regression* 7.10
PLS. See *partial least squares regression* 7.9
PLS-DA. See *partial least squares discriminant analysis* 6.27
PRESS. See *prediction error sum of squares* 5.25
QDA. See *quadratic discriminant analysis* 6.29
QSA/PR. See *Quantitative Structure-Activity Relationship* 5.30
QSAR. See *Quantitative Structure-Activity Relationship* 5.30
QSPR. See *Quantitative Structure-Activity Relationship* 5.30
RMSEC. See *root mean squared error of calibration* 7.12
RMSECV. See *root mean square error of cross validation* 5.31
RMSEP. See *root mean squared error of prediction* 5.32
RMSPE. See *root mean squared error of prediction* 5.32
RSS. See *prediction error sum of squares* 5.25
SEC. See *root mean squared error of calibration* 7.12
SEP. See *root mean squared error of prediction* 5.32
SIMCA. See *soft independent modelling of class analogy* 6.34
SIMPLISMA. See *simple-to-use interactive self-modelling mixture analysis* 5.33

SOM. See *self organising map* 6.31
SSEP. See *prediction error sum of squares* 5.25
SSR. See *prediction error sum of squares* 5.25
SVD. See *singular value decomposition* 5.37
SVM. See *support vector machine* 6.36
TILI. See *take-it-or-leave-it data* 3.25
TLS. See *errors-in-variables regression* 7.3
UPGMA. See *unweighted pair group method with arithmetic mean* 6.39
WLSR. See *weighted least squares regression* 7.14

10 Membership of sponsoring bodies

Membership of the Analytical Chemistry Division Committee for the period 2014–2015 was as follows:

President: D. Brynn Hibbert (Australia); Vice President: Jan Labuda (Slovakia); Secretary: Zoltán Mester (Canada); Past President: M. Filomena Camões (Portugal); Titular Members: Christo Balarew (Bulgaria), Yi Chen (China), Attila Felinger (Hungary), Hasuck Kim (Korea), M. Clara Magalhães (Portugal), Heli Sirén (Finland); Associate Members: Resat Apak (Turkey), Peter Bode (Netherlands), Derek Craston (United Kingdom), Yook Heng Lee (Malaysia), Tatyana Maryutina (Russia), Nelson Torto (South Africa); National Representatives: Othman Chande (Tanzania), Laurence Charles (France), Paul DeBièvre (Belgium), Marcos Eberlin (Brazil), Ales Fajgelj (Slovenia), Kate Grudpan (Thailand), Javed Hanif (Pakistan), Daniel Mandler (Israel), Predrag Novak (Croatia), David Shaw (USA).

This manuscript (PAC-REP-15-06-05) was prepared in the framework of IUPAC project 2008-002-1-500.

References

[1] International Organization for Standardization. *3534-1:2006 Statistics – Vocabulary and symbols – Part 1:General statistical terms and terms used in probability*:2006 ISO, Geneva.
[2] International Organization for Standardization. *3534-2:1993 Statistics – Vocabulary and symbols – Part 2: Applied statistics*:1993 ISO, Geneva.
[3] E. R. Cohen, T. Cvitas, J. G. Frey, B. Holmstrom, K. Kuchitsu, R. Marquardt, I. Mills, F. Pavese, M. Quack, J. Stohner, H. L. Strauss, M. Tamaki, A. Thor. *Quantities, Units and Symbols in Physical Chemistry (Green Book)*. The Royal Society of Chemistry, Cambridge (2007).
[4] D. B. Hibbert, P. Minkkinen, N. M. Faber, B. M. Wise. *Anal. Chim. Acta* **642**, 3 (2009).
[5] International Organization for Standardization. *Surface chemical analysis Vocabulary Part 1: General terms and terms used in spectroscopy*, ISO 18115:2010:2010 International Organization for Standardization, Geneva.
[6] P. Geladi, K. Esbensen. *J. Chemom.* **4**, 337 (1990).
[7] K. Esbensen, P. Geladi. *J. Chemom.* **4**, 389 (1990).
[8] B. R. Kowalski. *J. Chem. Inf. Comput. Sci.* **15**, 201 (1975).
[9] B. G. M. Vandeginste. *Chemometricopendium, a chemometrics thesaurus*. http://www.vicim.com/chemometrics%20thesaurus.web/index.html, accessed 1st November 2008.
[10] C. E. Miller. *Am. Pharm. Rev.* **2**, 41 (1999).
[11] International Organization for Standardization. *Surface chemical analysis Vocabulary Part 1: General terms and terms used in spectroscopy*, ISO 18115-1:2010 International Organization for Standardization, Geneva.
[12] J. W. Tukey. *Exploratory Data Analysis*, Addison-Wesley, Boston, MA (1977).
[13] P. Geladi, D. MacDougall, H. Martens. *Appl. Spectrosc.* **39**, 491 (1985).
[14] R. Kramer. *Chemometric Techniques for Quantitative Analysis*, Marcel Dekker, New York (1998).
[15] D. B. Hibbert. *J. Chromatogr. B* **910**, 2 (2012).
[16] International Organization for Standardization. *3534-3:2015 Statistics – Vocabulary and symbols – Part 3: Design of experiments*:2013 ISO, Geneva, Switzerland.
[17] G. W. Cobb. *Introduction to Design and Analysis of Experiments*, Springer-Verlag, New York (1998).
[18] E. Morgan. *Chemometrics: Experimental Design*, Wiley, Chichester (1991).

[19] National Institute of Standards and Technology. *NIST/SEMATECH e-Handbook of Statistical Methods – 5.3.3.4.4. Fractional factorial design specifications and design resolution*. http://www.itl.nist.gov/div898/handbook/pri/section3/pri3344.htm, accessed September 2015.
[20] G. E. Box, K. Wilson. *Journal of the Royal Statistical Society. Series B (Methodological)* **13**, 1 (1951).
[21] R. Wehrens, H. Putter, L. M. Buydens. *Chemometrics Intellig. Lab. Syst.* **54**, 35 (2000).
[22] B. Efron, R. Tibshirani. *An Introduction to the Bootstrap,* Chapman & Hall / CRC Press, Boca Raton, FL (1993).
[23] R. Bro, H. A. L. Kiers. *J. Chemom.* **17**, 274 (2003).
[24] M. Maeder, A. Zilian. *Chemometrics Intellig. Lab. Syst.* **3**, 205 (1988).
[25] H. R. Keller, D. L. Massart. *Chemometrics Intellig. Lab. Syst.* **12**, 209 (1992).
[26] P. D. Wentzell, D. T. Andrews, D. C. Hamilton, K. Faber, B. R. Kowalski. *J. Chemom.* **11**, 339 (1997).
[27] R. Bro. *Chemom. Intell. Lab. Syst.* **38**, 149 (1997).
[28] B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. D. Jong, P. J. Lewi, J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part B*, Elsevier Science B.V., Amsterdam (1998).
[29] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. D. Jong, P. J. Lewi, J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier Science B.V., Amsterdam (1997).
[30] W. Windig, B. Antalek, J. L. Lippert, Y. Batonneau, C. Brémard. *Anal. Chem.* **74**, 1371 (2002).
[31] H. Martens, M. Martens. *Multivariate Analysis of Quality. An Introduction*, John Wiley and Sons, Chichester (2001).
[32] R. Hecht-Nielsen. *Neurocomputing*, Addison Wesley, Boston (1990).
[33] M. Caudill. In *AI Expert,* Miller Freeman Publications, San Francisco (1989).
[34] D. E. Rumelhart, G. E. Hinton, R. J. Williams. in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, D. E. Rumelhart, J. L. McClelland, (Eds.), pp. 318–382. MIT Press, Cambridge, MA (1986).
[35] K. Fukunaga. *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, CA (1990).
[36] R. Brereton. in *Chemometrics for Pattern Recognition*, pp. 236–239. John Wiley & Sons, Chichester, UK (2009).
[37] R. O. Duda, P. E. Hart, D. G. Stork. *Pattern Classification*, Wiley-Interscience, New York (2001).
[38] R. G. Brereton. *J. Chemom.* **25**, 225 (2011).
[39] R. G. Brereton, G. R. Lloyd. *J. Chemom.* **28**, 213 (2014).
[40] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D. L. Massart, S. Heuerding, F. Erni. *Anal. Chim. Acta* **329**, 257 (1996).
[41] T. Kohonen, T. Honkela. *Kohonen network*. http://www.scholarpedia.org/article/Kohonen_network, accessed September 2015.
[42] J. Luts, F. Ojeda, R. Van de Plas, B. De Moor, S. Van Huffel, J. A. Suykens. *Anal. Chim. Acta* **665**, 129 (2010).
[43] Joint Committee for Guides in Metrology. *International vocabulary of metrology – Basic and general concepts and associated terms VIM*, JCGM 200:2012 BIPM, Sèvres, www.bipm.org/en/publications/guides/vim.html.
[44] S. van Huffel, P. Lemmerling. *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*. Springer, Netherlands (2013).
[45] National Institute of Standards and Technology. *NIST/SEMATECH e-Handbook of Statistical Methods – 4.1.4.3 Weighted least squares regression* http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd143.htm, accessed May 2015.