Food Engineering Series
Series Editor: Gustavo V. Barbosa-Cánovas

Colm P. O'Donnell
Colette Fagan
P.J. Cullen *Editors*

Process Analytical Technology for the Food Industry

Springer

Food Engineering Series

Springer's Food Engineering Series is essential to the Food Engineering profession, providing exceptional texts in areas that are necessary for the understanding and development of this constantly evolving discipline. The titles are primarily reference-oriented, targeted to a wide audience including food, mechanical, chemical, and electrical engineers, as well as food scientists and technologists working in the food industry, academia, regulatory industry, or in the design of food manufacturing plants or specialized equipment.

More information about this series at http://www.springer.com/series/5996

Colm P. O'Donnell • Colette Fagan • P.J. Cullen
Editors

Process Analytical Technology for the Food Industry

Springer

Editors
Colm P. O'Donnell
University College Dublin
Belfield
Ireland

Colette Fagan
University of Reading
Reading
United Kingdom

P.J. Cullen
Dublin Institute of Technology
Dublin
Ireland

ISSN 1571-0297
ISBN 978-1-4939-0310-8 ISBN 978-1-4939-0311-5 (eBook)
DOI 10.1007/978-1-4939-0311-5
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014940790

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Contributors

J. Burger BurgerMetrics SIA, Jelgava, Latvia

Fabien Chauchard Indatech, Clapiers, France

Qiaofen Cheng Department of Food and Nutritional Sciences, Whiteknights, Reading, UK

P.J. Cullen School of Food Science and Environmental Health, Dublin Institute of Technology, Dublin 1, Ireland

School of Chemical Engineering, University of New South Wales, Sydney, Australia

Cheng-Jin Du Warwick Systems Biology Centre, University of Warwick, Coventry, UK

Colette C. Fagan Department of Food and Nutritional Sciences, University of Reading, Reading, UK

Food and Nutritional Science, Department of Food and Nutritional Sciences, University of Reading, Reading, UK

Tat Hean Gan Brunel University, Middlesex, UK

E. Gaston IRIS-Innovació i Recerca Industrial i Sostenible, Castelldefels, Barcelona, Spain

Jarka Glassey School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne, UK

A. A. Gowen School of Food Science and Environmental Health, Dublin Institute of Technology, Dublin 1, Ireland

Joseph Irudayaraj Agricultural & Biological Engineering, Purdue University, West Lafayette, IN, USA

Digvir S. Jayas Department of Biosystems Engineering, University of Manitoba, Winnipeg, MB, Canada

Julie Lundtoft Johnsen Arla Strategic Innovation Centre, Arla Foods amba, Aarhus, Denmark

Ramazan Kizil Chemical Engineering Department, College of Chemical and Metallurgical Engineering, Istanbul Technical University, Maslak, Istanbul, Turkey

Jordane Lallemand Ondalys, Clapiers, France

Kathryn L. McCarthy Department of Food Science and Technology, University of California, Davis, CA, USA

Michael J. McCarthy Department of Food Science and Technology, University of California, Davis, CA, USA

Colm P. O'Donnell School of Biosystems Engineering, University College Dublin, Dublin 4, Ireland

Sébastien Preys Ondalys, Clapiers, France

Sylvie Roussel Ondalys, Clapiers, France

R. Vadivambal Department of Biosystems Engineering, University of Manitoba, Winnipeg, MB, Canada

Chapter 1
Benefits and Challenges of Adopting PAT for the Food Industry

P.J. Cullen, Colm P. O'Donnell and Colette C. Fagan

1.1 Introduction

Process analytical technology (PAT) is a framework for innovative process manufacturing and quality assurance. The concept is to design, analyse and control manufacturing processes through the measurement of identified critical control parameters which govern product variability. The identified benefits of the framework include increased process efficiency, reduced operating costs, increased process validation and ultimately improved final product quality and safety.

1.1.1 Evolution of PAT

Process analytical chemistry (PAC) is a term which developed during the 1940s to describe the application of analytical chemistry with techniques, algorithms and sampling equipment to solve developing problems related to various chemical processes. Although industrial process analysers have been in use for more than 60 years, the modern period of PAC essentially began with the formation of the Centre for Process Analytical Chemistry (CPAC) in 1984. The goal of PAC was to "supply quantitative and qualitative information about a chemical process" for monitoring, control and optimization. They went on to define five "eras" of PAC: (1) off-line, (2) at line, (3) on-line (4) in-line and (5) non-invasive, which describe the evolution

P.J. Cullen (✉)
School of Chemical Engineering, University of New South Wales,
Sydney, Australia
e-mail: patrick.j.cullen@dit.ie

C. P. O'Donnell
School of Biosystems Engineering, University College Dublin, Dublin 4, Ireland

C. C. Fagan
Department of Food and Nutritional Sciences, University of Reading,
P.O. Box 226, Reading RG6 6AP, UK

C. P. O'Donnell et al. (eds.), *Process Analytical Technology for the Food Industry,*
Food Engineering Series, DOI 10.1007/978-1-4939-0311-5_1,

of sensor technologies (Mishra et al 2008). Its definition has evolved over the years to encompass analytical measurements and understating of chemical, physical and microbiological parameters governing processing. Changing the term "chemistry" to "technology" allowed a broader scope of the approach to other processes. The pharmaceutical industry, in particular, has adopted the approach as a strategy to understand and control variability within the sector. The broad definition given by the US Food and Drug Administration (FDA): "A system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes with the goal of ensuring final product quality" covers the requirements and desires of manufacturing within the food industry.

Since 1987, PAT has had a dedicated international conference (International Forum Process Analytical Chemistry, IFPAC, which brings together instrumentation manufacturers, researchers and industry users.

1.1.2 Learning From Other Process Industries

The food industry has always been to the fore with regard to adoption of sensors and the use of risk analysis strategies. By comparison, the pharmaceutical industry has been more restricted in the adoption of advanced control strategies due to validated batch production processes, high-value-added products and lack of specialised technologies. Nevertheless, it was identified that an improved production process was required and PAT emerged as a platform of future good manufacturing practice (GMP). Industrial adoption has still remained relatively low; however, the ideology and desire for the approach is evident between regulatory agencies and industry alike.

The pharmaceutical industry also recognised that PAT could provide additional benefits such as continuous validation. Q7A GMP Guidance for Active Pharmaceutical Ingredients (API) defines validation as: *a documented program that provides a high degree of assurance that a specific process, method, or system will consistently produce a result meeting predetermined acceptance criteria.* PAT tools are capable of continuously measuring product *acceptance criteria* and critical control points (CPPs), thereby continuously evaluating if the process is behaving *consistently.*

Food treatment processes such as pasteurization need validation; the emergence of novel treatment processes such as high-pressure processing, pulsed electric fields, etc. necessitates process validation to ensure adequate treatment. A PAT strategy may facilitate such validation. Also, similar complementary benefits as found in the pharmaceutical industry may occur within the food industry with increased PAT implementation. For example, Hazard Analysis Critical Control Point (HACCP) is used in the food industry to identify potential food safety hazards so that key actions can be taken to reduce or eliminate the risk of the hazards being realized. HACCP is *a systematic preventive approach to food safety that addresses physical, chemical, and biological hazards as a means of prevention rather than finished product inspection.* Integrating a PAT strategy within HACCP may support the overall goal of ensuring food safety through the use of process monitoring.

Comparing both industries, we can see similarities and differences which influence the drivers to PAT adoption. The pharmaceutical industry is highly regulated and is risk averse with actives synthesised via both chemical and biological routes; many processes are operated under sterile conditions in highly controlled environments. Batch production continues to dominate with low levels of automation. Product quality is typically monitored off-line using laboratory-based methods of analysis. Food production within developed countries is increasingly regulated and is risk adverse to microbial or chemical contamination. Production is typically non-sterile.

1.1.3 PAT Drivers in the Food Industry

Consumer: For the food industry, the consumer is a key driver for food production methodologies. Produce taste, nutrition, appearance, cost and shelf life are important parameters influencing consumer purchase and thereby the choice of production methodology. Technologies which can optimize the process will ultimately lead to consumer loyalty and repeat purchase. Consequently, the food industry has traditionally employed technologies to monitor food produce particularly as end-point quality control strategies.

Regulators Pharmaceutical regulators have played a fundamental role in the push to adopt PAT strategies within the pharmaceutical industries. For pharmaceutical current GMP (cGMP), both the FDA and the European Medicines Agency (EMA) openly encourages the use of PAT. The FDA formed a PAT advisory committee that includes industry participation which is intended to facilitate dialogue between the regulators and industry. They have issued a PAT guidance document and created a PAT training program for industry. By comparison, there has been little involvement from food regulators in such a strategy. However, as PAT technologies begin to show promise as food safety prevention techniques this may change. Apart from production optimization, many of the PAT tools discussed within this book such as hyperspectral imaging have showed potential for contamination identification of foods within production facilities (Chap. 9). Similarly, PAT data may be useful in food traceability or potentially have a role to play in the development of proactive hazard alert system in ensuring food security throughout the food chain. If the approach can show reduced risk to consumers, food regulators may become increasingly interested in the approach.

Business One of the common reasons forwarded as to explain the limited uptake of PAT tools by the pharmaceutical industry has been the lack of a business case to improve current production processes given the high margins that drug companies traditionally operated within, coupled with the validation implications with adopting new production approaches. By comparison, the food manufacturing industry has typically been governed by lower profit margins and production efficiency. PAT tools may offer improved production cost and energy efficiency through process optimization along with increased quality control.

Sustainability The concept of environmental sustainability in food processing has become a key issue in recent times as awareness of the importance of environmental protection, and the possible impacts associated with the manufacture and consumption of food products, has increased. Environmental sustainability can be achieved by developing and implementing alternative environmental best-practice technologies and products which maximise the efficient use of resources and achieve cost savings, while minimising negative human and environmental impacts (Clark 2011). Recently, PAT has been linked to green production strategies:

> The Process Analytical Technology (PAT) initiative has made its name as a mechanism for monitoring processes in real time, facilitating process understanding and, in some cases, real-time release of product. With PAT, the focus has been on monitoring Critical Quality Attributes and controlling Quality Critical Process Parameters, and to a lesser degree manufacturing efficiency. It may be time to add sustainability to the list of PAT's objectives. (Thomas 2009)

1.1.4 Technology Advances

Originally, PAC measurements were performed off-line; however, they moved continually closer to production to result in: at line, on-line and in-line measurements. The principal benefit from such a paradigm shift was a significant decrease in the time delay between sampling and analysis along with monitoring of more representative samples. The past two decades have seen significant progress in the incorporation of on- and in-line process monitoring using advanced instrumentation. Optical and spectroscopic technologies have been to the fore in this advancement including computer vision, ultraviolet–visible (UV–Vis), near- and mid-infrared (NIR and MIR) and Raman spectroscopy. This rise was facilitated by related technological advances in solid-state detectors, fibre optics and instrumentation innovations for in situ sampling (Chew and Sharrat 2010) along with a parallel evolution of computer processing power.

This use of increasingly sophisticated process analysers resulted in increasingly large data sets that require appropriate numerical strategies to unravel chemical information (or process signatures) and associated process states encoded within the analytical data. With increasing know-how and affordability of spectroscopic and chromatographic instrumentation for on-line and in-line process analysis, PAT data sets are often intrinsically multivariate in nature (Chew and Sharrat 2010). Chemometrics has emerged and is widely embraced as a useful tool to unravel the data obtained by PAT.

1.1.5 Challenges

To facilitate widespread adoption of PAT within the food industry, a number of challenges need to be overcome. Technologies must meet the challenges posed by the food manufacturing environment including: cleaning-in-place (CIP) compatibility,

harsh environments, real-time analysis, low-cost and ease of use. Technologies which provide food safety information directly from the production environment would be advantageous, with identification of microbial or chemical contamination. Suitable data management systems need to be developed and integrated with production to ensure that the benefits offered by PAT are achieved.

Unlike the pharmaceutical industry, variability with food production is sometimes welcomed. Indeed, one of the admirable aspects of artisan or "home-made" food is the variable nature of product. However, adoption of PAT does not mean that we have to produce foods which appear perfectly similar to some defined optimum. It is up to the manufacturers themselves to decide on what parameters they wish to optimize; indeed, PAT could be used to produce foods which display more artisan-like features.

Finally, the food industry needs to be made more aware of PAT as a framework for innovative process manufacturing and quality assurance. More collaboration between industry, academia and regulators is required to unify the disperse efforts currently underway. Adoption of PAT as a strategy would bring together process engineers, food scientists, technologists and microbiologists under one umbrella with the goal of providing the industry with a manufacturing framework for the twenty-first century. Here, we can learn from and cooperate with other industries such as the pharmaceutical and petrochemical to further develop the strategy.

References

Chew W, Sharrat P (2010) Trends in process analytical technology. Anal Methods 2:1412–1438

Clark J (2011) Introduction to green chemistry. In: Proctor A (ed) Alternatives to conventional food processing. RSC Publishing, London, p 1–10. http://pubs.rsc.org/en/content/ebook/978-1-84973-037-2

Mishra A, Banerjee S, Bhatwadekar N, Mahajan P, Karode P (2008) Process analytical technology (PAT): boon to pharmaceutical industry. Pharm Rev 6:3

Thomas P (2009) Is it time for PAT to go green? Pharma Manufacturing. http://www.pharmamanufacturing.com/articles/2009/148.html. Accessed 20 Oct 2011

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

C. P. O'Donnell et al. (eds.), *Process Analytical Technology for the Food Industry,*
Food Engineering Series, DOI 10.1007/978-1-4939-0311-5_2,

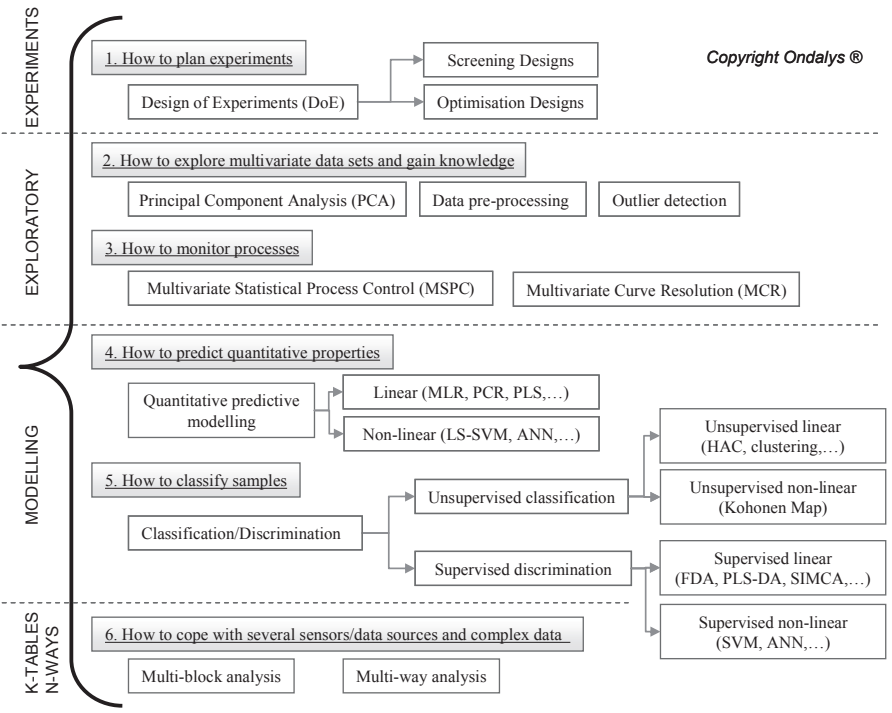

Fig. 2.1 The multivariate data analysis approach: classification of chemometrics methods

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.

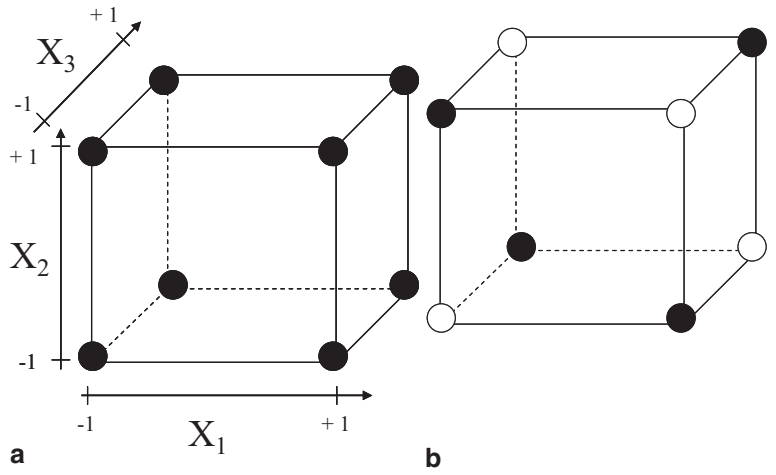

Fig. 2.2 **a** Full factorial design. **b** Fractional factorial design with three factors

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:

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

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 $t_{(1-\alpha,\nu)}.\sigma_E$, 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 aliase 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 aliase 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 aliase generators. With three factors, only one aliase 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:

$$\mathbf{I} = \mathbf{X_{123}}. \qquad (2.2)$$

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

analysis days, aliase 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.

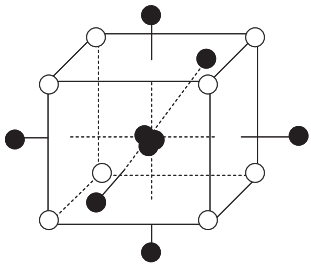

Fig. 2.3 Central composite design with three factors

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:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{23} x_{23} + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \varepsilon. \tag{2.3}$$

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.

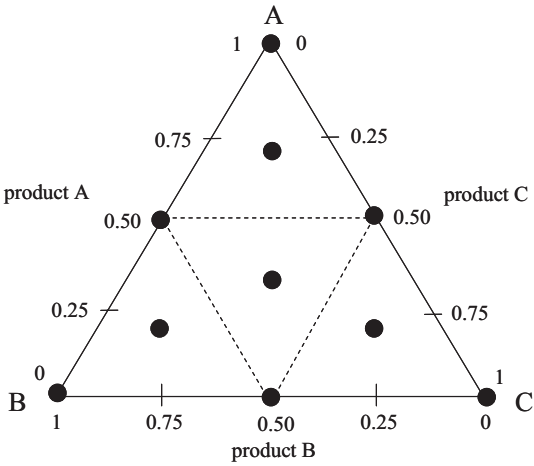

Fig. 2.4 Example of mixture design for three products (augmented simplex-centroid design for quadratic models)

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):

$$\sum_{i=1}^{c} x_i = 1, \tag{2.4}$$

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 **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):

$$\mathbf{x}_i = a * \overline{\mathbf{x}} + b \tag{2.5}$$

$$\mathbf{x}_i^{\mathrm{MSC}} = \frac{\mathbf{x}_i - b}{a}, \tag{2.6}$$

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

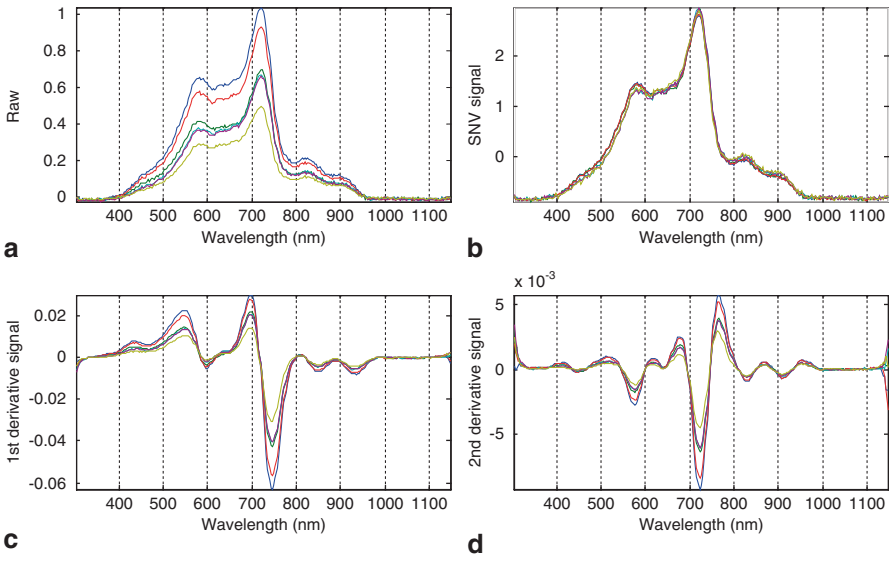

Fig. 2.5 Raw spectra (**a**), SNV (**b**), first derivative (**c**) and second derivative (**d**) spectra for a visible-NIR spectra data set

most widespread algorithm in chemometrics. The wider the window the higher the smoothing while a polynomial degree increase will better fit tiny spectral features and enhancing noise. Thus, a balance between these parameters must be found.

- *Derivatives* are generally used in spectroscopy to enhance spectral features; however, they also correct for additive effect, as a constant baseline (first derivatives, see Fig. 2.5c) or both the offset and the slope of the baseline (second derivatives, see Fig. 2.5d). The most classical algorithms used are the SG (Savitzky and Golay 1964) and the Norris–Williams (Norris and Williams 1984) algorithms. Since computing derivatives enhances the noise, these two methods also smooth the spectrum.

2.3.1.3 Dimensionality Reduction Methods

Dimensionality reduction methods aim at eliminating uninformative signal, thus enhancing information and reducing collinearity:

- *Variable selection* is often used to remove uninformative or noisy variables and keep the relevant ones. For data with only a few variables (e.g. process data), the classical method is the stepwise-multiple linear regression (stepwise-MLR) method, used to select the most informative variables during the model development. For data with a higher number of variables and containing more collinearities (e.g. spectroscopic data), other methods are preferred. A common method in spectroscopy is the interval-partial least squares (i-PLS), which selects the spectral regions in testing all possible combinations with one or several mov-

ing window(s) of a fixed size (Nørgaard et al. 2000). A plethora of different methods have been developed. For uninformative variable elimination (UVE), the addition of dummy random variables allows the identification of spectral variables that are as uninformative as noise (Centner et al. 1996). The genetic algorithm (GA) is a stochastic method using the principle of evolution theory to select a subset of variables (Leardi and González 1998). A large number of parameters must be tuned to apply this method. Furthermore, being a stochastic technique, each run provides a different result.

Some comparison studies can be found in Abrahamsson et al. (2003).

- For predictive modelling purposes, orthogonal pretreatments, such as *orthogonal signal correction* (OSC), remove variations which are not linked to **Y** (Wold et al. 1998a). Some variants deriving from the classical OSC and differing in the way the non-relevant part is modelled, can be cited: direct orthogonalisation (Andersson 1999), direct OSC (Westerhuis et al. 2001), piecewise OSC (Feudale et al. 2002a) and orthogonal-PLS (Trygg and Wold 2002). These methods have for objective to extract the net analyte signal (NAS), i.e. the part of the spectra related to the quantity of interest and which is orthogonal to the other compounds. It has been shown that these pretreatments do not always provide better model performances than partial least-squares (PLS) models (see Sect. 2.4.2.4) based on raw data. However, the models based on these preprocessed data provide a better understanding of the model. In this way, Svesson et al. propose some discussions and a comparison on different "OSC" algorithms (Svensson et al. 2002).
- *Data compression* methods aim at reducing the dimensionality of large data sets. Latent structures are extracted and used to rebuild a "cleaned" signal without noise. PCA is the most common method (see Sect. 2.4.2). Other methods are also widely used, like Fourier transformation (FT) working on the frequency domain (McClure et al. 1977; Wu et al. 1996) or wavelet transform (WT) working on time and frequency domains (Daubechies 1990; Alsberg et al. 1997). Both methods present the advantage of working on one spectrum at a time, whereas PCA needs all spectra. These methods are useful to reconstruct the corrected signal or to extract the latent structures (e.g. PCA scores, FT or WT coefficients) and use them as inputs to derive predictive or discrimination models.

A comparison of several preprocessing methods for improving the determination of moisture and protein contents of forage samples is given by Azzouz et al. (Azzouz et al. 2003).

2.3.2 Principal Component Analysis

2.3.2.1 Introduction—Objective of PCA

The PCA is the major workhorse of the chemometrics tools. The PCA method can be used for the following goals:

- Visualisation of X in the multivariate space
- Outlier detection
- Variable selection
- Data compression, when reducing X dimensionality by removing noise
- Be the basics for other multivariate methods, such as unsupervised classification or MSPC

2.3.2.2 Geometrical Interpretation

The PCA can be seen as a better way to visualise samples represented by numerous variables, by projecting their original coordinates into a new set of axes, called principal components (PC). These axes satisfy a number of properties, which make the sample visualisation easier.

The following graphs explain how PCA works for a simple **X**-matrix, composed of only three variables (Fig. 2.6a):

- Each sample is located in the original space with its three coordinates (Fig. 2.6b). The X-matrix can then be visualised as a cloud of points in the three-dimensional (3D) space (Fig. 2.6c).
- The coordinate system is translated to the barycentre of the sample cloud (star in Fig. 2.6d), for a better visualisation, by mean centring the variables.
- A new axis, called first PC, is built following the direction of the maximum spread of the samples (Fig. 2.6d); this helps better visualise the maximum variability of the sample set. The new coordinates of the samples are called Scores.
- A second axis, orthogonal to the first one, is then searched, to represent the maximum of the remaining sample set variance (Fig. 2.6g). This axis can be visualised by placing the eye facing to the first PC (Fig. 2.6e–f).
- This process is done iteratively for the number of PCs equal to the number of original variables (three in this graphical example) (Fig. 2.6h).

Thus, PCA can be seen as a change of axes, designed to better visualise the sample variability, but maintaining the distances and scales between samples. For more convenience, the samples are usually visualised on a 2D plane, corresponding to the projection of the samples on this set of two axes (see Fig. 2.7, an example of plane PC1 × PC3).

2.3.2.3 Mathematical Computation

The spread, inertia or variance of the cloud of samples seen in the previous section is expressed mathematically by the variance–covariance matrix of **X**. Thus, the PCA decomposes the sample set space in the direction of the maximum of **X**-variance. The matrix of variance–covariance ($\mathbf{V}_{(p,\,p)}$) is computed as shown in Eq. 2.7, with $\tilde{\mathbf{X}}$ being the centred matrix of **X**:

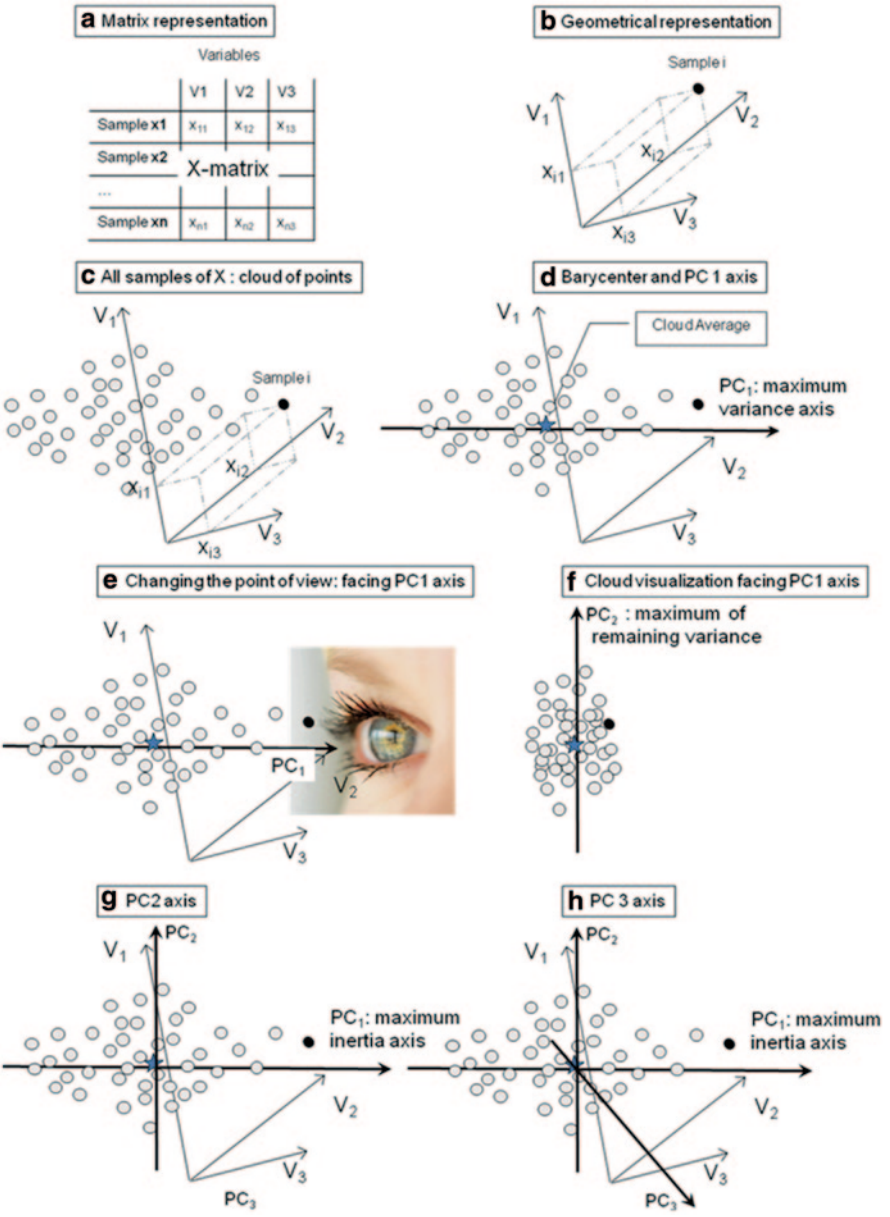

Fig. 2.6 Principal component analysis (PCA) geometrical visualisation

$$\mathbf{V} = \frac{\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}}{n-1}. \tag{2.7}$$

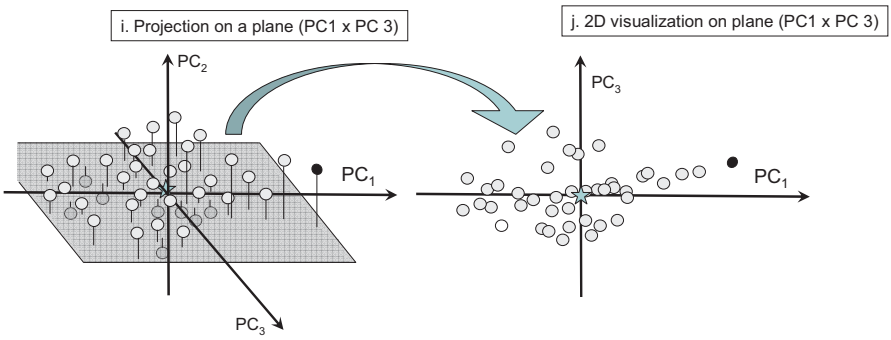

Fig. 2.7 PCA visualisation on a 2D plane

To find the maximum inertia axes, the PCA algorithm diagonalises the matrix of variance–covariance, computing the eigenvalues and eigenvectors of **V**. Then, the eigenvalues are sorted in a descending order, since they are directly related to the variance explained by each axis.

The eigenvectors are called loadings (**P**) and correspond to the weights of each original variable to build the PC: the PC is thus a linear combination of the original variables.

Then, the initial data set **X** can be decomposed with the following expression:

$$\mathbf{X}_{(n,p)} = \mathbf{T}_{(n,p)}\mathbf{P}^{T}_{(p,p)}, \tag{2.8}$$

where **P** is the loading matrix and **T** the score matrix of the new coordinates in the PC subspace. Matrix dimensions are noted in parenthesis.

Since data contains a part of the information and a part of the noise (error), we can express **X** with the first k components, the p–k last components remaining in the error matrix (see Eq. 2.9 and Fig. 2.8a):

$$\mathbf{X}_{(n,p)} = \mathbf{T}_{(n,k)}\mathbf{P}^{T}_{(k,p)} + \mathbf{E}_{(n,p)}, \tag{2.9}$$

where **P** is the loading matrix for k components and **T** the corresponding scores. Matrix dimensions are noted in parenthesis.

Thus, when a centred PCA is performed, each sample can be decomposed as shown in Fig. 2.8b. The variables are generally centred (see Sect. 2.3.1.1), to translate the coordinate system origin to the barycentre of the sample cloud, and provide a better visualisation.

2.3.2.4 Interpretation of PCA

When a PCA is performed, scores and loadings are visualised to understand: