

Chapter 2

Multivariate Data Analysis (Chemometrics)

Sylvie Roussel, Sébastien Preys, Fabien Chauchard and Jordane Lallemand

2.1 Introduction

2.1.1 Definition of Chemometrics

Chemometrics, or multivariate data analysis, is the science which applies optimal mathematical and statistical methods to process data. Chemometrics includes the design of experiments upstream and the analysis of data to get valuable information after measurements have been taken. The need for chemometrics tools mainly comes from the development of analytical instruments providing large amounts of increasingly complex data.

This scientific arena consists of a large variety of mathematical methods, aiming at processing numerous data sets to achieve diverse objectives. The scheme below is an overview of the chemometrics approach any scientist should follow when facing a multivariate data analysis issue (Fig. 2.1).

Even though the principles of chemometrics are based in mathematics and statistics, one does not need to have deep knowledge of either of these disciplines to analyse multivariate data. However, thorough knowledge of the application as well as common sense are required in order to analyse the outputs of the chemometrics software packages and avoid pitfalls and misinterpretations.

2.1.2 PAT and Chemometrics

Process analytical technology (PAT), as defined in Chap. 1, includes appropriate measurement devices, that can be placed at-, in- or on-line, combined with mul-

S. Roussel (✉) · S. Preys · J. Lallemand
Ondalys, ZA La Plaine, 4 rue Georges Besse, 34830 Clapiers, France
e-mail: contact@ondalys.fr

F. Chauchard
Indatech, ZA La Plaine, 4 rue Georges Besse, 34830 Clapiers, France

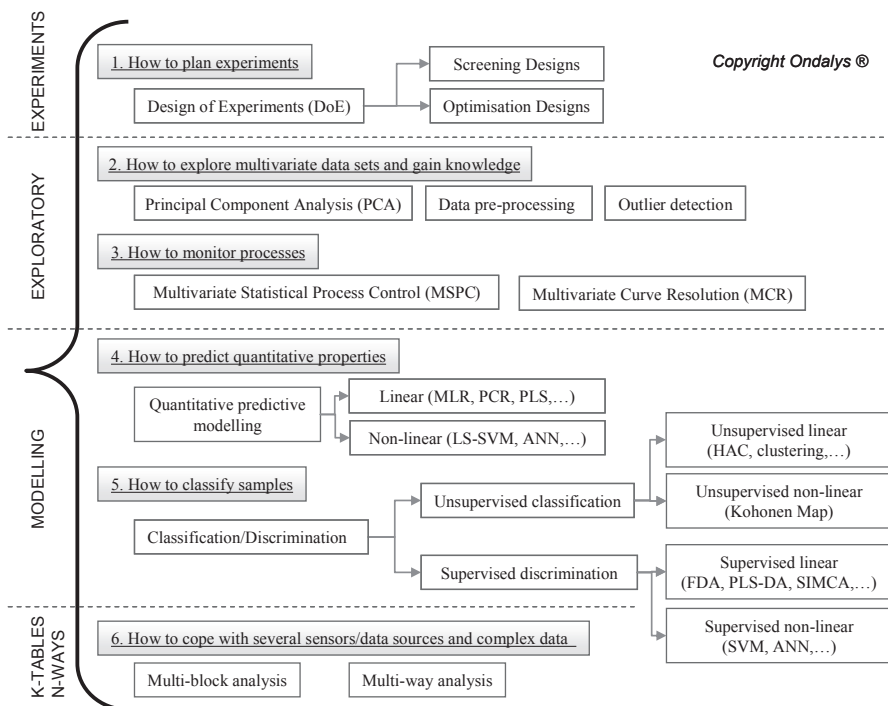


Fig. 2.1 The multivariate data analysis approach: classification of chemometrics methods

tivariate statistical (chemometrics) tools to analyse data and monitor and control processes. Chemometrics is therefore essential to understanding and diagnosing real-time processes, and keeping them under multivariate statistical control. PAT is also strongly linked to the quality by design (QbD) concept, which implies quality integration from the product development stage. Within the PAT framework, Wold et al. identified five levels of chemometrics analysis corresponding to different data and objective complexity levels (Wold et al. 2006):

- PAT-1: calculating critical quality attributes (CQA), such as concentrations, from rapid and real-time multivariate measurements, such as spectra, by multivariate calibration (predictive modelling).
- PAT-2: sorting samples (raw materials, intermediate or final products) as acceptable or not, based on multivariate measurements, such as spectra or property profiles, using multivariate statistical process control (MSPC).
- PAT-3: monitoring and classifying batch processes as acceptable or not from real-time multivariate measurements, such as process data, raw material data and spectra, using batch SPC (BSPC).
- PAT-4: combining data from all the critical process steps and raw materials to assess the final product quality using multi-block analysis.
- PAT-5: including feedback control to the process settings from the multivariate models, using process dynamic identification and time series modelling among others. This last level is not discussed further in this chapter.

For each of these levels, the modelling approach works well if the training set used to build the model is representative of the acceptable (in-control) samples. This is ensured by covering the desired variability, either by using design of experiments (DoE) during the process development at laboratory and pilot scales or by using huge historical laboratory or production databases with sufficient variability. Finally, the robustness of the model has to be regularly evaluated during its life cycle through maintenance and updating.

2.2 Design of Experiments

Using historical databases to model processes requires a very large amount of observations to ensure a minimum of variability. When it is possible, a more rational way consists in choosing the observations or experiments to span the whole desired operating conditions, i.e. the design space, with a maximum of variability. DoE (experimental designs) corresponds to that part of chemometrics which aims at planning the relevant experiments, minimising the cost without decreasing information quality, quantifying the different factor effects, modelling and optimising the processes (Gacula and Jagbir Singh 1984; Box and Draper 1987; Lundstedt et al. 1998; Leardi 2009). Different designs corresponding to different objectives are discussed in the following sections, such as screening and optimisation designs.

2.2.1 Problem Formulation

Understanding and modelling a process requires first to determine its multivariate inputs and outputs. On the one hand, the inputs represent the different factors or parameters which may have an influence on the outputs, such as temperature, pressure or the type of catalyst for a chemical reaction. They correspond to the independent measurements or variables which can be set independently of one another. The user's expertise and some tools, such as the Ishikawa diagram, are needed to determine an exhaustive list of the potential variability sources. Factors which cannot be precisely set by the user, i.e. uncontrolled factors, cannot be considered as inputs. On the other hand, outputs correspond to the response measurements (or dependent variables) which have to be optimised, such as the yield of a chemical reaction.

2.2.2 Screening Designs

The DoE methodology often includes a first step which consists in implementing a screening design (Araujo and Brereton 1996a). The experiments are chosen in order to quantify the influential factors among a large number of factors.

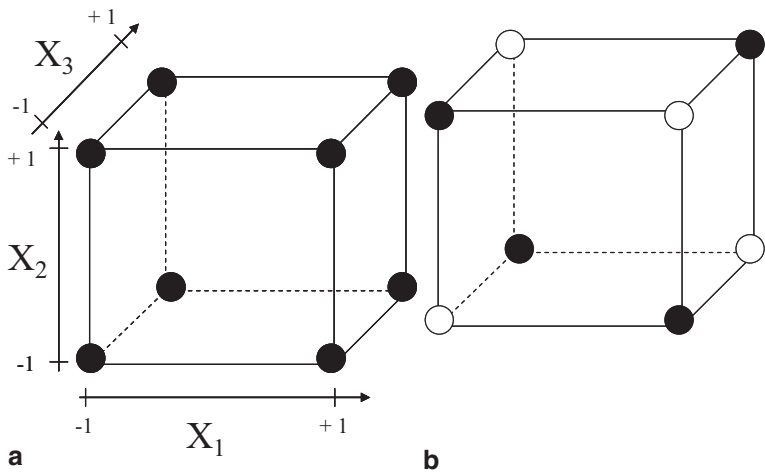


Fig. 2.2 a Full factorial design. b Fractional factorial design with three factors

Table 2.1 Example of an experimental design for a full factorial design with three factors

Experiment	Average I	X_1	X_2	X_3	X_{12}	X_{13}	X_{23}	X_{123}	Response Y
1	+1	-1	-1	-1	+1	+1	+1	-1	60
2	+1	+1	-1	-1	-1	-1	+1	+1	72
3	+1	-1	+1	-1	-1	+1	-1	+1	54
4	+1	+1	+1	-1	+1	-1	-1	-1	68
5	+1	-1	-1	+1	+1	-1	-1	+1	52
6	+1	+1	-1	+1	-1	+1	-1	-1	83
7	+1	-1	+1	+1	-1	-1	+1	-1	45
8	+1	+1	+1	+1	+1	+1	+1	+1	80
Effects	64.25	11.5	-2.5	0.75	0.75	5	0	0.25	

2.2.2.1 Full Factorial Designs (2^k)

Full factorial designs are the basic designs which carry out all possible experiments with k two-level factors, low and high levels. All experiments at the boundaries of the design space are planned, as illustrated for three factors in Fig. 2.2a. The corresponding experimental matrix with its encoding system is shown in Table 2.1.

The main effects for each factor are calculated as the semi-difference between the high-level average and the low-level average. They represent the average direct impact of each factor on the response when increasing the encoded factor level from 0 to 1.

First-degree interaction effects between two factors are then processed as the semi-difference between the effect of factor 1 at factor 2 high level and the effect of factor 1 at factor 2 low level. The second-degree interaction corresponds to the interaction between three factors. Interactions with a degree higher than 1 are however often small and difficult to interpret.

Table 2.2 Experimental design for the fractional factorial design with three factors coming from the full factorial design in Table 2.1

Experiment	Average I	X ₁	X ₂	X ₃	X ₁₂	X ₁₃	X ₂₃	X ₁₂₃
2	+1	+1	-1	-1	-1	-1	+1	+1
3	+1	-1	+1	-1	-1	+1	-1	+1
5	+1	-1	-1	+1	+1	-1	-1	+1
8	+1	+1	+1	+1	+1	+1	+1	+1

The resulting model equation for the experimental matrix in Table 2.1 is illustrated in Eq. 2.1:

$$y = 64.25 + 11.5x_1 - 2.5x_2 + 0.75x_3 + 0.75x_{12} + 5x_{13} + 0.25x_{123} + \varepsilon. \quad (2.1)$$

The significance of each effect has then to be determined by means of statistical tests. Estimating the experimental uncertainty s_y from m replicate measurements, the uncertainty of each calculated effect value is $\sigma_E = s_y / \sqrt{n}$, where n is the number of designed experiments. An effect is thus significant at a risk α if the calculated effect value (model parameter, e.g. for factor 1) is greater than $t_{(1-\alpha, \nu)} \cdot \sigma_E$, where $t_{(1-\alpha, \nu)}$ is found in the Student's law table and $\nu = m - 1$ is the degree of freedom. Another way to assess the significance of each effect is to run an analysis of variance (ANOVA) and observe the calculated p values associated with each effect. The resulting significant effects, meaning that these factors are statistically influent on the response, have to be maintained for the remainder of the model development.

2.2.2.2 Fractional Factorial Designs (2^{k-p})

Fractional factorial designs are used to screen factors when the number of experiments has to be lowered (Fig. 2.2b). The alias principle allows selection of which experiments from the full factorial design must be run without losing significant information. The idea is to choose the experiments which lead to confound important effects, such as main and first-degree interaction effects, with smaller and less interpretable effects, such as second (and more)-degree interaction effects. For example, with three factors, Eq. 2.2 shows the alias generator. The resulting experimental matrix is in Table 2.2. The number of experiments is reduced to 2^{k-p} , where p is the number of alias generators. With three factors, only one alias generator is allowed, dividing the number of experiments by two for similar model accuracy. The experimental matrix shows that main effects are confounded with first-order effects as the encoding is the same two by two. This only allows the interpretation of principal effects:

$$\mathbf{I} = \mathbf{X}_{123}. \quad (2.2)$$

Randomisation of the experiment order is usually needed to correct eventual systematic response errors. When experimental blocks are clearly identified, such as

analysis days, alias generators are used to confound the block effects with high-order interaction effects.

Some food applications were developed by Ellekjaer et al., who studied the effects of process variables and ingredients on sensory variables for processed cheese (Ellekjær et al. 1996), and Christiansen et al., who implemented a fractional factorial design to model food dressings (Christiansen et al. 2004).

2.2.2.3 Other Screening Designs

Other screening designs using linear models are also commonly used to identify the few significant factors among many.

The Plackett–Burman designs are two-level saturated designs where all interaction effects are neglected (Plackett and Burman 1946). The number of experiments is a multiple of four, and “saturated” means that this number is equal to the number of model parameters, i.e. the number of factors plus one (model constant), without any degree of freedom left. For example, a six-factor Plackett–Burman design requires theoretically a minimum of seven experiments, running finally eight experiments (multiple of four). The total number of experiments is hence drastically reduced.

The Rechtschaffner screening designs correspond to saturated two-level fractional factorial designs to estimate main and first-order interaction effects (Rechtschaffner 1967). For example, a six-factor Rechtschaffner design requires 22 experiments ($1 + k + k(k-1)/2$, with k the number of factors).

2.2.3 Optimisation Designs: Response Surface Methodology

When two-level factorial designs have difficulties to model a process, showing a significant lack-of-fit when observing ANOVA results or when using validation experiments at the centre of the experimental domain, second-order designs, also called optimisation designs, are used. These designs propose to carry out experiments at more than two levels, allowing curvature modelling. Non-linear response surfaces can thus be drawn to achieve the main goal of these designs, i.e. estimating the experimental area corresponding to the response optimum (Araujo and Brereton 1996b).

2.2.3.1 Central Composite Designs

Central composite designs are widely used, since they can complete an existing full factorial screening design with $2k$ additional experiments, designing a “star” around the existing hyper-cube. Additional experiments to the centre can be required (Fig. 2.3 and Table 2.3). Five levels for each factor are thus investigated to model non-linearity.

Fig. 2.3 Central composite design with three factors

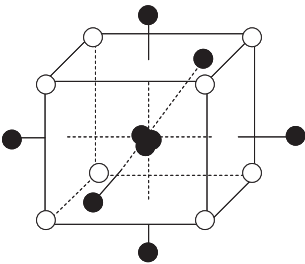


Table 2.3 Number of experiments for central composite designs

k	Full factorial design 2^k	Star points		Centre	Total
		Number $2k$	α		
2	4	4	1.414	3	11
3	8	6	1.682	3	17
4	16	8	2	3	27
5	32	10	2	4	46

The model equation for a second-degree design is shown in Eq. 2.3:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_{12}x_{12} + \beta_{13}x_{13} + \beta_{23}x_{23} + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{33}x_3^2 + \varepsilon. \tag{2.3}$$

2.2.3.2 Other Optimisation Designs

Central composite designs are the most popular optimisation designs. However, when the number of factors increases, the number of experiments rapidly becomes very large. Thus other second-degree designs are also commonplace.

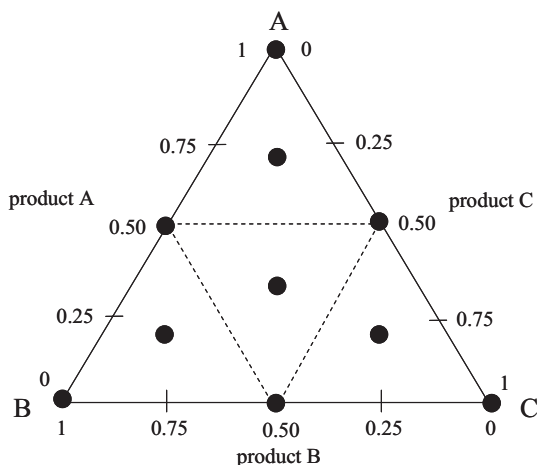
The 3^k three-level full factorial designs are an extension of the two-level full factorial designs seen in Sect. 2.2.2.1.

The Rechtschaffner optimisation designs, similar to the screening Rechtschaffner designs in Sect. 2.2.2.3, correspond to saturated three-level fractional factorial designs to estimate main and first-order interaction effects. For example, a six-factor Rechtschaffner design requires 28 experiments $(1 + k + k + k(k - 1))/2$, with k the number of factors). Additional experiments in the centre are always recommended.

The Box–Behnken designs are incomplete three-level full factorial designs, without experiments in the corners of the experimental domain (Ferreira et al. 2007). Application on several responses related to bread-making quality is illustrated in Rouillé et al. (2000).

Doehlert designs allow the estimation of all main effects, first-order interactions and quadratic effects without any confounding effects (Ferreira et al. 2004). Their geometric shape is polyhedronic based on hyper-triangles (simplexes). The specificity of the Doehlert designs is related to the ability to extend them to contiguous experimental domains for one or more factors in a sequential way. The number of levels is finally not the same for each factor.

Fig. 2.4 Example of mixture design for three products (augmented simplex-centroid design for quadratic models)



2.2.3.3 Mixture Designs

When dealing with formulation optimisation, the closure constraint in the mixture (Eq. 2.4) has to be taken into account (Cornell 1990; Eriksson et al. 1998):

$$\sum_{i=1}^c x_i = 1, \quad (2.4)$$

where x_i is a compound (factor) of the mixture and c the total number of the compounds.

This mixture constraint leads to an interdependency between all factors and hence has several consequences. First, the representation of the experiments does not imply hyper-cubes but hyper-tetrahedrons (Fig. 2.4). Second, the underlying models are simplified. The linear model loses its constant term and the second-order model loses its constant and quadratic terms.

2.3 Exploratory Analysis

The first step of chemometrics analysis consists in performing an exploratory analysis in the multivariate space, also called descriptive analysis or unsupervised analysis which occurs without prior knowledge concerning neither the nature nor the group membership of the samples. Initially, data have to be preprocessed or “cleaned” before the exploratory treatment. This is often performed using principal component analysis (PCA).

2.3.1 Data Preprocessing

Data preprocessing techniques are often used prior to modelling in order to reduce noise and undesired perturbations in the signal. Several preprocessing methods have been initially developed in near-infrared spectroscopy, due to its sensitivity to the external environment (temperature changes, humidity, etc.). The most suitable preprocessing technique will depend on the conditions; these must be compared to find the optimal combination on a given data set.

2.3.1.1 Classical Preprocessing Methods

The most widely used preprocessing methods consist in mean centring or scaling the data. They can be used for all types of multivariate data: continuous, discrete, spectroscopic or process data.

- *Mean centring* is the most common preprocessing. The principle is to subtract the variable mean to each value. Mean centring is quasi-systematic in projection methods such as PCA or PLS. It is used in order to centre the subspace to the barycentre of the original data set, for a better data visualisation (see Sect. 2.3.2). When building a predictive model, mean centring $\mathbf{X}$ data set implies that the constant term (b_0) of the equation is not equal to zero (see Sect. 2.4.2.1, Eq. 2.13). Thus, in the cases where the intercept is expected to be null, the data should not be centred.
- *Scaling* is used to make the different variables comparable when included in a global multivariate analysis. The most common scaling technique is the *unit-variance scaling* which divides each variable by its standard deviation, like a columnwise normalisation. The method must be systematically applied to data sets containing variables of different scales (e.g. pH, temperature) in order to give them equal weights in further processing. Scaling should not be applied to spectroscopic data because each variable is comparable and the intensity variations between wavelengths constitute the important information (e.g. spectral peaks). Other kinds of scaling are possible for this data, for instance, to stress the importance of specific variables by giving them higher weights.
- *Auto-scaling* is the combination of mean centring and unit-variance scaling.

2.3.1.2 Signal Correction Methods

The signal correction methods aim at correcting the influence of different perturbations and/or enhancing information. In spectroscopic data, perturbations can be additive, i.e. a constant, which can be wavelength dependent, is added to the spectrum, or multiplicative, where each element of the spectrum is multiplied by a constant. These phenomena are typical of light scattering effects, which induce a photon loss

(additive effect) and an increased path length (multiplicative effect), among others. Scatter correction must not be applied if the parameter of interest is physical in nature (e.g. particle size, turbidity).

Almost all the methods cited below are “rowwise” methods, i.e. the preprocessing is carried out sample by sample. It is not the case for mean centring and scaling, where all (calibration) samples are required in order to preprocess the data set, i.e. they are “columnwise” treatments.

A recent review of some of the mentioned signal correction preprocessing methods can be found in Rinnan et al. (2009).

- *Baseline correction* subtracts the undesired spectral background. The classic way is to subtract the lowest value of each spectrum from all the variables. *Detrending* removes curvilinear baseline by approximating it with a wavelength-dependent second-degree polynomial fit.
- *Normalisation* is used rowwise when there is a non-desired intensity variation between objects due to multiplicative effects. This allows focus on the data profile rather than the global intensity. Normalisation is done by dividing each spectrum by an estimation of its spectral intensity. This can be done using the following properties: its area (area normalisation), its maximal peak (maximum normalisation), a specific spectral point (peak normalisation), its length (unit vector normalisation), or the sum of the spectral values.
- *Standard Normal Variate (SNV)* is a path-length variation correction method used, like normalisation, to limit the spectral intensity variation problem (Fig. 2.5b). It is a rowwise auto-scaling, thus removing the spectrum mean value to all the spectrum variables and dividing them by the spectrum standard deviation (Barnes et al. 1989).
- *Multiplicative Signal Correction (MSC)* is also a very common method for correcting multiplicative scattering effects (Geladi et al. 1985). The principle is to fit each spectrum to a reference spectrum (generally, the average calibration database spectrum, Eq. 2.5), and then to correct them as shown in Eq. 2.6. The reference spectrum must be representative to avoid an ill-fitting model. Different versions have been derived. For instance, the extended MSC (EMSC) is based on a polynomial baseline correction depending on the wavelength; it can also allow for the introduction of prior information in the spectra (Martens and Stark 2001):

$$\mathbf{x}_i = a * \bar{\mathbf{x}} + b \quad (2.5)$$

$$\mathbf{x}_i^{\text{MSC}} = \frac{\mathbf{x}_i - b}{a}, \quad (2.6)$$

where $\mathbf{x}_i$ is the measured spectrum, $\bar{\mathbf{x}}$ is the mean spectrum (or a reference spectrum), a is the intercept, b the slope and $\mathbf{x}_i^{\text{MSC}}$ the corrected spectrum.

- *Smoothing* is used to remove random noise. The principle is to use an average of neighbouring points. For example, the moving average method uses the average of a neighbouring window to calculate the new value. The Savitzky–Golay (SG) algorithm uses a polynomial fit (Savitzky and Golay 1964). The latter is the

Process Analytical Technology for the Food Industry

O'Donnell, C.P.; Fagan, C.; Cullen, P.J. (Eds.)

2014, VIII, 301 p. 107 illus., 45 illus. in color., Hardcover

ISBN: 978-1-4939-0310-8