B978-0-444-59528-7.00003-X, 00003

Chapter 3

C0015 # Exploratory Data Analysis

Mario Li Vigni, Caterina Durante and Marina Cocchi[1]
Department of Chemical and Geochemical Sciences, University of Modena and Reggio Emilia, Modena, Italy
[1]*Corresponding author: marina.cocchi@unimore.it*

Chapter Outline

s0005 ## 1 THE CONCEPT (LET YOUR DATA TALK)

p0005 In the food area, as in most research fields, the system complexity to be faced is increasing both in the way of producing food and in the consumer expectations evolved, together with the targets of regulatory authority and society needs and issues. Food production is connected to environmental, socio-economic challenges; food consumption with health, safety and nutritional attitudes. Among the emerging research areas there are the study and making of functional food, the use of nanotechnology, the assessment of food authenticity, including provenance and organic production, and the monitoring and improvement of the quality of food processing. From the point of view of food and food-processing characterization this implies that we need to extract information and obtain models capable of inferring the underlying relationships that link the compositional profile and the processing conditions to very general end properties of foodstuff, such as the healthiness, the consumer perception, the link to a territory and so on. Moreover, the implication of the production chain on food quality has also to be assessed.

p0010 In this respect, the research attitude cannot be purely 'deductive': theory-driven hypothesis could be not only inefficient but even difficult to formulate.

Data Handling in Science and Technology, Vol. 28. http://dx.doi.org/10.1016/B978-0-444-59528-7.00003-X

DHST, 978-0-444-59528-7

B978-0-444-59528-7.00003-X, 00003

This is why and where researchers may benefit from new technological tools (analytical instrumentation, hardware and algorithms/software development) to come back to an 'inductive' data-driven attitude with a minimum of *a priori* hypothesis as a first efficient step to progress faster and further.

p0015 To this aim exploratory data analysis (EDA) is well suited. EDA is well known in statistics and sciences as that operative approach to data analysis aimed to improve understanding and accessibility of the results. Without forgetting the soundness of statistical models and hypothesis formulation, which is intrinsically connected to the concept of 'analysis' in its scientific meaning, the focus is moved to 'exploration', which, as a word, leads to more exotic thoughts and feelings, such as unravelling mysterious threads or discovering unknown worlds. As a matter of fact, EDA does relate to the process of revealing hidden and unknown information from data in such a form that the analyst obtains an immediate, direct and easy-to-understand representation of it. Visual graphs are a mandatory element of this approach, owing to the intrinsic ability of the human brain to get a more direct and trustworthy interpretation of similarities, differences, trends, clusters and correlations through a picture, rather than a series of numbers. As a matter of fact, our perception of reality is that we believe what we are able to see.

p0020 The other axiom of EDA is that the focus of attention is on the data, rather than the hypothesis. This means, figuratively, that it is not the analyst 'asking' questions to the data, as in an interrogation, instead the data are allowed to 'talk', giving evidence of their nature, the relationships which characterize them, the significance of the information which lies beneath what has been evaluated on them – or even the complete absence of any of this, if it is the case.

p0025 One of the milestone references for EDA is the comprehensive book by Tukey [1]. Tukey, in his work, aimed to create a data analysis framework where the visual examination of data sets, by means of statistically significant representations, plays the pivotal role to aid the analyst to formulate hypotheses that could be tested on new data sets. The stress on two concepts such as dynamic experimenting on data (e.g. evaluating the results on different subsets of a same data set, under different data-preprocessing conditions) and exhaustive visualization capabilities offers researchers the possibility to identify outliers, trends and patterns in data, upon which new theories and hypothesis can be built. Tukey's first view on EDA was based on robust and nonparametric statistical concepts such as the assessment of data by means of empirical distributions, hence the use of the so-called five-number summary of data (range extremes, median and quartiles), which led to one of his most known graphical tools for EDA, the box plot.

p0030 This approach well denotes the conceptual shift from confirmatory data analysis, where a hypothesis and a distribution are assumed on the data, and statistical significance is used to test the hypothesis on the basis of the data (where the less reliable the results, the more the data divert from the postulated distribution), to EDA, where the data are visualized in a distribution-free

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:11:59:50 Page Number: 2

B978-0-444-59528-7.00003-X, 00003

approach, and hypotheses arise from the observation, if any, of trends and clusters or correlations among them. In practice, the objectives of EDA aim to

u0005 • Highlight phenomena occurring in the observations so that hypotheses about the causes can be suggested, rather than 'forcing' hypotheses on the observations to explain phenomena known *a priori*. 'The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data' [2].

u0010 • Provide a basis to assess the assumption for statistical inference, for example, by evaluating the best selection of statistical tools and techniques, or even new sampling strategies, for further investigations. 'Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone as the first step' [1].

p0045 The tools and techniques of EDA are strongly based on the graphical approach mentioned so far. Data visualization is given by means of box plots, histograms and scatter plots, all distribution-free instruments which can be extremely useful to probe if the data follow a particular distribution.

p0050 At the beginning of its development, EDA represented a kind of Copernican revolution, in the sense that it put data and the information they bring, not the hypothesis and the information it seeks, at the centre of attention. However, using it in a framework where the common approach was to reduce problems into simpler forms that were solvable, usually by constraining the experimental domains to uni- or oligovariate models nowadays, shows huge limitations. When dealing with scientific fields such as chemistry, and in particular analytical chemistry, and food science, where instrumental analysis can provide at least thousands of variables for each sample, often in a fast way, data complexity has exponentially increased to the point that a multivariate approach (i.e. the evaluation of the simultaneous effect of all the variables which characterize a system on the relationships among its samples) is mandatory. The use of graphical instruments is limited to human ability to interpret two-dimensional (2D) and three-dimensional (3D) spaces, which is impossible to apply when variability is represented through, for example, an analytical signal. Correlation tables, albeit offering a direct view of which variables are related to each other, are often complex to read and interpret. Multivariate analysis methods, especially those based on latent variables projection, provide the best tool to combine the analysis of variable correlations and sample similarities/differences, the reduction of variable space to lower dimensions and the possibility of offering graphical outputs that are easy to read and interpret. Thus, the passage from EDA to exploratory multivariate data analysis (EMDA) is conceptually easier than the one from confirmatory data analysis to EDA, as it only represents a shift towards the use of methods which are based on a multivariate approach to data.

p0055 EMDA stays on the track opened by EDA, in order to grasp the data structure without imposing any model. It has to be stressed that when dealing with

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:11:59:50 Page Number: 3

B978-0-444-59528-7.00003-X, 00003

a multidimensional space, visualization requires either a projection step or a domain change, for example, from acquired variables to a similarity/dissimilarity space. Thus, if *a priori* hypotheses (imposed models) are avoided, most often some assumptions on data are adopted. This brings a diversity of complementary instruments that can be used, and we may say that EMDA is a road that several tools allow you to travel.

p0060 The aim of this chapter is to illustrate the most used and effective tools for the analysis of food-related data, so that the reader is offered some clues about which tool to choose and what is possible to get out of it.

s0010 2 DESCRIPTIVE STATISTICS

p0065 All those visualization tools which allow the exploration of uni- and oligovariate data can be considered as instruments of descriptive statistics. Descriptive statistics is usually defined as a way to summarize/extract information out of one or a few variables: compared to inferential statistics, whose aim is to assess the validity of a hypothesis made on measured data, descriptive statistics is merely explorative. In particular, some salient facts can be extracted about a variable:

o0005 **i.** A measure of the central tendency, that is, the central position of a frequency distribution of a group of data. In other words, a number which is better suited to represent the value of the investigated samples with respect to the measured property. Typical statistics are mean, median and mode.

o0010 **ii.** A measure of spread, describing how spread out the measured values are around the central value. Typical statistics are range, quartiles and standard deviation.

o0015 **iii.** A measure of asymmetry (skewness) and peakedness (kourtosis) of the frequency distribution, that is, if the spread of data around the central value is symmetric in both left/right directions, and how sharp/flat is the distribution in the central position, respectively.

p0085 While useful, these statistics are only a summarization of data and do not offer a direct interpretation benefit when compared to a graphical representation of the data. The main graphical tools for descriptive statistics are frequency histograms, box-whisker graphs and scatter plots. These tools are useful to inspect the statistics reported earlier, the presence of outliers and multiple modes in the data (histograms), to highlight location and variation changes between different groups of data or among several variables (box-whisker), to reveal relationships or associations between two variables (scatter plots), as well as to highlight dependency with respect to some ordering criterion, such as run order, time, position, etc.

p0090 Albeit simple and in spite of the high degree of summarizing they carry with them, these tools can also be particularly useful prior to EMDA.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:11:59:52 Page Number: 4

B978-0-444-59528-7.00003-X, 00003

It may seem a paradox, but they are very effective to identify gross errors: for example, a huge difference between median and mean for a variable could be due to a misprinted number, or could suggest the need to transform variables (e.g. log transform) and help choosing the appropriate data pretreatment.

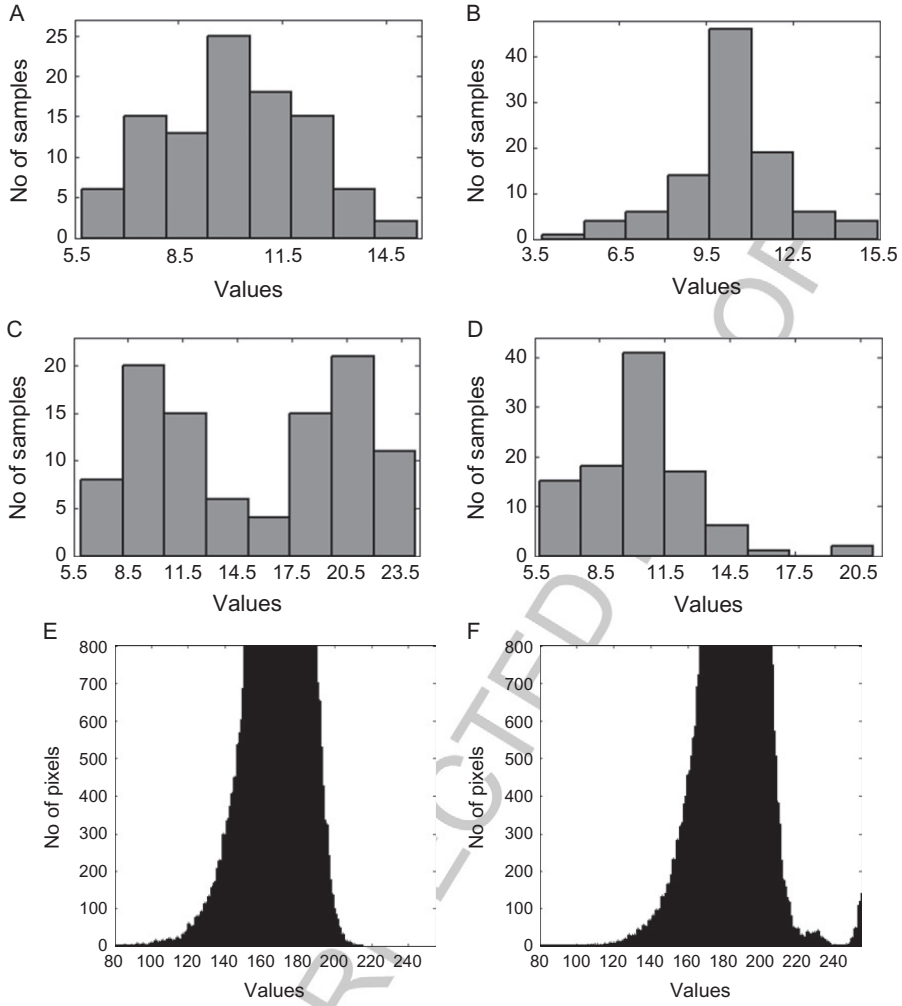
2.1 Frequency Histograms

p0095 To draw a histogram, the range of data is subdivided in a number of equally spaced bins. Thus, frequency histograms report on the horizontal axis the values of the measured variable and on the vertical axis the frequencies, that is, the number of measurements, which fall into each bin. The number of bins influences the efficacy of the representation, thus some attention must be given in their choice. Some common rules have been coded, among which the most used considers a number of bins k equal to the square root of the number of samples n, or equal to $1+\log_2(n)$. Theoretically derived rules are reviewed in Scott's book [3], and iterative methods have also been proposed [4]. In most cases, one of the two rules cited earlier is enough to obtain a nice representation of data distribution, but the choice of k becomes critical when n is huge, for example, if you want to represent a frequency histogram of pixels intensity of an image, where the number of 'samples' easily goes beyond several hundreds of thousands.

p0100 Figure 1 reports some examples of histograms which are quite common to find for discrete variables. Figure 1A shows what to expect when the variable has an almost normal distribution, that is a maximum frequency of occurrence for a given value (close to the average of the values) and decreasing frequencies for higher and lower values. Skewness of the distribution (Figure 1B) is indicated by a higher frequency of occurrence for values which are higher or lower than the most frequent one. Histograms can show the presence of clusters in the data according to a given value, as can be seen in Figure 1C: here it is possible to see two values of higher frequency, around which two almost normal distributions suggest the existence of two clusters. In addition, the presence of outliers (Figure 1D) can be highlighted. An outlier usually has a value way higher or lower than all the other samples, hence it will appear in the histogram as a bar both well separated from the main cluster of values and showing a low frequency of occurrence. As mentioned, histograms can also be used when the number of observations is high (on the other hand, they lose any exploratory meaning when used for data sets where the number of variables is very high and correlated, such as in instrumental signals), as shown in the last two parts of Figure 1. Here, the distribution of pixels of images is used. In particular, Figure 1E shows the zoomed view of pixel distribution for an image acquired on a product (in this case, a bread bun) which is considered a production target (i.e. the colour intensity and homogeneity of its surface are inside specification values for that product): it is possible to see that frequencies of occurrence are almost symmetrically distributed across the average value (data have been centred across the mean intensity value).

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:11:59:52 Page Number: 5

B978-0-444-59528-7.00003-X, 00003



f0005 **FIGURE 1** Examples of histograms. (A) Almost normal distribution of a discrete variable; (B) skewed distribution (higher values have a higher frequency of occurrence); (C) overlapping of two distributions centred across a different mean value (possibly indicating the presence of two clusters); (D) presence of outliers (low frequency of occurrence for high values); (E) pixel distribution of a reference image; and (F) pixel distribution of an image where defects are detected (defective pixels bring to the bump in the right tail of frequency distribution and to the frequency bars detected for values >240). In the pixel distribution cases, a zoom has been taken to highlight the differences.

A different shape is manifest in Figure 1F, where an image of a sample with surface defects (such as darker or paler colour, or the presence of spots and blisters) is considered. In this case, the pixel distribution is skewed towards positive values and lower frequency occurrence features appear, which are an index of phenomena which deviate from the bulk of the data, such as darker localized spots.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:11:59:52 Page Number: 6

B978-0-444-59528-7.00003-X, 00003

s0020 2.2 Box and Whisker Plots

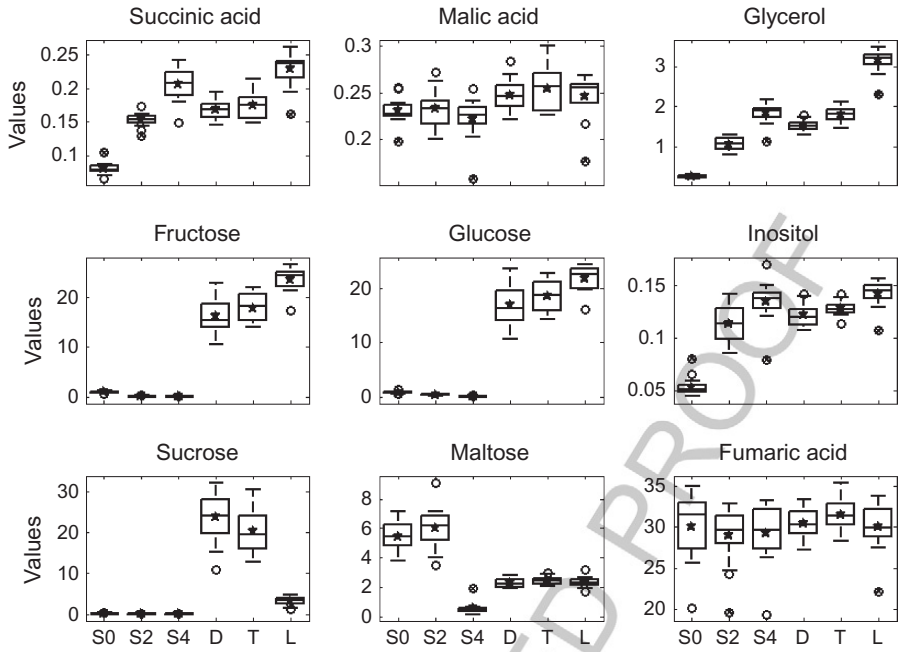
p0105 Box and whisker plots (box plot in short) [5–7] are very useful to summarize the kind of information that in inferential statistics we seek by means of analysis of variance (ANOVA). Indeed, they allow a direct comparison of the distribution of several variables (on the order of tens, visualization becomes inefficient) in terms of both central location and variation. Thus, it is a quick way to estimate if a grouping factor has potentially a significant effect on the measured variables. Typical questions that can be answered are: Does the location differ between subgroups? Does the variation differ between subgroups? Are there any outliers?

p0110 The construction of a box plot requires calculating the median and the quartiles of a given variable for the samples: the lower quartile (LQ) is the 25th percentile and the upper quartile (UQ) is the 75th percentile. Then a box is drawn (hence the name) whose edges are the lower and upper quartiles: this box represents the middle 50% of the data and the difference between the upper and lower quartile is indicated as the inter quartile range (IQR). Generally, the median is represented by a line drawn inside the box, and in some representations the mean is also drawn as an asterisk, for example, to better evaluate the differences between central tendency descriptors. Then a line is drawn from the LQ to the minimum value and another line from the UQ to the maximum value and typically a symbol is drawn at these minimum and maximum points (whiskers). In most implementations, if a data point presents a value higher than $UQ + 1.5 * IQR$ or smaller than $LQ - 1.5 * IQR$, it is represented by a circle. This helps pointing out potential outliers (the circle may be drawn with a higher dimension if the LQ or UQ is exceeded by $3 * IQR$). A single box plot can be drawn for one set of samples with respect to one variable; alternatively, multiple box plots can be drawn together to compare several variables, groups in a single set of samples or multiple data sets. Box plots become difficult to draw and interpret in those cases where it is necessary to deal with continuous data, such as spectra or signals.

p0115 Figure 2 shows a box plot representation of each of the nine variables which characterize the GCbreadProcess data set (see Section 3.1.4 for more details on the data set), the concentration of chemical compounds determined in gas chromatography (GC) at six points of an industrial bread-making process, namely, S0, S2, S4, D, T and L. As it is possible to regroup data according to the sampling point, the representation is useful to obtain a screening evaluation of which variables show different distributions across the phases of the production process. For example, fumaric acid and malic acid have similar distributions and values for all six points (the 'box', that is the IQR, and the 'whiskers', that is the 95th and 5th percentile range, are almost overlapped for all the sampling points), thus they will be of little use to differentiate the process phases. On the contrary, fructose and glucose show a clear difference in both range and mean and median value (respectively, the star and the

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:11:59:55 Page Number: 7

B978-0-444-59528-7.00003-X, 00003



f0010 **FIGURE 2** oxplot representation of the nine variables which characterize the GCbreadProcess data set (see text for details).

horizontal line inside the box) for points S0, S2 and S4 with respect to points D, T and L. The presence of potential outliers, that is points which fall beyond the 95th and 5th percentile limits, is indicated by circles (crossed circles are characterized by a 3*IQR distance). This representation can be a starting point to decide which variables, or combination of more, are the best to differentiate the six points.

s0025 3 PROJECTION TECHNIQUES

p0120 EMDA pursues the same objectives illustrated for uni- and oligovariate EDA, namely giving a graphical representation of multivariate data highlighting the data patterns and the relationships among objects and variables with no *a priori* hypothesis formulation. The importance of this step and its relevance in food analysis is worth being stressed. In fact, the multivariate exploratory tools make it feasible to generate hypotheses from the data, notwithstanding how complex they are, opening to the researcher a way towards the formulation of new ideas. In other words, intuition is inspired by the synergy of data reduction and graphical display. In fact, by compressing the data to a few parameters, without losing information, it becomes possible to look at data, so that the researcher's mind can capture data variation in terms of grouping and patterns in the *natural way*.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:11:59:56 Page Number: 8

B978-0-444-59528-7.00003-X, 00003

p0125 It is indeed very different, with respect to the possibility of enhancing discovery, to operate simplification by reduction at the problem level or data level [8]. In the first case, prior knowledge is used to isolate or split the complex system into subsystems or steps, for example, in the case of food, to focus on the quantification of specific constituents or on the modelling of a simplified process, such as thermal degradation or ageing, at laboratory scale, discarding the food processing and the production chain. In the second case, prior knowledge is used, after data reduction by EMDA, to interpret the patterns which appear and validate possible conclusions which will guide to new hypothesis generation. In the first case, interactions among the reduced subsystems or steps are lost and, at most, the *a priori* hypothesized mechanistic behaviour may be confirmed or rejected; reformulation of the hypothesis will require *a priori* adoption of a different causal model. Differently, in the second case, the salient features of the system under investigation as a whole, including interactions, interconnections and peculiar behaviours, are learned from data by comparing conclusions induced by graphs to prior knowledge; it is then possible to validate the model and new hypotheses can be generated, so that an interactive cycle of multivariate experiments planning, multivariate systems characterization and multivariate data analysis is enabled.

p0130 Which food area would require explorative multivariate data analysis tools? We have seen in the introduction section that food science today embraces a wide multidisciplinary ambit, involving chemistry, biology/microbiology, genetics, medicine, agriculture, technology and environmental science, and also sensory and consumer analysis as well as economy.

p0135 Moreover, the investigation of the food production chain in an industrial context requires the assessment of not only the chemical/biological parameters but also the process parameters, irregularities, the influence of raw materials, etc. From an industrial perspective, the goal is not to produce a given product with constant technology and materials, which is impossible in practice, but rather to be able to control the specific, transient traits of production in order to ensure the same product quality.

p0140 Accordingly, the data used for food and food-processing characterization are changing [9–11] from traditional physical or chemical data, such as conductivity, thermal curves, moisture, acidity and concentrations of specific chemical substances, to fingerprinting data. Examples of this kind of data range from chromatograms or spectroscopic measurements, that is, complete spectra obtained by infrared (IR) [12–14], nuclear magnetic resonance (NMR) [15,16], mass spectrometry (MS), ultraviolet–visible (UV–vis) or fluorescence spectrophotometry, to landscapes obtained by any hyphenated combination of the previous techniques [17–21]; from signals obtained by means of sensor arrays such as electronic noses or tongues [22], microarrays and so on to imaging and hyperspectral imaging techniques [23–25].

p0145 The nature of this kind of data, the need to consider the many sources of variability due to the origin of raw materials, seasonality, agricultural

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:11:59:58 Page Number: 9

B978-0-444-59528-7.00003-X, 00003

practices and so on, together with the objective of studying the complex food processes as a whole, explain why EMDA is mandatory.

p0150 Multivariate screening tools are needed in order to model the underlying latent functional factors which determine what happens in the examined systems, and are the basis for an exploratory, inductive data strategy.

p0155 These tools have to accomplish two tasks:

o0020 **i.** Data reduction, that is, compression of all the information to a small set of parameters without introducing distortion of data structure (or at least keeping it to a minimum and maintaining control of the disturbance that has been introduced);

o0025 **ii.** Efficient graphical representation of the data.

p0170 By far the most effective techniques to achieve these objectives are based on projection techniques, that is methods to project the data from its *J*-variables/conditions space to lower dimensionality, that is, *A*-latent factors/components space. The commonly most used one is principal component analysis (PCA) and its extensions.

s0030 3.1 Principal Component Analysis

p0175 An exhaustive description of PCA historical and applicative perspectives, including a comparative discussion of PCA with respect to related methods, has been given by Joliffe [26] and Jackson [27]; other basic references are the dedicated chapters in Massart's book [28] and *Comprehensive Chemometrics* [29], and a more didactical view, with reference to the R-project code environment may be found in Varmuza [30] and Wehrens [31]. A description of PCA strictly oriented to spectroscopic data may be found in the *Handbook of NIR Spectroscopy* [32], Beebe [33] and in Davies' column in *Spectroscopy Europe* [34,35]; other salient references are Wold *et al*. [36] and Smilde *et al*. [37].

p0180 Here, PCA will be presented as a basic multivariate explorative tool with emphasis on the data representation and interpretation aiming at giving practical guidelines for usage in this specific context; the reader is referred to the literature cited earlier for more details.

p0185 PCA is a bilinear decomposition/projection technique capable of condensing large amounts of data into few parameters, called principal components (PCs) or latent variables/factors, which capture the levels, differences and similarities among the samples and variables constituting the modelled data. This task is achieved by a linear transformation under the constraints of preserving data variance and imposing orthogonality of the latent variables.

p0190 The underlying assumption is that the studied systems are 'indirectly observable' in the sense that the relevant phenomena which are responsible for the data variation/patterns are hidden and not directly measurable/observable. This explains the term latent variables. Once uncovered, latent variables (PCs) may be represented by scatter plots in a Euclidean plane.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:00 Page Number: 10

B978-0-444-59528-7.00003-X, 00003

p0195 An almost unique feature of PCA and strictly related projection techniques is that it allows a simultaneous and interrelated view of both samples and variables spaces, as it will be shown in detail in the following section.

p0200 For clarity of presentation, the PCA subject will be articulated in subsections: definition and derivation of PCs, including main algorithms; preprocessing issues; PCA in food data analysis practice.

s0035 3.1.1 Definition and Derivation of PCA

p0205 PCA decomposes the data matrix as follows:

$$\mathbf{X}_{(I,J)} = \mathbf{T}_A \cdot \mathbf{V}_A{}^{\mathrm{T}} + \mathbf{E}_{(I,J)} \quad (1)$$

where A is the number of components, underlying structures or 'chemical' (effective) rank of the matrix; the score vectors, $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_A]$, give the coordinates of samples in the PC space, hence score scatter plots allow the inspection of sample similarity/dissimilarity, and the loadings vectors, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_A]$, represent the weight with which each original variable contributes to the PCs, so that the correlation structure among the variables may be inspected through loading scatter plots. $\mathbf{E}$ is the residual, or noise or error matrix, the part of the data which was not explained by the model; it has the same dimensions as $\mathbf{X}$ and it is often used as a diagnostic tool for the identification of outlying samples and/or variables.

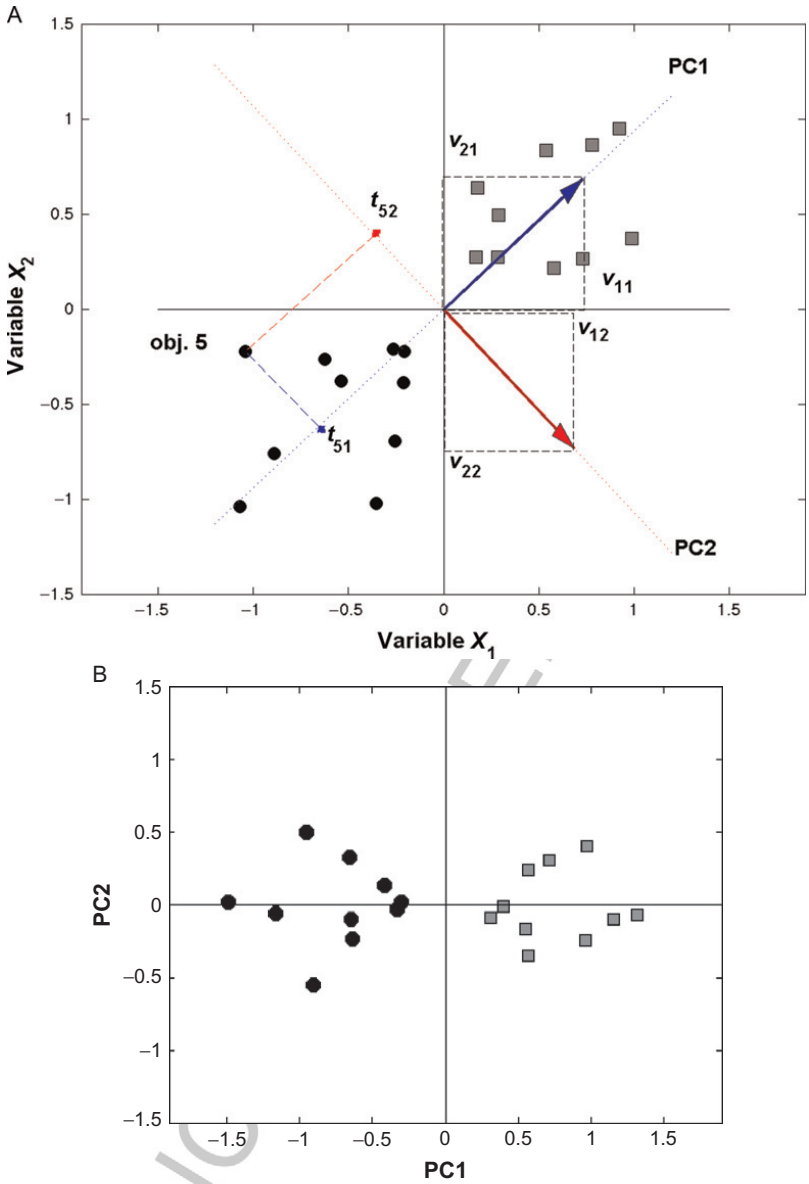
p0210 From a geometrical point of view, PCA is an orthogonal projection (a linear mapping) of $\mathbf{X}$ in the coordinate system spanned by the loading vectors $\mathbf{V}$. Figure 3A reports an example of a set of samples characterized by two variables $\mathbf{x}_1$ and $\mathbf{x}_2$, projected on the straight lines defined by the loading vector $\mathbf{v}_1$ and $\mathbf{v}_2$. For each of the I samples, a score vector $\mathbf{t}_i$ is obtained containing the scores for the sample (i.e. the coordinates on the PC axes).

p0215 Considering the projection of these samples on the PC space (Figure 3B), it emerges that the two categories (black circle and grey squares, respectively) are well separated on the first PC, while the second PC describes mainly non-systematic variability; thus one component ($A = 1$) is sufficient to retain information on this set of data.

p0220 Thus, PCA operates a reduction of dimensions from the number of variables J in $\mathbf{X}$ to A underlying virtual variables describing the structured part of data. Hence, a representation of the scores by means of 2D or 3D scatter plots allows an immediate visualization of where the samples are placed in the PC space, and makes the detection of sample groupings or trends easier (Figure 4A–C). The loadings represent the weight of each of the original variables in determining the direction of each of the PCs or, which is the same as PCs are defined as the maximum variance directions, which of the original variables varies the most for the samples with different score values on each of the components. A 2D or 3D plot of the loadings can be read as follow: variables that present loadings, which are equal or have close values, result correlated (anti-correlated if the signs

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:02 Page Number: 11

B978-0-444-59528-7.00003-X, 00003



f0015 **FIGURE 3** Geometry of PCA. A simulated example with 20 samples characterized by two variables. (A) The samples are plotted in the space of the original variables x_1 and x_2. The blue (grey, dashed) and red (black, dot–dashed) lines represent the directions of PC1 and PC2 axes, respectively. The coordinates of the blue (grey) arrow are v_{11} and v_{21}, the loading values of variables x_1 and x_2 on the first PC, respectively. The coordinates of the red (black) arrow are the v_{12} and v_{22}, the loading values of variable x_1 and x_2 on the second PC, respectively. The scores values are the orthogonal projection of the sample coordinates on the PC axes, as an example the scores of sample 5 are shown: t_{51} (PC1 score) and t_{52} (PC2 score). (B) The 20 samples represented in PC space: PC1 versus PC2. (For interpretation of the references to colour in this figure legend, the reader is referred to the online version of this chapter.)

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:04 Page Number: 12

B978-0-444-59528-7.00003-X, 00003

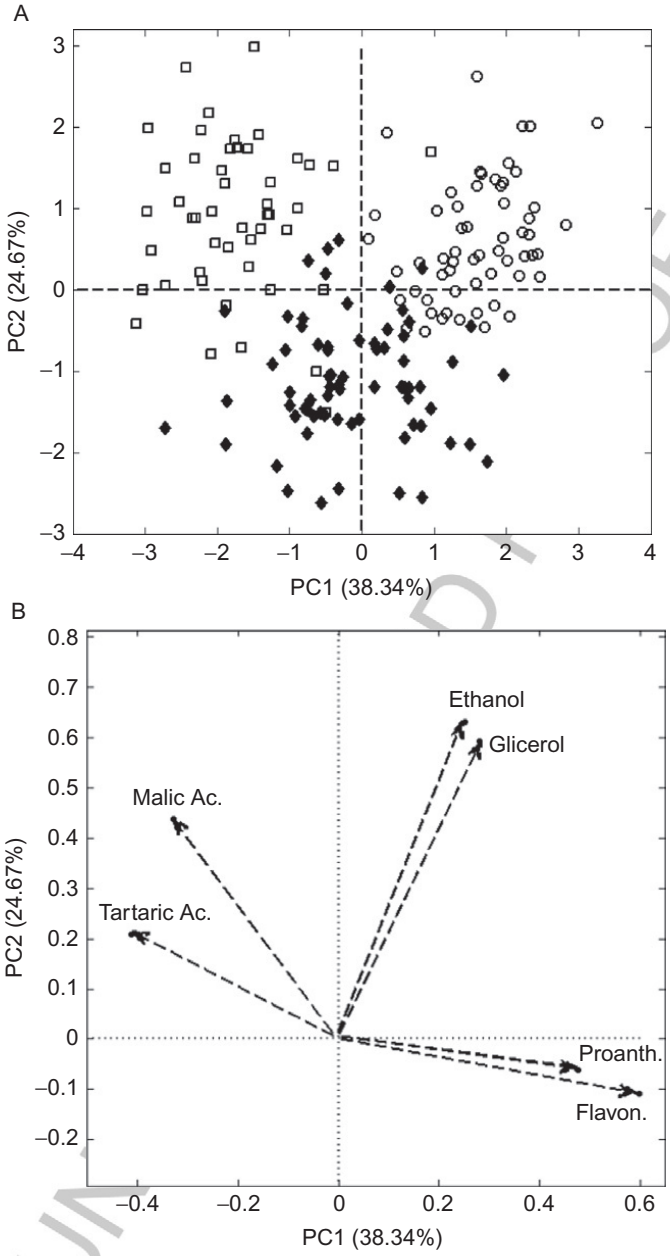
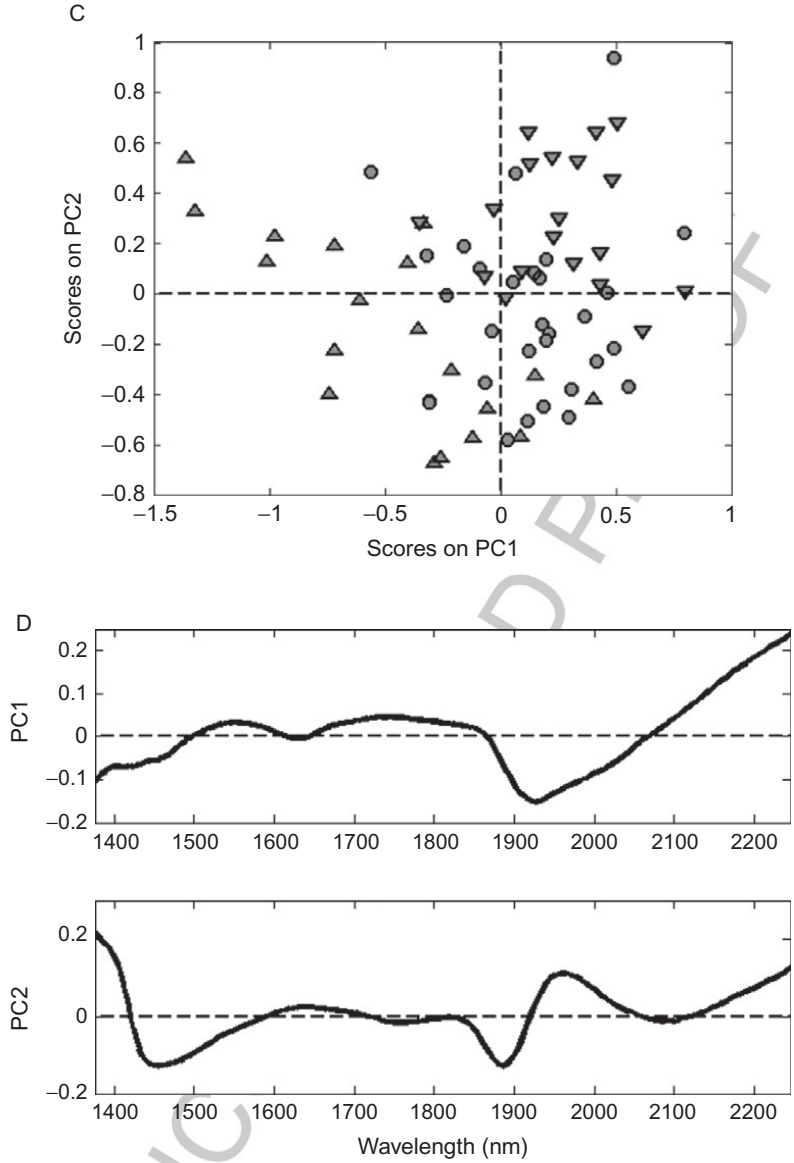


FIGURE 4—Cont'd

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:14 Page Number: 13

B978-0-444-59528-7.00003-X, 00003



f0020 **FIGURE 4** Examples of PCA. PCA of concentrations of six chemical compounds determined in samples of wine from three different cultivars. (A) PC1 versus PC2 scores plot (cultivar E: squares; cultivar B: circles; cultivar G: black diamonds). (B) PC1 versus PC2 loadings plot. PCA of NIR signals acquired at different leavening times of dough bread obtained from several flour mixtures. (C) PC1 versus PC2 scores plot. Upward triangles: beginning of the leavening (0–10 min); circles: middle time (10–40 min); downward triangles: end of the leavening (40–60 min). A slight trend with leavening time can be observed from negative to positive values of PC1 and towards positive values for PC2 at the end of leavening. (D) Loadings on PC1 (top) and loadings on PC2 (bottom): the separate visualization helps interpreting which spectral region influences each component the most.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:17 Page Number: 14

B978-0-444-59528-7.00003-X, 00003

are opposite); an example is illustrated in Figure 4B. When dealing with instrumental signals it is usually impossible to visualize the loadings with scatter plots, and more information can be obtained by analyzing them component-wise, as shown in Figure 4D and E. In this way it is possible to obtain a profile which can be directly compared to the original signal, so that regions which are more important for that PC (higher absolute values of the loadings) can be individuated.

p0225 A loading plot can be discussed together with the corresponding score plot, that is, drawn for the same couple of PCs, or directly represented in the same figure, which is then named a biplot (the mathematics of a biplot, i.e. how to render coherent, in the same coordinate space, the scale of score and loading values will be discussed following the mathematical formulation of PCA). In this way, it is easier to explain the groupings or trends one may notice in the PC space in terms of the original variables, as shown in Figure 5A. Although the biplot representation for spectral data is hard to visualize, it is possible to highlight some spectral regions which are responsible for the separation of the process steps, as reported in Figure 5B (small crosses).

p0230 From an algebraic point of view, PCA can be formulated as a mathematical maximization problem with constraints. We have seen that PCs are a linear combination of the original variables:

$$\mathbf{t}_a = \mathbf{X} \cdot \mathbf{v}_a \tag{2}$$

where $\mathbf{v}_a$, the loadings vectors, are subjected to $\mathbf{v}_a{}^{\mathrm{T}}\mathbf{v}_a = 1$ (normalization), $\mathbf{v}_a{}^{\mathrm{T}}\mathbf{v}_b = 0$ (orthogonalization) and maximization of var($\mathbf{t}_a$); hence the expression to be maximized, for $a = 1 \ldots J$ is

$$(\mathbf{X}\mathbf{v}_a)^{\mathrm{T}}(\mathbf{X}\mathbf{v}_a) = \mathbf{v}_a{}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}_a = \mathbf{v}_a{}^{\mathrm{T}}\mathrm{cov}(\mathbf{X})\mathbf{v}_a \tag{3}$$

where 'cov' stands for covariance (assuming $\mathbf{X}$ has been column mean centred) and the solution can be formulated as an eigenvectors/eigenvalues problem, for each value of a:

$$\mathrm{cov}(\mathbf{X})\mathbf{v}_a = \boldsymbol{\lambda}_a\mathbf{v}_a \tag{4}$$

p0235 This means that the unknown values for the loadings correspond to the eigenvectors of the $\mathbf{X}$ covariance matrix and $\boldsymbol{\lambda}$ are the corresponding eigenvalues.

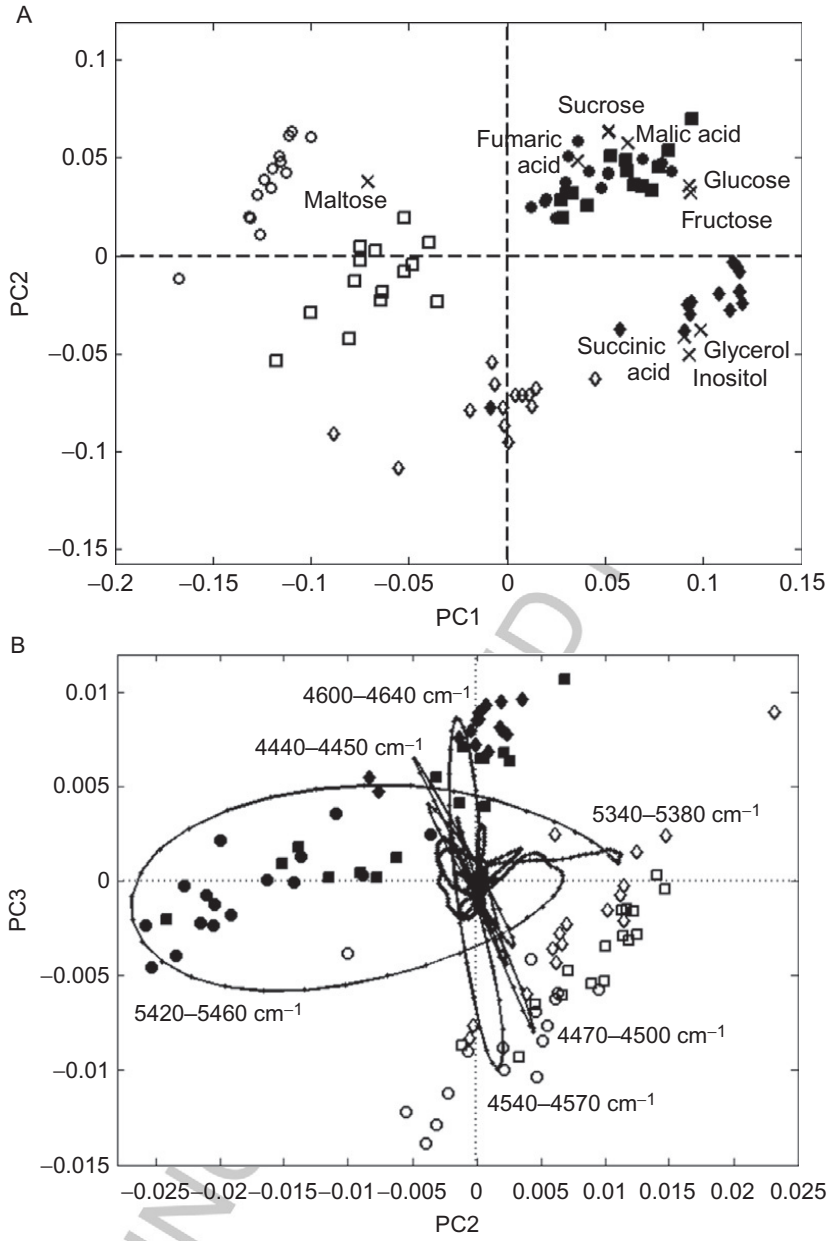
p0240 In other words, PCs calculation brings us to the diagonalization of the covariance matrix of $\mathbf{X}$, when $\mathbf{X}$ is column mean centred; in the case that $\mathbf{X}$ has been autoscaled (for autoscaling procedure, see Section 3.1.3), it brings us to the diagonalization of the $\mathbf{X}$'s correlation matrix.

p0245 As a consequence, PCs are sort in decreasing variance order and considering the algebraic property of the conservation of the trace, that is, for any non-singular square matrix, $\mathbf{B}$, given its diagonal $\mathbf{D}$, it holds: trace($\mathbf{B}$) = trace($\mathbf{D}$), the sum of the eigenvalues equals the total variance of the $\mathbf{X}$ matrix:

$$\sum\nolimits_a \lambda_a = \mathrm{var}(\mathbf{X}) \tag{5}$$

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:20 Page Number: 15

B978-0-444-59528-7.00003-X, 00003



f0025 **FIGURE 5** Examples of biplot representation. (A) GCbreadProcess data set. The biplot representation of PC1 versus PC2 allows assessing which chemical compounds (loadings, crosses) are more present in each of the six process steps monitored, visualized as classes for the scores according to the following coding: empty circles, S0; empty squares, S2; empty diamonds, S4; filled circles, D; filled squares, T; filled diamonds, L. In particular, all compounds, except for maltose, are more present in the second phase (filled symbols), and, in both phases, the content of sucrose, glucose and fructose decreases moving from S0 to S4 and D to L, respectively, while succinic acid, glycerol and inositol increase. (B) NIRbreadProcess data set. Scores have been represented according to the same code as in (A). Loadings correspond to each of the 1336 wavelengths recorded in the NIR signal (small points).

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:22 Page Number: 16

B978-0-444-59528-7.00003-X, 00003

p0250 In the case that $\mathbf{X}$ has been autoscaled and is full rank, the eigenvalues sum up to the number of variables J.

p0255 Adopting this derivation, the loadings can be obtained by any method for eigenvectors/eigenvalues calculation. Then, score vectors are obtained from

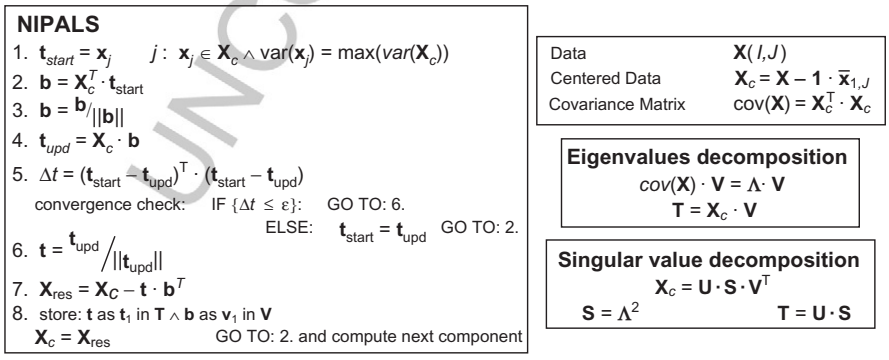
$$\mathbf{T} = \mathbf{X} \cdot \mathbf{V}^{\mathrm{T}} \tag{6}$$

p0260 The main algorithms used for eigenvectors/eigenvalues computation differ in two aspects: the matrix to work on, either $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ (eigenvalue decomposition (EVD) and the POWER method) or $\mathbf{X}$ (singular value decomposition (SVD) and non-linear iterative partial least squares (NIPALS)). However SVD may work as well on $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ (giving the same results as eigenvalue decomposition). Another difference is whether PCs are obtained simultaneously (EVD and SVD) or sequentially (POWER and NIPALS): for details and comparison of efficiency see Wu *et al*. [38]. In all the cases for which rows dimension I is much smaller than columns dimension J, one can operate on $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ instead (EVD, POWER, SVD), and on $\mathbf{X}^{\mathrm{T}}$ (NIPALS).

p0265 The two most widely used algorithms, NIPALS [39,40] and SVD [41,42], are schematically depicted in Figure 6, where the equivalence of loadings, eigenvalues and scores is also illustrated.

p0270 The main advantage of using NIPALS is in it being sequential, so that, especially when J is much larger than I (fat data matrices, such as with spectroscopic or chromatographic data), it can be stopped after a few components are derived. Indeed, in EDA two to four PCs are often what is needed, and generally an automatic stopping criterion can be implemented in NIPALS such as a desired percentage of explained variance or the reaching of a monotonous trend in eigenvalues versus the number of components plot. A disadvantage, however, may be that convergence is not always ensured.

p0275 We have mentioned in the previous section that reporting scores and loadings values in the same graph, namely a biplot [26,37,43–45], is very useful to discuss sample trends as a function of variable importance and their synergy in determining them. Biplots are based on SVD:



f0030 **FIGURE 6** PCA derivation according to different algorithms.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:26 Page Number: 17

B978-0-444-59528-7.00003-X, 00003

$$\mathbf{X} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^{\mathrm{T}} \tag{7}$$

p0280 $\mathbf{U}(I \times A)$ and $\mathbf{V}(J \times A)$ are orthonormal and $\mathbf{S}(A \times A)$ is a diagonal matrix with elements equal to ${\lambda_a}^{1/2}$, where λ_a are the eigenvalues collected in the diagonal matrix $\mathbf{\Lambda}$, $\mathbf{V}$ is the loadings matrix and the product $\mathbf{U} \cdot \mathbf{S}$ gives the scores. We can rewrite Equation (7) as

$$\mathbf{X} = \left(\mathbf{U} \cdot \mathbf{\Lambda}^{1/4}\right) \cdot \left(\mathbf{\Lambda}^{1/4} \cdot \mathbf{V}^{\mathrm{T}}\right) \tag{8}$$

p0285 We have in this way weighted 'scores' and 'loadings' equally by the eigenvalues making the lengths of objects and variables vectors in the biplot approximately equal, thus, in a biplot, for the corresponding components, $\mathbf{T}^* = (\mathbf{U} \cdot \mathbf{\Lambda}^{1/4})$ and $\mathbf{V}^* = (\mathbf{\Lambda}^{1/4} \cdot \mathbf{V}^{\mathrm{T}})$ are plotted simultaneously. A further option is to take into account the difference in dimensionality between rows and columns, and use a normalizing factor, for example, $(I/J)^{1/4}$ and $(J/I)^{1/4}$ factors for $\mathbf{T}^*$ and $\mathbf{V}^*$, respectively [44].

p0290 This, also called symmetric scaling, is not the only adopted choice in biplot representation; alternatives are obtained considering the general expression, with α varying between 0 and 1:

$$\mathbf{X} = \left(\mathbf{U} \cdot \mathbf{\Lambda}^{\alpha/2}\right) \cdot \left(\mathbf{\Lambda}^{(1-\alpha)/2} \cdot \mathbf{V}^{\mathrm{T}}\right) \tag{9}$$

p0295 The symmetric scaling just described corresponds to $\alpha = 0.5$. For a detailed discussion on the implication of the different choices see Refs. [26,27,37,44]. The most common alternatives are $\alpha = 0$, in which case the plot gives Euclidean distance among variables and Mahalanobis distance among objects; and $\alpha = 1$, in which case the plot gives Euclidean distance among objects and Mahalanobis distance among variables.

s0040 3.1.2 Extracting Information from the PCA Model

p0300 A PCA model is determined once the number of PCs to be retained has been fixed. The maximum number of PCs that can be calculated corresponds to the mathematical rank of the data matrix (if data are not significantly correlated, the rank is $\min(I, J)$, otherwise it can be lower), but the interest is generally in recovering the 'chemical' rank, that is, the number of underlying phenomena, latent variables sufficient to describe the problem/system at hand.

p0305 When, as illustrated here, PCA is used as an EMDA tool, as the purpose is graphical representation/inspection of data, the matter of choosing an appropriate number of PCs, at first sight, does not seem so relevant. This is true, and not true, at the same time. True because it is always possible to calculate all components up to the rank and identify the most significant PCs by sequential graphical inspection of score plots. Not true because the residual $\mathbf{E}$ (the errors part or the not systematic variation in data) does constitute a relevant part of what we also want to know about our data, such as outliers, noise content. More generally, we may see

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:29 Page Number: 18

B978-0-444-59528-7.00003-X, 00003

PCA decomposition not only as 'data structure' + 'noise' but as 'pertinent information' + 'other structured variation' + 'noise'. Thus, establishing the number of components corresponding to 'pertinent information', and hence establishing a PCA model, is useful. It has to be stressed that by retaining all components no data reduction is operated (except for the compression from J to the mathematical rank) and noise is not estimated; only an orthogonal rotation of the variables space is accomplished.

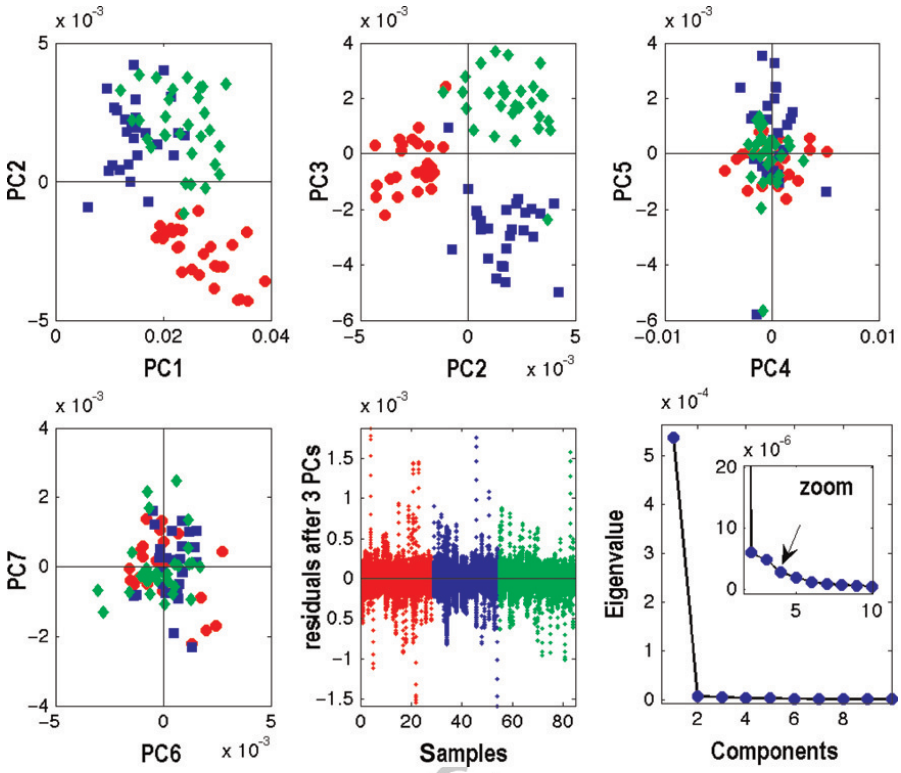
p0310 Several criteria and rules of thumb have been formulated [26,28,46] to answer the question: How many PCs? In EMDA, criteria based on statistical inference, that is, on formal tests of hypothesis, should be avoided as we do not want to assume, in the model estimation phase, our PCs to follow a specific distribution. In this context, more intuitive criteria, albeit not formal, but simple and working in practice, are preferable, especially graphics-based criteria, such as sequential exploration of scores plots and/or inspection of residuals plots; plots of eigenvalues (scree plots [47]) or cumulative variance versus number of components. Different consideration holds when PCA is used to generate data models that are further used, for example, for regression, classification tasks or process monitoring [48,49] (Section 3.1.5), where PCA model validation, for example, by cross-validation, in terms of performance on the assessment of future samples has to be taken into account.

p0315 By exploring scores plots for subsequent components, the number of PCs can be evaluated on the basis of data structure description and one can stop when no further salient information about sample patterns is gathered (Figure 7). Residual plots allow checking if some systematic variation is left in the unmodelled part of the data; in general, residuals should be normally distributed around zero with no specific trends (Figure 7, bottom third from left). The reasoning behind the use of scree or cumulative variance plots is that components describing systematic variation will both account for a larger portion of data variance and are not likely to account for an equal amount of variance each. Instead, components describing unsystematic variability, 'noise', reflect a situation of equivalent captured variance in almost all directions in the space of the original variables, that is, the situation of randomly or uniformly distributed data. Thus a change from steep to shallow slope in the line connecting the points reported in a scree graph can be considered to correspond to an optimal number of components. In Figure 8, for the same data set shown in Figure 3 the scree plot, reporting both eigenvalue and log(eigenvalues), the cumulative variance and the eigenvalue ratio plot are compared; the suggested number of components is 2 (classical scree) or 3 (scree variant) and 4 for cumulative variance. Two PCs, considering that now data have been centred, are sufficient to extract category-related information.

p0320 Another simple rule is retaining a number of components corresponding to a given percentage of accounted variance, for example, 80–90%. In this case, the kind of data and the type of pretreatment (Section 3.1.3) have to be taken into account: typically, if the data matrix is not centred, the first component

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:30 Page Number: 19

B978-0-444-59528-7.00003-X, 00003



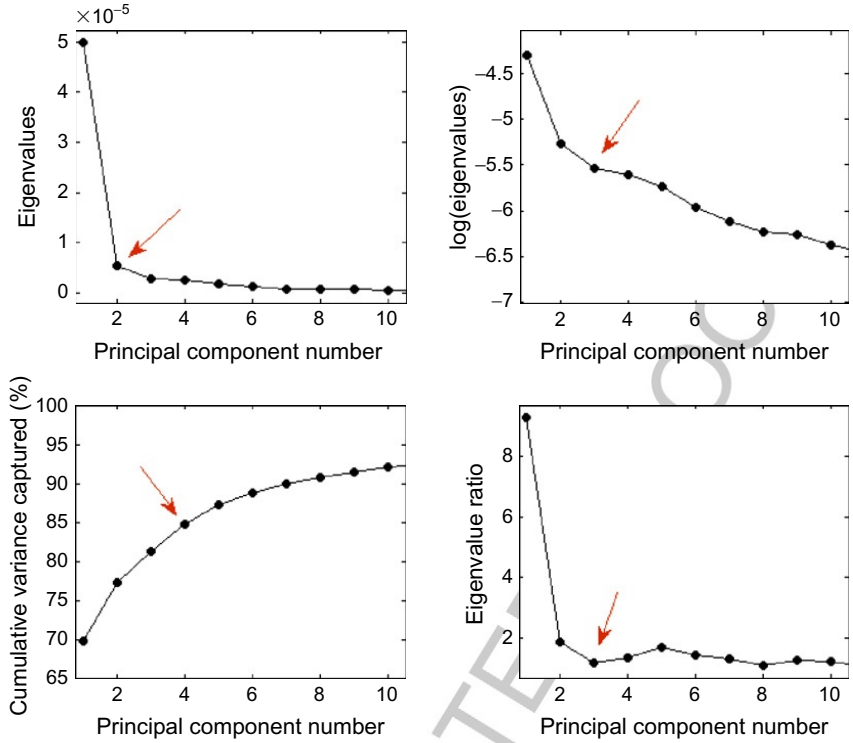
f0035 **FIGURE 7** The analysed data set consists of three categories of animal feed characterized by NIR spectroscopy. Data are first derivative spectra but are not centred. Scores plots for subsequent components top and first left bottom. PC1 accounts for 95.6% of data variance and describes just the distance from the origin (i.e. it resembles the average spectrum), PC2 (1%) and PC3 (0.9%) show the structure, distinguishing the three categories. PC4 describes some peculiar samples. Further PCs show almost uniform distribution. Residuals versus samples number, coloured according to category (bottom: second from left). Scree plot (bottom left) suggesting 3–4 PCs. (For colour version of this figure, the reader is referred to the online version of this chapter.)

will account for a very large percentage of data variance and will correspond to the average variable profile. Moreover, especially with some kind of spectral data, like near-infrared (NIR) signals, the interesting chemical variability may represent a very low portion with respect to other sources of physical variability: in these cases, if the sources of non-relevant variability have not been removed by pretreatment, patterns will emerge in the last PCs instead of the first few.

p0325 Other rules are based on numerical evaluation of the eigenvalues: it is assumed that in the case of perfect independence among variables, the PC will be the same as the original variables (PCA represents an invariant rotation of axes) and will account for unitary variance in case of autoscaled data, thus a PC with an eigenvalue less than 1 contains less information of one original variable and could be discarded (this rule sometimes is also taken as eigenvalues

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:30 Page Number: 20

B978-0-444-59528-7.00003-X, 00003



f0040 **FIGURE 8** Same set of data as in Figure 7 but PCA has been computed on centred data. Top left: scree plot, taking 2 PCs seems appropriate. Top right: logarithm of the eigenvalues versus PC number, the trend is smoothed and suggests taking 3 PCs. Bottom left: plot of cumulative variance versus PC number, a plateau is reached with 4 PCs (85% variance explained). Bottom right: ratio of eigenvalues, starting from eigenvalue PC1/eigenvalue PC2; suggestion is to stop at 3 PCs, as the ratio of 3 PCs/4 PCs is very small compared to the previous ones. (For colour version of this figure, the reader is referred to the online version of this chapter.)

less than 2). This rule has also been extended to not autoscaled data considering that PC accounting for a percentage variance less than $100*(V_{tot}/J)$, where V_{tot} is the total variance of the data set, can be discarded.

p0330 It may be argued that these criteria are subjective and it may be difficult for the user to take a decision. However, it should be remembered that EDA is always a subjective exercise composed of several steps: a chemometric tool may be wrong or right for a given purpose/set of data, but it will never be the only one that can be applied, thus the user will have the honour, and the burden, of experimenting different ones. The same holds true for the number of PCs to be retained: while it is mandatory to be aware of the consequences, in terms of how (which feature of) a data set will be modelled as a consequence of that choice, the choice itself is a responsibility the user has to take.

p0335 Once a PCA model is obtained, information may be retrieved from eigenvalues (explained variance, redundancy), scores (on samples, systems, conditions,

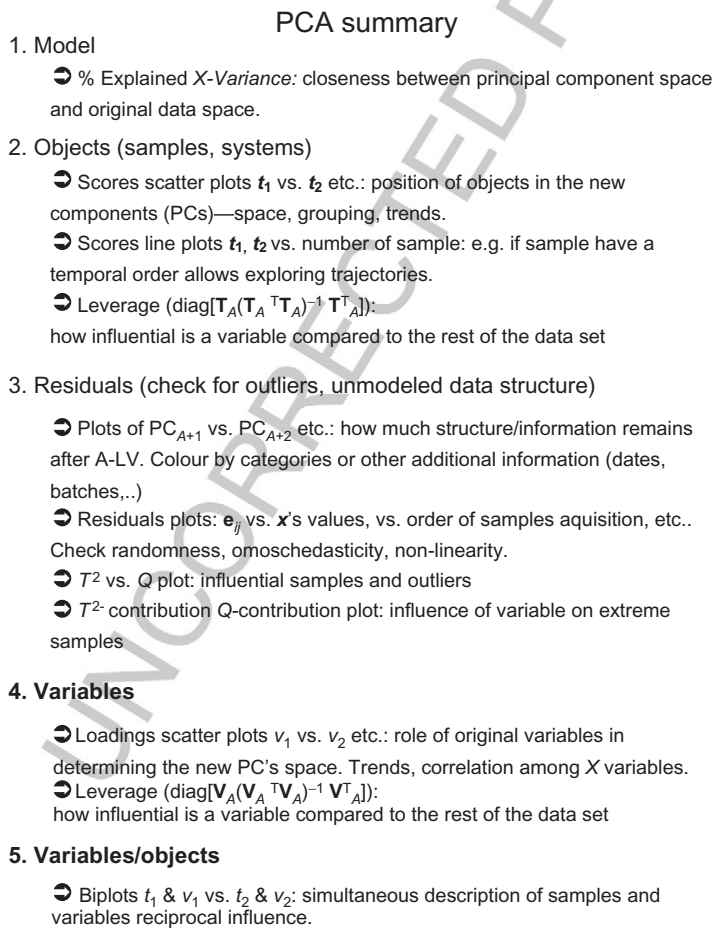
Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:44 Page Number: 21

B978-0-444-59528-7.00003-X, 00003

e.g. time, ageing, etc., depending on what is reported on data table rows), loadings (on variables, signal regions depending on the kind of data), biplot (reciprocal behaviour and trends in samples/variables) and residuals matrix **E** (anomalous/unmodelled samples and variables, model diagnostic). Figure 9 offers a schematic view and a reference summary.

p0340 Furthermore, in the specific case in which the PCA score plot reveals the absence of grouping and clusters, that is, the data set being analyzed is composed of samples of the same nature which represent the same system or population, it may be assumed that the calculated PCA model represents a homogeneous set of samples, a specific category, and such a question can be formulated: How fit is a sample (in the actual data set or a future one) to this model?

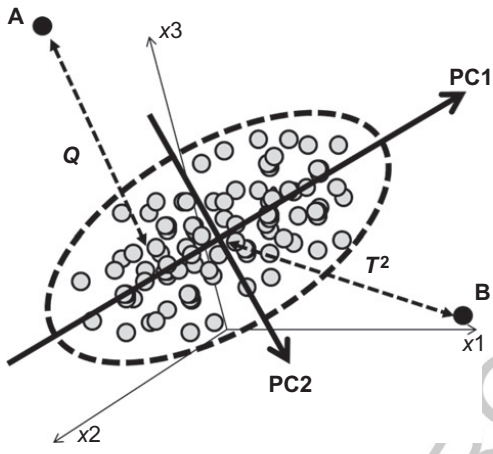
p0345 To get an answer it is useful to calculate two distances: the distance of a given sample from the PCA model hyper plane and from the centre of the



f0045 **FIGURE 9** Summary of PCA outputs.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:45 Page Number: 22

B978-0-444-59528-7.00003-X, 00003



f0050 **FIGURE 10** Graphical representation of the PCs space for a two-component model on a three-variable data set.

model (Figure 10). The sum of squared residuals for each sample, here named Q (other commonly encountered names are SPE, DModX), is a measure of the distance of a sample from the PCA model (i.e. the higher Q is, the lower the fit of the model). Q is the sum of squares of each row (sample) of $\mathbf{E}$; that is, for the ith sample in $\mathbf{X}$, $\mathbf{x}_i$:

$$Q_i = \mathbf{e}_i \cdot \mathbf{e}_i^{\mathrm{T}} \tag{10}$$

where $\mathbf{e}_i$ is the ith row of $\mathbf{E}$. Q values indicate how well each sample conforms to the PCA model, that is, it is a measure of the difference, or *residual*, between a sample and its projection into the A PCs retained in the model (the distance of point A, Figure 10, from the PC's plane).

p0350 The squared elements of a single row of the, I by J, $\mathbf{E}$ matrix, $\mathbf{e}_i^2$, represent the Q contributions for a given sample, which is an indication of how much each variable contributes to the overall Q for the sample, and is particularly useful in identifying the variables which contribute most to a given sample's sum-squared residual error. To retain information about the sign of the deviation for a given variable, in some of the most common softwares it is possible to find some implementations, such as $\mathrm{sign}(\mathbf{e}_i) * \mathbf{e}_i^2$, or simply $\mathbf{e}_i$, instead of $\mathbf{e}_i^2$ representing the Q contribution plot.

p0355 The sum of normalized squared scores, T^2, known as Hotelling's T^2 statistic [50], is a measure of the variation in each sample within the PCA model (the distance of point B, Figure 10, from the centre of the PC's plane). T^2 is defined as

$$T_i^2 = \mathbf{t}_i \lambda^{-1} \mathbf{t}_i^{\mathrm{T}} \tag{11}$$

where $\mathbf{t}_i$ refers to the ith row of $\mathbf{T}$, the scores matrix from the PCA model, and λ is a diagonal matrix containing the eigenvalues (λ_1 through λ_A) corresponding to the A eigenvectors (PCs) retained in the model. T^2 contributions describe

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:48 Page Number: 23

B978-0-444-59528-7.00003-X, 00003

how individual variables contribute to the distance of Hotelling's T^2 value for a given sample. The contributions to T_i^2 for the ith sample, $\mathbf{t}_{\mathrm{con},i}$, is a vector calculated from

$$\mathbf{t}_{\mathrm{con},i} = \mathbf{t}_i \lambda^{-1/2} \mathbf{P}^{\mathrm{T}} \tag{12}$$

and can be considered a scaled version of the data within the PCA model to equalize the variance captured by each PC.

p0360 Assuming a normal distribution of the scores (which may be reasonable in this specific case, as PCs are derived for a single category and PCs are in general more normally distributed than original variables), two statistics can be associated to the two distances: Q statistics [51,52] and Hotelling's T^2 statistics [50]; for a discussion and further details on these statistics and their application the reader is referred to Chapter XXX [53–55]. Au1

p0365 From an explorative point of view, the T^2 versus Q plot (Figure 11) allows the inspection of peculiar samples. When the confidence limits, calculated according to the respective statistics, are added, the plot is split into four regions:

o0030 **I.** The model space (bottom left) where normally behaving samples belong;

o0035 **II.** The region of extreme samples (bottom right), which show an extreme behaviour because they respect the variable correlation structure captured by the PCA model but get high values in scores space. These samples, whose T^2 values are high, are also said to have high leverage because they pull the PC axes towards them;

o0040 **III.** The region far from model samples (top left): these samples, whose Q values are high, look 'well behaving' once projected on model space, because they share some features with the modelled category, but are not well modelled because part of their variation is not accounted for by the model (e.g. a sample which has the same composition as the modelled category but contains a chemical compound which is not present in the other ones);

o0045 **IV.** The outliers region (top right), where all the anomalous, extreme and not modelled samples belong, having both T^2 and Q high values.

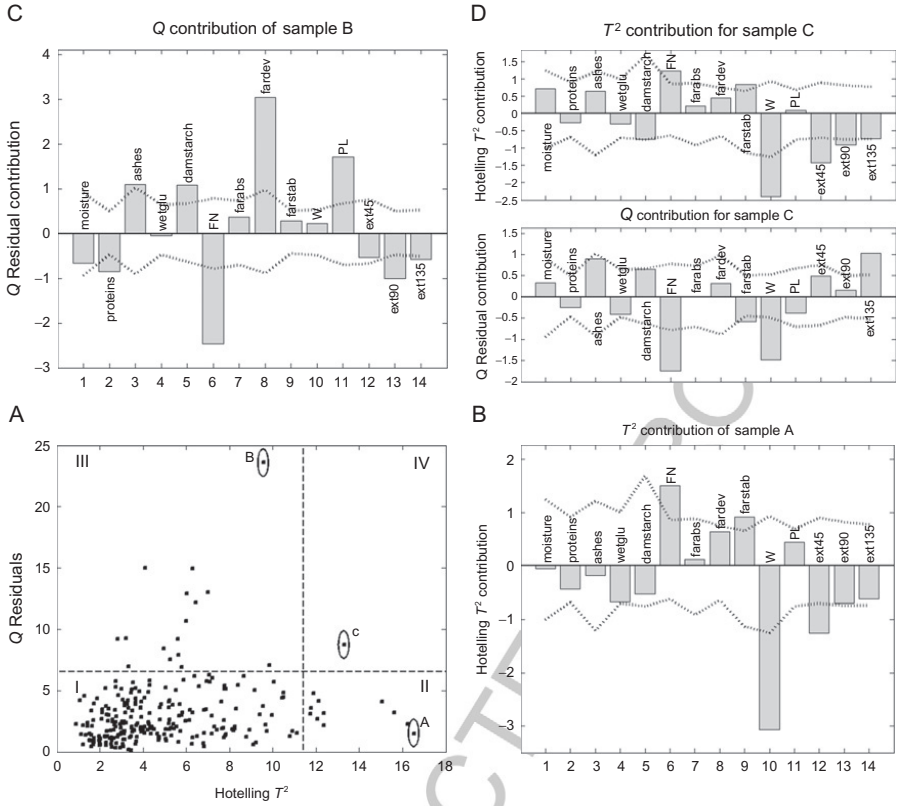
p0390 Moreover, by means of contribution plots [56] it is possible to come back to the original variables and their contribution to each distance, thus understanding why the samples behave differently or extremely with respect to the others. The use of confidence limits [57], additionally, can help in the identification of which contribution is statistically significant: often, considerations on the highest absolute value are not sufficient to highlight contributions for variables which have a wide variability range and also for not extreme samples, as can be seen in Figure 11.

s0045 *3.1.3 Data Pretreatment*

p0395 Explorative data analysis aims at looking at/into data. It is a common experience that things may be seen from different angles and perspectives, and

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:50 Page Number: 24

B978-0-444-59528-7.00003-X, 00003



f0055 **FIGURE 11** Representation of T^2 versus Q values for the samples in the PCA model of Flour-Rheo Data set and contribution plots to the distances for some illustrative samples. Confidence limits (dotted black lines) are computed at 95%; for contribution plots they are based on the 5th to 95th percentile range of values for samples in region I. (A) The T^2 versus Q plot: the four regions of the graph (I–IV) are described in the text. (B) Contribution to T^2 distance for sample A in region II: the sample presents an extreme value for parameters W and ext45 (lower value than the mean of the model), FN and farstab (higher value). (C) Contribution to Q distance for sample B in region III: correlation structure is not respected for several parameters, for example, FN, fardev and PL. (D) Contribution to T^2 and Q distances for sample C in region IV: the sample presents extreme values for several properties (e.g. W, FN and ext45) and correlation structure is not respected mostly by FN, W and ext135.

resolution changes depending on the distance from which things are observed, illumination and so on. The way to introduce different perspectives and resolution in looking at data is to apply data pretreatments or preprocessing. In this respect, EMDA offers different views depending on which data set has been given as input, raw or processed according to the type of processing, and by comparing these different views it also becomes possible to assess the effects the applied pretreatment has introduced. Thus EMDA results, while depending on data pretreatment, also provide a diagnostic tool to orient pretreatment choice.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:51 Page Number: 25

B978-0-444-59528-7.00003-X, 00003

p0400 Here, the subject will not be entirely covered, but the main data pretreatment tools will be illustrated following the two general cases, considering the data table is organized as samples/objects in the rows and variables in the columns, rows and columns pretreatments.

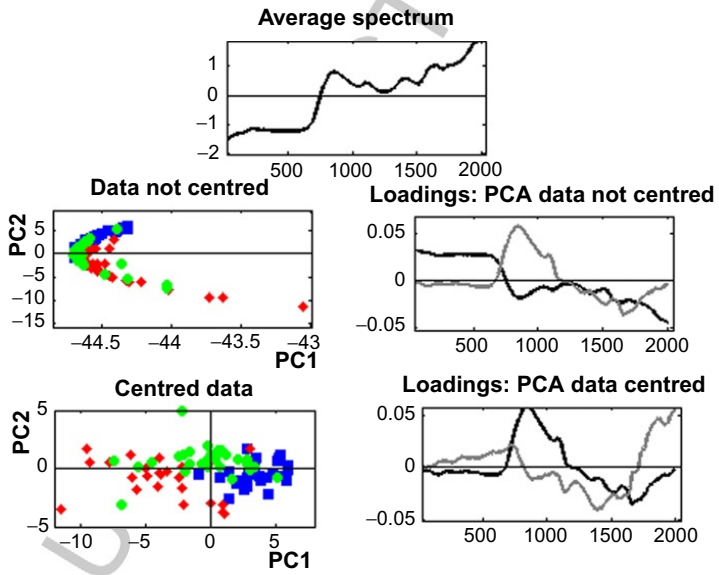
s0050 ### 3.1.3.1 Column Pretreatment

p0405 These pretreatments include data centring and scaling.

p0410 Data centring across columns consists of subtracting a constant term, also called offset, for all samples (rows) from each variable:

$$\overline{x_j} = \sum_i x_{ij}/I \tag{13}$$

p0415 From a geometric point of view this operation corresponds to setting the centre of the coordinates system (both variables and PCs spaces) equal to zero. From a practical point of view this means to look at data from the inside of the data, not from a distance; this is illustrated in Figure 12 where the scatter plot of the first two PCs is shown before and after columns centring the data. Or, in other words, if we considered the variables average as a rough summary of our samples data, centring may help focusing on the 'differences' discarding what is a common pattern. Columns mean centring keeps the distances among samples in variable/PCs space unchanged.



f0060 **FIGURE 12** The analysed data set consists of three categories of animal feed characterized by NIR spectroscopy. Spectra are SNV preprocessed. The first PC loadings of not centred data (middle right, black line) closely resemble the average spectrum (top). The PC1 versus PC2 scores plot (middle left) shows the distance from the 0,0 point. The first PC loadings of centred data (bottom left, black line) are like the PC2 loadings (middle left, grey line) of not centred data. (For colour version of this figure, the reader is referred to the online version of this chapter.)

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:54 Page Number: 26

B978-0-444-59528-7.00003-X, 00003

p0420 Thus, centring may be generally useful, in situations where the previously mentioned considerations hold true. When centring is appropriate, data rank is generally reduced. Also, centring may increase the efficiency of some algorithms for PCA estimation, such as NIPALS or POWER method.

p0425 Of course, centring is not a sensible choice when it is meaningful to consider the distance of the objects from the zero origin, or we are interested to obtain models under non-negativity constraints, for example, because 'chemically' interpretable scores values in terms of chemical concentrations are needed.

p0430 Mathematically speaking, centring may be seen as projecting the data onto a space where the common offset represented by the column average is removed [58]. Fitting a PCA model to centred data can be seen as a two-step procedure: first, the average is removed; and second, the model is fitted and model parameters, for example, loadings and scores, are estimated. This may have some drawbacks in specific situations, for example, when dealing with incomplete data with missing elements. In this case, removing averages and then fitting a model is non-optimal and may alter the data structure, producing wrong results, as wrong data dimensionality estimation.

p0435 In food data analysis practice, one aspect has to be taken into consideration: the samples often belong to different conditions, such as same food sampled by different producers or at different locations and/or at different times (seasons, years) and so on. In these cases, different views/information are obtained by simply column centring or centring separately each 'block' of samples, even if the resulting column averages are zero in both cases. An example is given in Figure 13, where samples are casks of balsamic vinegar at different ageing (the time order, from the youngest to the oldest, is indicated by symbols: black asterisk, light blue cross, magenta triangle, green diamond, red square, blue circle, respectively) belonging to three different producers (C, M and S) characterized by organic sugars and acid content [59]. It can be seen how columns centring by subtracting the overall column average obscures the trends inside each series of casks: anchoring the variation to the overall average, the intra-producers differences are highlighted (Figure 13, top right). By subtracting the average for each series separately, the ageing trend within each series is better depicted (Figure 13, bottom left), which reflects typical features, such as the starting age of the series, the extent of yearly topping up and so on. Moreover, the PC1 versus PC2 scores plot now shows a distinction by cask ageing instead of by producer.

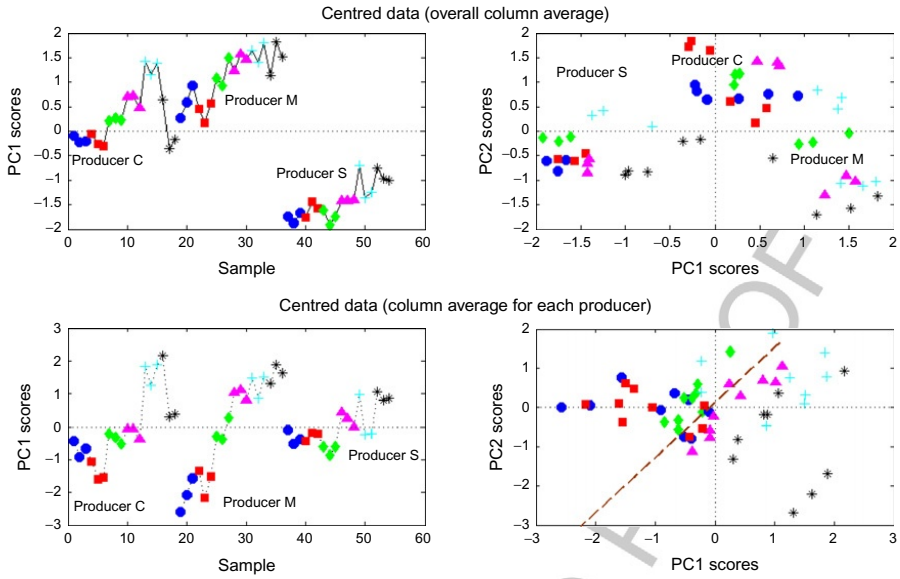
p0440 PCA seeks directions of maximum variance and variance depends on the measurement scale of the variables, thus it is important to focus our attention on the kind of variables we have measured to characterize our data: Are the scales comparable? Are we interested in allowing each variable the same chance to contribute to the PCA model? Which is the noise level?

p0445 Scaling will cope with these issues. Columns scaling means to apply a weight to each variable:

$$\mathbf{X}_{\text{scaled}} = \mathbf{X} \cdot \mathbf{W} \tag{14}$$

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:57 Page Number: 27

B978-0-444-59528-7.00003-X, 00003



f0065 **FIGURE 13** Scores plot of PCA of balsamic vinegar at different ageing corresponding to two different centring procedures. Top plots: overall average removed; bottom plots: series average removed. The different coloured symbols indicate ageing of each cask (three replicates for each one): the time order, from youngest to oldest, corresponds to the following symbols: black asterisks, light blue cross, magenta triangles, green diamonds, red squares, blue circle). Letters indicate the three different producers: C, M and S. The dashed oblique line in the bottom right figure highlights a possible separation by cask ageing of three oldest (above the line) from the three youngest (below the line). (For interpretation of the references to colour in this figure legend, the reader is referred to the online version of this chapter.)

p0450 The weights matrix $\mathbf{W}$ is a diagonal matrix of dimension $J \times J$ whose diagonal elements are weights to be applied to each column. The main purpose of scaling is to change the importance attached to different parts of the data in fitting the model. Thus, in general, scaling may be seen as a way to introduce our knowledge about the nature of the variables, their relevance to our data description and so on. The choice of the type of weight to apply depends on the specific data set and our aims. Three main objectives are pursued by scaling: (1) to adjust scale differences; (2) to take into account noise level; and (3) to consider in the same data set variables differing in size and/or kind, for example, punctual and spectral variables altogether.

p0455 *Case 1*. The first case arises because a PCA model is based on describing data variance: as a consequence, the variables showing large variation are implicitly important, that is, more important than the others. The point is, why do those variables have a bigger variability? In case the variation is solely due to a matter of scale (units of measure) or if it reflects the different amount of compounds (the presence of major and minor constituents albeit all potentially interesting to describe/differentiate the studied samples), it is fair to scale variables. On the contrary, if the variation is only due to noise

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:00:59 Page Number: 28

B978-0-444-59528-7.00003-X, 00003

(which has to be estimated by replicate measures, unless the uncertainty associated with the specific method of measurement of the given variable is known by previous experimentation), the result of scaling enhances unimportant variables, thus leading to a poorer model. In order to adjust for scale differences, it is common to use as variables weights (scaling factors) a measure of the data dispersion, such as:

u0015 – The inverse of the standard deviation (scaling to unit standard deviation or unit variance);

u0020 – Pareto scaling [60] (uses the inverse of the square root of the standard deviation, has an effect, that is, somehow intermediate between using raw variables and scaling, in the sense that large variance variables are less down-weighted);

u0025 – Range scaling [61] (uses the inverse of variable range as weight, thus resulting more sensitive to outliers, a problem which can be avoided by using robust range estimators);

u0030 – Va.st. (variable stability) scaling [62] (it consists of down-weighting those variables that are the least stable and focusing on stable variables, using the standard deviation and the so-called coefficient of variation as scaling factor).

p0480 *Case 2*. In this case, scaling may be used to account for non-constant noise level (error variance), and the applied weight corresponds to the inverse of residual variance of each variable (it seems similar to unit variance scaling but instead of the standard deviation of the variable, the residuals standard deviation, as assessed by replicates, is used), down-weighting variables whose uncertainty is higher.

p0485 *Case 3*. In this case, subjective weights are given to variables. The most common case is the one in which it is necessary to consider subsets of variables of different nature, for instance concentration variables together with spectral variables, such as NMR fingerprint. Because the number of spectral variables is very high, and each one contributes to data variance, the model will focus on explaining only the spectral variables. To solve this issue, the spectral variables can be block-scaled [62] so that their total variation is set equal to the total variation of the metal content variables: in this way both blocks of data are given an equal chance of influencing data variance. The weights to accomplish block-scaling are defined as

$$w_{jB} = \sqrt{\frac{\mathrm{SS_{TOT}}}{\mathrm{SS_{BLOCK}} \cdot n_{\mathrm{BLOCK}}}} \tag{15}$$

where w_{jB} is the weight to be assigned to each variable in a given block, B; $\mathrm{SS_{TOT}}$ is the total sum of squares (alternatively, variance can be used) over all J's variables, $\mathrm{SS_{BLOCK}}$ is the sum of squares over the J's variables belonging to the given block and n_{BLOCK} is the number of variables inside the block.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:02 Page Number: 29

B978-0-444-59528-7.00003-X, 00003

p0490 Most often, block-scaling is coupled with scaling to unit variance, meaning that each variable is weighted by its inverse standard deviation, but a correction term taking into account the number of variables in each block is also included; that is, the weight to be applied to each variable is

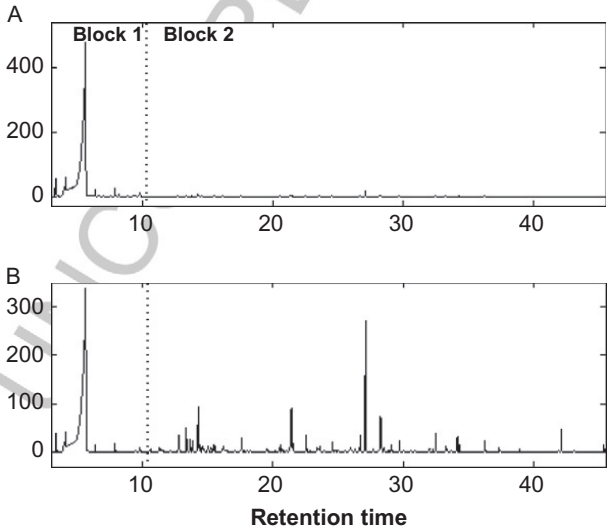
$$w_{jB} = \left[J/\left(\sigma_{jB} \cdot n_{\mathrm{BLOCK}}\right)\right]^{1/2} \tag{16}$$

p0495 Block scaling is also useful in the case where the variables are of the same nature, for example, chromatographic or spectral signals, but their variation reflects the presence of peaks/bands that account for constituents present at different concentration, such as main/secondary metabolites, metals/trace metals and so on. In this case, the blocks are defined as signal regions, not necessarily contiguous, each including peaks of similar variation together with some baseline regions (in order to avoid up-weighting of noise) as shown in Figure 14.

p0500 Scaling is most often coupled with centring; autoscaling, for instance, refers to columns centring plus dividing by columns standard deviation. The distance among samples is not preserved by scaling: Figure 15 shows how PCA scores and loadings are affected by data scaling (data set: FlourRheoData, see Section 3.1.4 for more information on it).

s0055 ### 3.1.3.2 Row Pretreatment

p0505 In this section pretreatments applied in the rows direction are discussed; this is also called data preprocessing and most applications concern data signals, in particular spectra. Here, we will refer mainly to the preprocessing of NIR



f0070 **FIGURE 14** GC-FID chromatograms of vinegar samples (A) before and (B) after block-scaling; two blocks were defined and are indicated by numbers and separated by vertical lines.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:02 Page Number: 30

B978-0-444-59528-7.00003-X, 00003

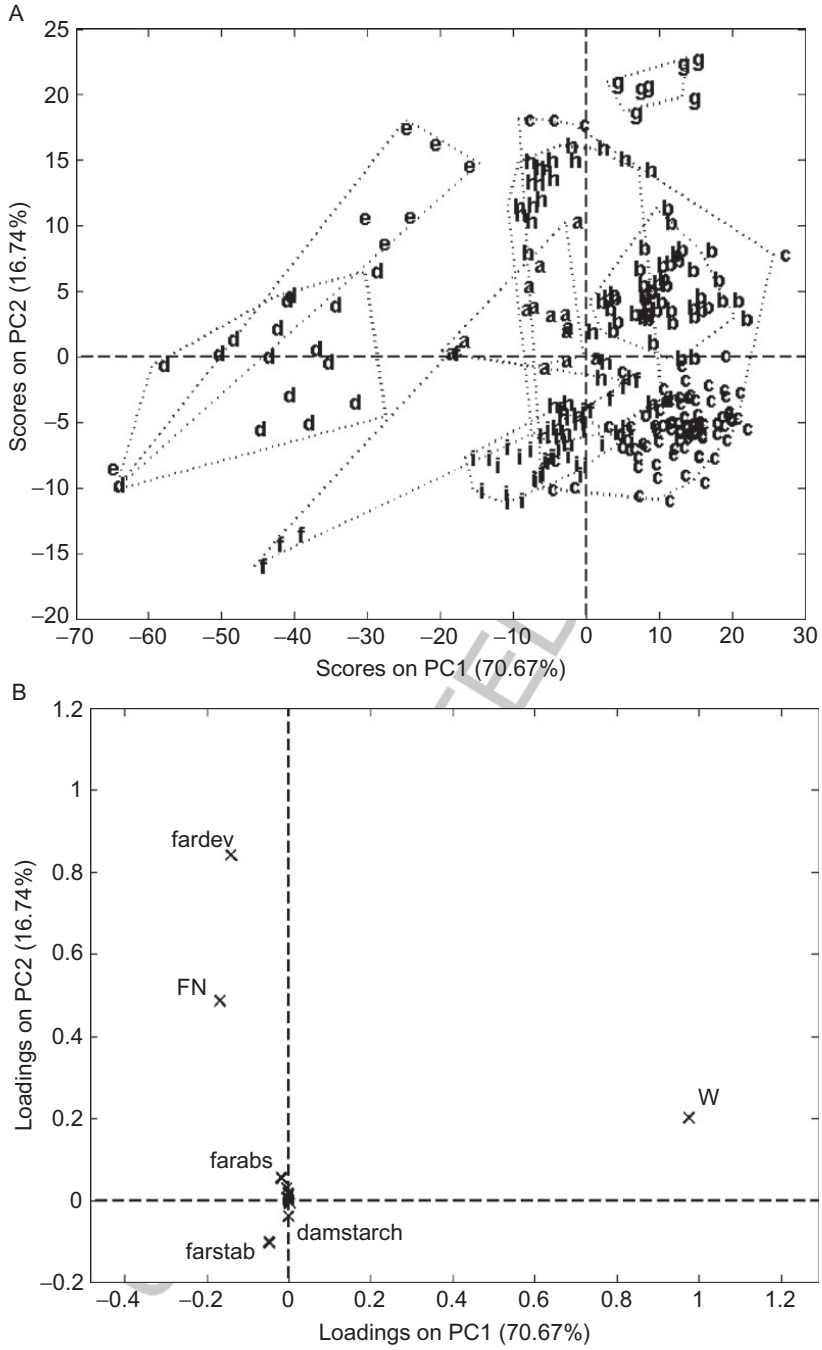
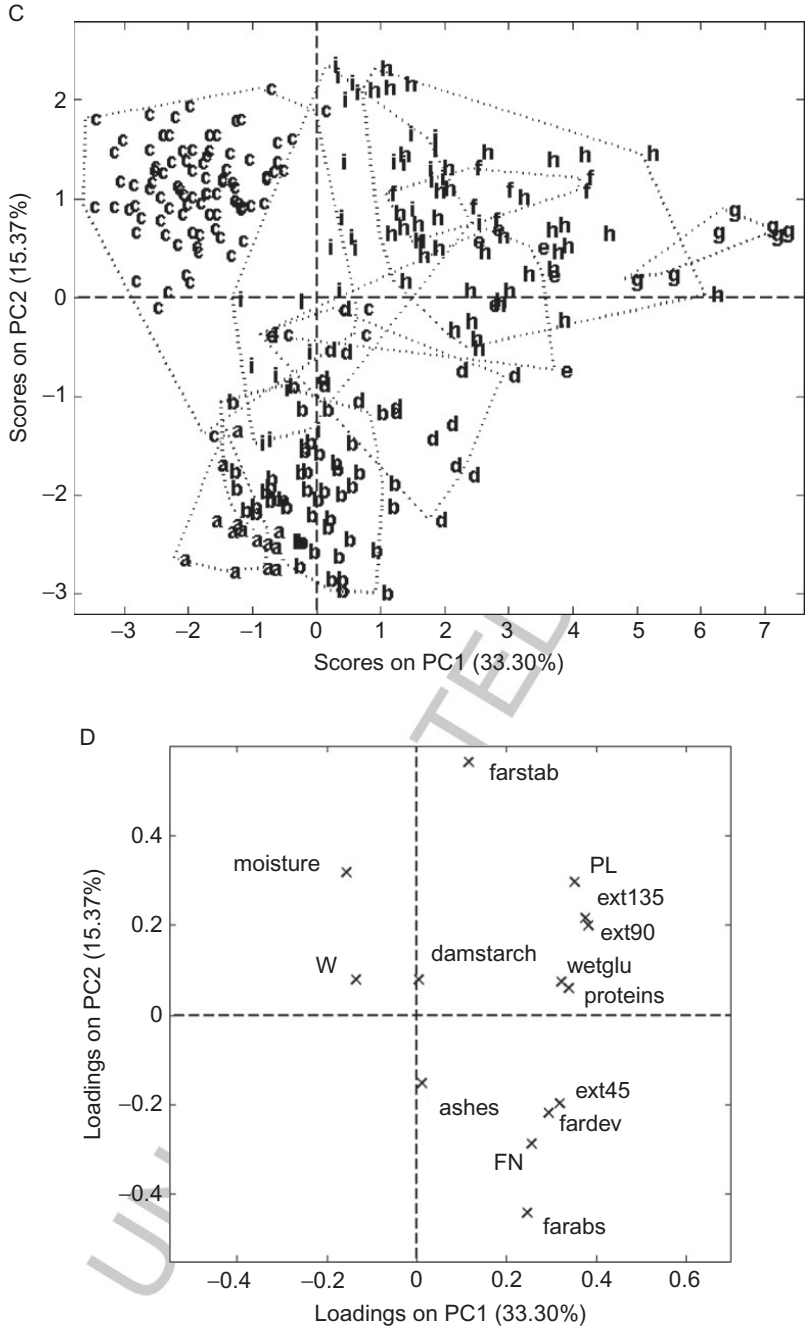


FIGURE 15—Cont'd

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:05 Page Number: 31

B978-0-444-59528-7.00003-X, 00003



f0075 **FIGURE 15** Example of the effect of data scaling on data set FlourRheoData: (A) scores for PC1 versus PC2 and (B) loadings for PC1 versus PC2 for the mean centred data; (C) scores for PC1 versus PC2 and (D) loadings for PC1 versus PC2 for the autoscaled data.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:07 Page Number: 32

B978-0-444-59528-7.00003-X, 00003

spectra, given the high relevance this technique has acquired in food analysis; however, the concepts are generally applicable.

p0510 Signal preprocessing is applied to correct/remove the contribution of undesired phenomena ranging from stochastic measurement noise to various sources of systematic errors: non-linear instrument responses, shift problems and interfering effects of undesired chemical and physical variations. These operations are also known as de-noising, smoothing, background and baseline corrections, normalization (transforming to a scale of relative intensity), alignment (removing horizontal shift), and correction for scatter in NIR. Moreover, transforming the signal, for example, by derivative operations, can implicitly accomplish normalization, baseline removal and partial band deconvolution. As far as removing horizontal shift is concerned, which is a frequent issue arising in chromatography, in NMR signals, as well as in time series data, several algorithms [63–67] which can aid to remove misalignments have been proposed.

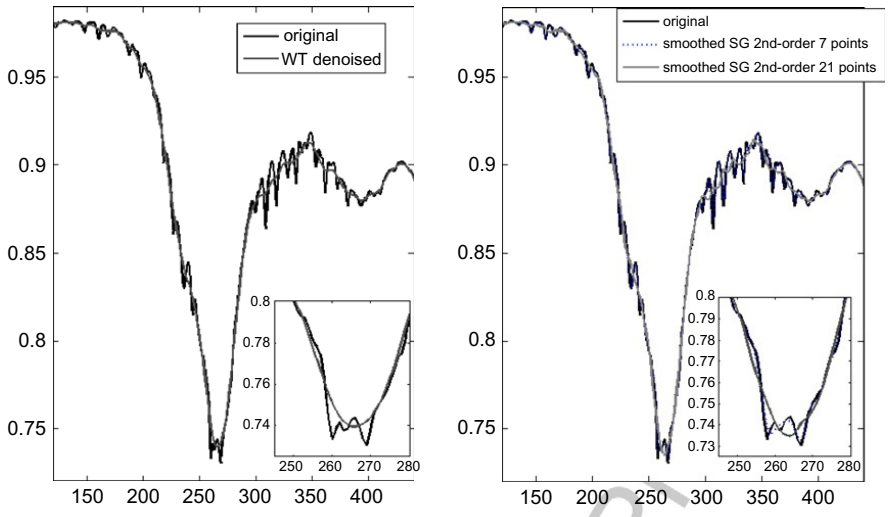
p0515 A general distinction in preprocessing methods may be in terms of filtering methods, which transform measured data mathematically into a presumably 'better' version of the same data, leaving out some undesired types of variation, and model-based methods, where the 'better' version is obtained based on a more explicit mathematical model in such a way that the information filtered out is not lost, as statistical estimates of the mathematical parameters involved in the filtering are also obtained.

p0520 Among the most used filtering methods for de-noising/smoothing, that is, removing uninformative high-frequency variation, there are moving average and polynomial Savitsky–Golay filtering [68], which works on the assumptions that the signal is smooth compared to noise (sum of monotonic functions); noise is mainly uncorrelated and will be eliminated by mild methods. Alternatively high-frequency contributions may be removed in frequency (Fourier transform) or wavelet (wavelet transform) domain. Some examples are given in Figure 16.

p0525 The need to normalize signals, which consists in passing from the measurement scale to a relative one, may arise for different reasons in different contexts. The normalization issue is relevant for signals where peak intensity/area is proportional to concentration and it is not possible to use exactly the same amount of matter for each sample, for example, in high-resolution magic angle spinning-NMR signals where semi-solid samples are used. Other examples are those situations in which the intensity of the signal is affected by physical or chemical variability different from the one we are interested to model, such as water content, temperature, and particle size in NIR, which may be due to the acquisition condition being different from sample to sample. Normalization avoids that these differences in concentration overwhelm the variability due to actual differences in the samples. Signals could be normalized to unit length, unit area, maximum intensity, or according to a reference, intense, well-resolved peak, which is present in all samples and whose

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:09 Page Number: 33

B978-0-444-59528-7.00003-X, 00003



f0080 **FIGURE 16** An example of a noisy band in a medium-IR spectrum of a balsamic vinegar sample. Left plot: black line corresponds to original signal; grey line the same signal de-noised by a db1 wavelet filter at level 3 (all detail coefficients from level 1 to 3 were set to zero, only approximations at level 3 were reconstructed). Right plot: black line corresponds to original signal; blue dashed line to the same signal after smoothing with an SG filter, 7-point window, second-order polynomial; grey line to smoothing with an SG filter, 21-point window, second-order polynomial. (For interpretation of the references to colour in this figure legend, the reader is referred to the online version of this chapter.)

variation is uninteresting (as it will get the unit value for all samples, thus remaining constant over the series of samples considered).

p0530 In cases where a constant offset (vertical shift) is present, it is common to perform row autoscaling, that is, centring and scaling to unit variance along time points and wavelength direction and so on (depending on the kind of signal). For NIR signals, this operation is better known as standard normal variate (SNV). This preprocessing, to some extent, implicitly accomplishes normalization too.

p0535 In the presence of a linear or curvilinear offset, the detrend method can be used. Detrending consists of fitting a polynomial of a given order to the entire signal range. As this algorithm fits the polynomial to all points, baseline and signal, it tends to work only when the largest source of variability in each sample is baseline/background interference, as in NIR signals; this means that it may remove variations which are interesting to model whenever the variation of interest is a reasonably significant portion of the overall variance.

p0540 When the polynomial is fitted to signal points, which are manually selected, that can be attributed only to baseline (background), then baseline (background) removal is achieved and the drawbacks of detrending are overcome. A variant for baseline correction is to do adopt a weighted least squares automatic procedure (asymmetric least squares [63]). This is an automatic approach to determine which points most likely belong to baseline only, by

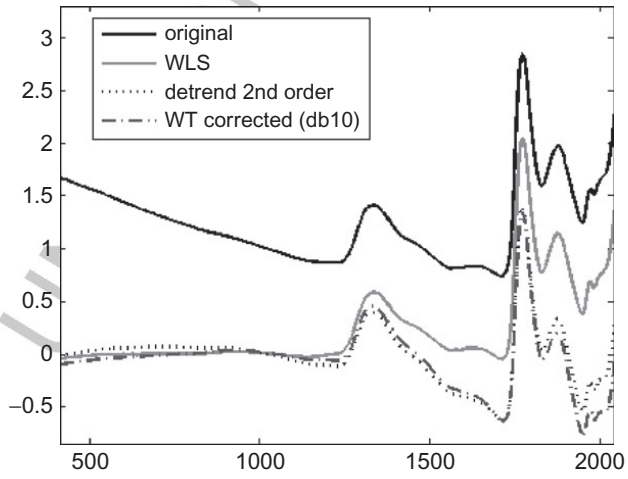
Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:10 Page Number: 34

B978-0-444-59528-7.00003-X, 00003

iteratively fitting a polynomial to each signal and determining which signal points are clearly above the 'fitted baseline' and which fall below. Then the points below are assumed to be more significant in fitting the baseline and get higher weights in the next iteration of fit. The baseline is usually approximated by low-order polynomial, but a specific baseline reference (background profile) can also be supplied.

p0545 A different approach, based on the assumption that the baseline/background is a low-frequency contribution, is to filter out this low frequency, for example, by using the wavelet transform [69] with a wide wavelet filter and removing approximation coefficients at a low scale. In Figure 17 a comparison of the same signals of different baseline correction methods is reported.

p0550 Computation of signal derivatives (also referred to as derivative spectra) allows the removal of constant (first derivative) and linear (second derivatives) offsets from the signal data [70,71]. Derivatives return the slope of a line at any given point: thus the first derivative of a signal shows a maximum where the signal has a maximum slope and crosses zero where the signal has a peak, which renders spectral interpretation more difficult; the second derivative accounts for the rate of slope change in the original signal and has negative peaks corresponding to peaks in the original signal, which allows them to be better interpreted. Drawbacks of derivatives are that scale is decreased and when, as in general, they are computed as differences, noise is increased, so they are often preceded by smoothing. Moreover, derivatives implicitly accomplish band deconvolution and, in general, tend to be far more amenable to use before multivariate analysis of NIR spectra.



f0085 **FIGURE 17** An example of baseline/background removal. In case of WT the approximations at level 9 obtained by a db10 wavelet filter have been removed.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:11 Page Number: 35

B978-0-444-59528-7.00003-X, 00003

p0555 Model-based preprocessing methods [72] assume a mathematical model to account for the different contributions to a signal, which can be expressed in the general form:

$$\mathbf{x} = b\mathbf{x}_{\text{ref}} + t\mathbf{P}^{\text{T}} + \varepsilon \tag{17}$$

where $\mathbf{x}_{\text{ref}}$ is a reference signal which represents at best the sought unknown 'true' signal ($\mathbf{x}_{\text{true}}$) that has to be recovered, and $\mathbf{P}$ contains terms modelling different phenomena: $\mathbf{P} = [\mathbf{P}_{\text{Phys}}, \mathbf{P}_{\text{chem\_irr}}, \mathbf{P}_{\text{chem\_i}}]$, such as physical variability ($\mathbf{P}_{\text{Phys}}$) due, for example, to light scattering (NIR), or thermal conditions; variability due to chemical effects but not relevant to the goals of data analysis ($\mathbf{P}_{\text{chem\_irr}}$), for example, compositional/structural changes due to seasonal variability not important for variety discrimination or water content and so on; and informative chemical variability ($\mathbf{P}_{\text{chem\_i}}$), for example, due to changes in concentrations of the analytes to be determined, in general variability which represents interesting information about samples to be captured/modelled by data analysis. In other words, it is assumed that each signal profile (chromatographic signals, absorbance spectra, etc.) for a set of related samples may be approximated as physical/chemical modifications of a 'true' signal.

p0560 The simplest method which corresponds to these terms is multiplicative scatter correction (MSC). In this case, it is assumed that chemical variation is small compared to physical variation (i.e. variation introducing a constant (additive)/proportional (multiplicative) baseline effect) and thus the true 'signal' may be replaced by a constant reference signal, usually the mean (or median) spectrum, $\mathbf{m}$ (it may also be a specific spectrum of the data set). The previous equation becomes

$$\mathbf{x} = b\mathbf{m} + a + \varepsilon \tag{18}$$

p0565 Solving by least squares, the multiplicative, b, and additive, a, parameters can be estimated and the signal profile corrected to

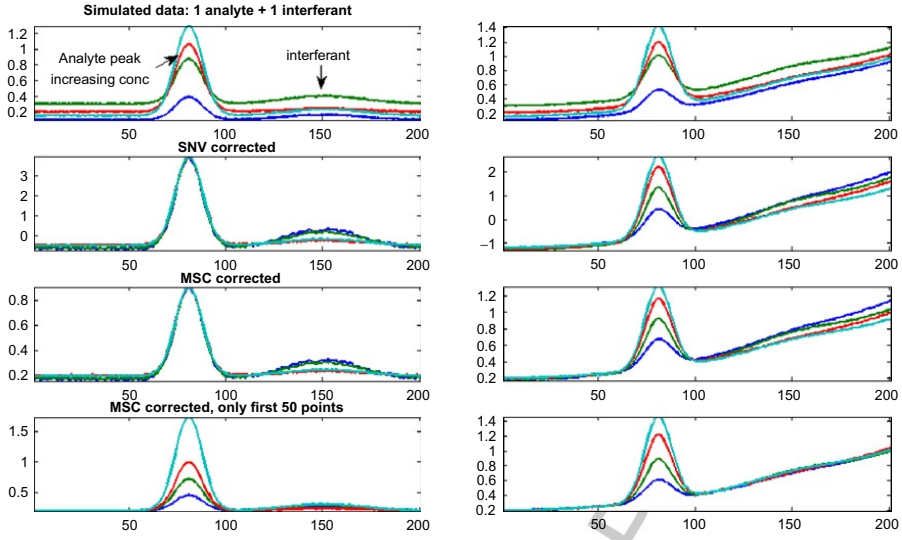
$$\mathbf{x}_{\text{corr}} = (\mathbf{x} - a)/b \tag{19}$$

p0570 This is a very strong simplification because it assumes that chemical variations, for example, analyte and possible interferant concentrations, have little effect on the observed spectra. However, it often works with NIR spectra, where the chemically relevant induced changes mostly manifest as 'shoulders' on a very strong background profile. A situation where MSC (as well as SNV for analogous reasons) does not work [72,73] is shown in Figure 18. A possible solution could be to apply MSC only on a given portion of the signal known to have constant chemical variation, for example, where an interferant at constant concentration is present, and/or belonging only to baseline (absence of chemical variation), as shown in Figure 18.

p0575 It has to be noted that, in the cases shown, derivative spectra would have solved the problem (see Figure 19).

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:14 Page Number: 36

B978-0-444-59528-7.00003-X, 00003



f0090 **FIGURE 18** Simulated data with two peaks: one analyte of interest at increasing concentration and one interferant of varying intensity. Left subplots: vertical shift and white noise present, background absent. Right subplots: vertical shift, noise and background present. (For colour version of this figure, the reader is referred to the online version of this chapter.)

p0580 Going back to the general formulation of model-based preprocessing, the model can be extended (the method in this case is called extended multiplicative scatter correction (EMSC) [74]) to include other physical variations, for example, non-linear terms (not only proportional/multiplicative) with respect to wavelengths:

$$\mathbf{x} = b\mathbf{m} + a + c\boldsymbol{\lambda} + d\boldsymbol{\lambda}^2 + \cdots + \varepsilon \tag{20}$$

where $\boldsymbol{\lambda}$ is the abscissa vector and higher-order terms in $\boldsymbol{\lambda}$ may as well be considered.

p0585 Finally, to cope with situations which differ from the case where the assumption is made of very small effects of 'chemical' variation on spectra, the $\mathbf{P}_{\text{chem\_i}}$ terms may be introduced in the EMSC model. This can be done as an example by considering the signal profile of the pure analyte:

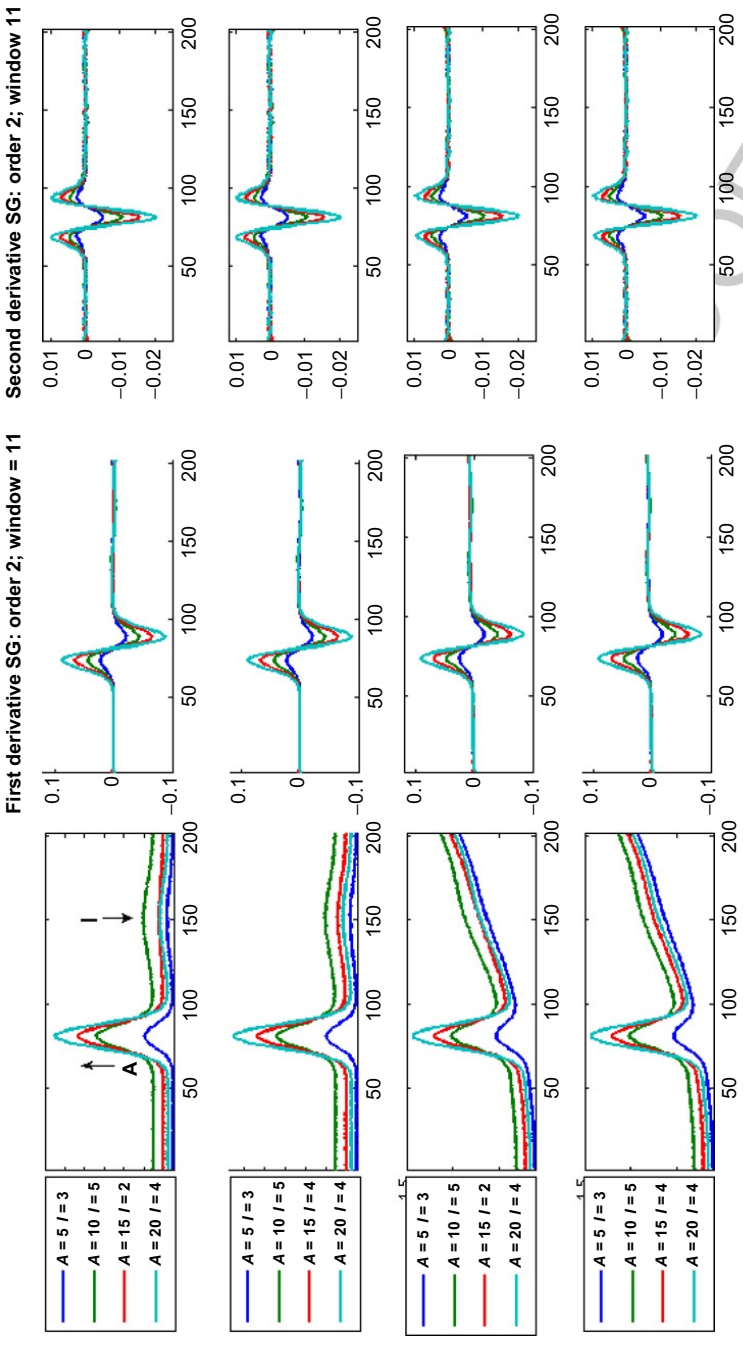
$$\mathbf{x} = b\mathbf{m} + t\left[\mathbf{P}_{\text{phys}}\mathbf{P}_{\text{chem\_i}}\right]^{\mathrm{T}} + \varepsilon = b\mathbf{m} + a + c\boldsymbol{\lambda} + d\boldsymbol{\lambda}^2 + h\mathbf{x}_{\text{pure}} + \varepsilon \tag{21}$$

p0590 In this case the a, b, c and d parameters estimated by linear fitting are not confused with the chemical variation because this is considered in the h $\mathbf{x}_{\text{pure}}$ term; hence we can correct for physical unwanted variation by

$$\mathbf{x}_{\text{corr}} = \left(\mathbf{x} - a - c\boldsymbol{\lambda} - d\boldsymbol{\lambda}^2\right)/b \tag{22}$$

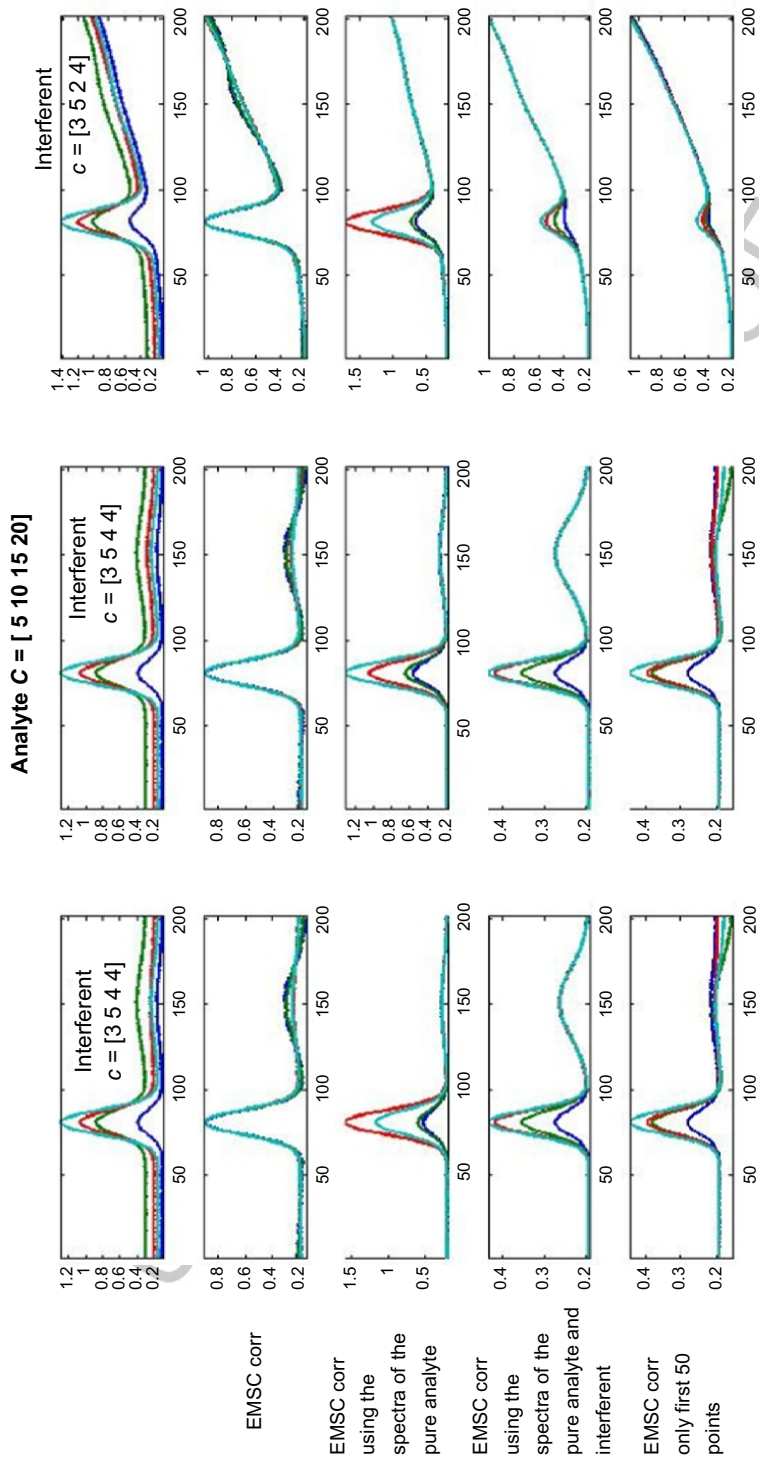
p0595 In Figure 20, an example is shown where this approach has been applied; however, because the interferant concentration is not constant, in order to

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:15 Page Number: 37



f0095 **FIGURE 19** Simulated data with two peaks and one analyte of interest (A) at increasing concentration and one interferant (I) of varying intensity, concentrations are reported on the legend. Left: raw signals. Middle: first derivative spectra. Right: second derivative spectra. (For colour version of this figure, the reader is referred to the online version of this chapter.)

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:18 Page Number: 38



f0100 **FIGURE 20** Simulated data with two peaks and one analyte of interest at increasing concentration and one interferant of varying intensity, concentrations are reported on the legend. The subplots in each column refer to a different series of raw signals, while on the rows the different subplots refer to EMSC applied with different terms, as reported on the left of each series. (For colour version of this figure, the reader is referred to the online version of this chapter.)

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:20 Page Number: 39

B978-0-444-59528-7.00003-X, 00003

have a suitable correction the interferant pure spectrum information has also been considered in the EMSC model.

p0600 When $\mathbf{P}_{chem\_irr}$ is known, for example, by taking a difference spectra of two samples known to belong to different batches, years, producers and so on, we can take this term into account in the model and correct for it; in other terms, EMSC can be seen like a generalized linear model.

p0605 To illustrate the effects of the different preprocessing described until here on explorative PCA models, the data described in Section 3.1.4, namely NIR-BreadProcess, has been used.

p0610 The results are shown in Figure 21, which shows the effect of signal preprocessing on the average spectra (Figure 21A) for each leavening step and on the variance inside each step (Figure 21B), and in Figure 22, where the trends as captured by PCA computed on the different preprocessed data, followed by column mean centring (Figure 22), are shown. For each set, PC2 versus PC3 score plots are reported because these two components describe variability linked to the different leavening step, while for all preprocessing the first component is still describing overall variability not relevant for the leavening step separation. The bottom right sub-plot reports a score plot of the PCA on the concentration of some organic compounds involved in the leavening process determined by GC on the same samples. It is interesting to note that the preprocessing which best matches this trend is the second derivative.

p0615 In the more general case, when the preprocessed signals are not to be used for explorative purposes only, but for modelling tasks (classification, multivariate calibration) as well, the $\mathbf{P}_{chem\_irr}$, can also be estimated by multivariate modelling. For instance, a PCA model can be built with samples measured in those different conditions that we know *a priori* which may introduce unwanted variability (batches, seasonality, acidity of the media, humidity content, etc.); then the loadings of the few PCs where this variability is modelled can be used as $\mathbf{P}_{chem\_irr}$.

p0620 Other model-based preprocessing methods, which are more demanding from the computational and validation point of view, are, for example, orthogonal signal correction and orthogonal partial least squares [75]. When preprocessing it is not sufficient to distinguish relevant information from uninformative sources of variation, it is becoming common to employ variable selection techniques which may reveal to be a very useful tool, especially for multivariate calibration problems [76–78].

s0060 3.1.4 PCA in Practice: Exploring Food Data

p0625 As EMDA is a data-driven method, and according to the main target of this book which aims to showcase the potentialities of an application-oriented discipline like chemometrics in a challenging world such as food analysis, in this section PCA will be shown 'in action' on data sets obtained from real cases of food analysis, both at laboratory and plant scale.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:24 Page Number: 40

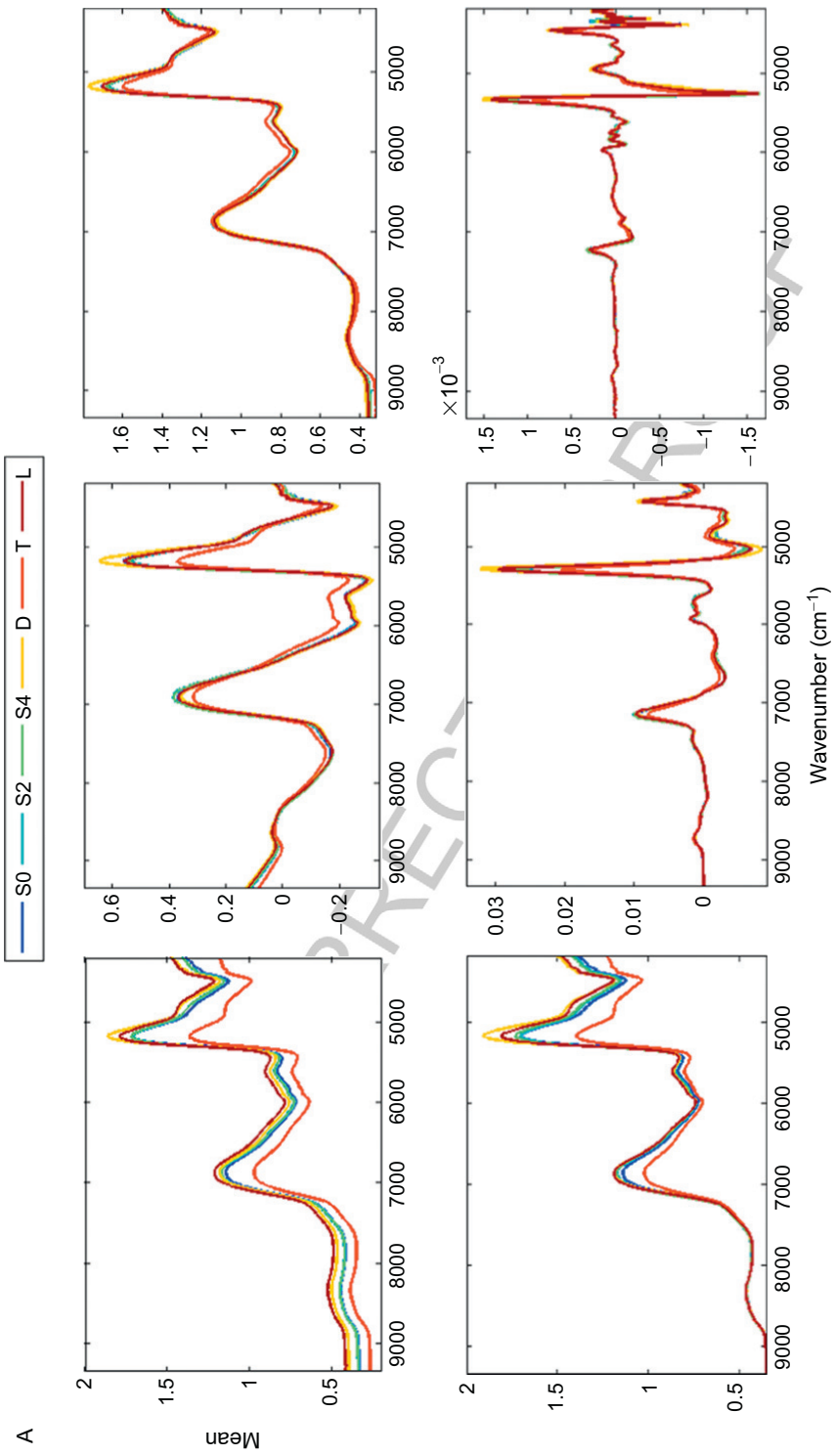
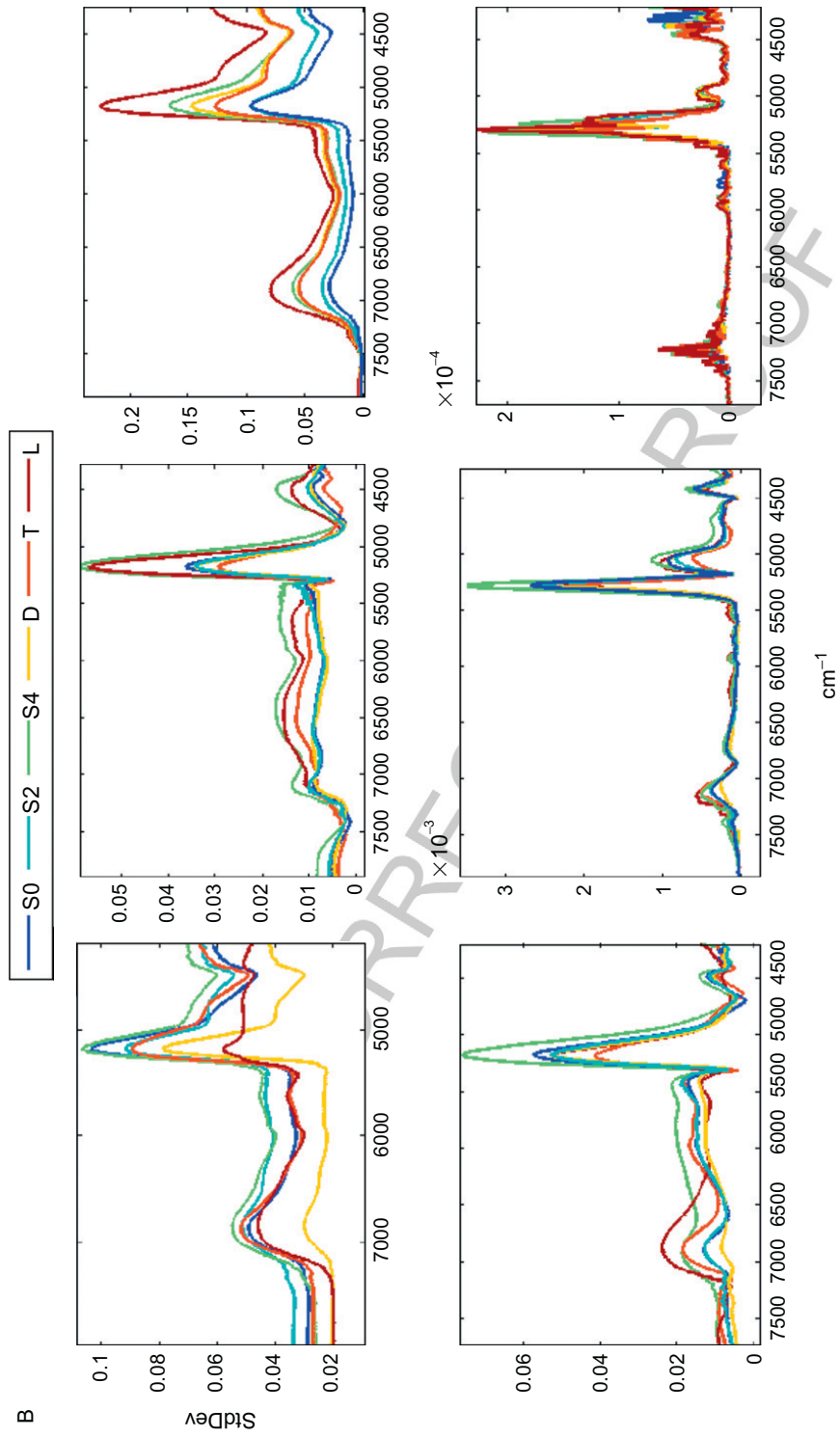


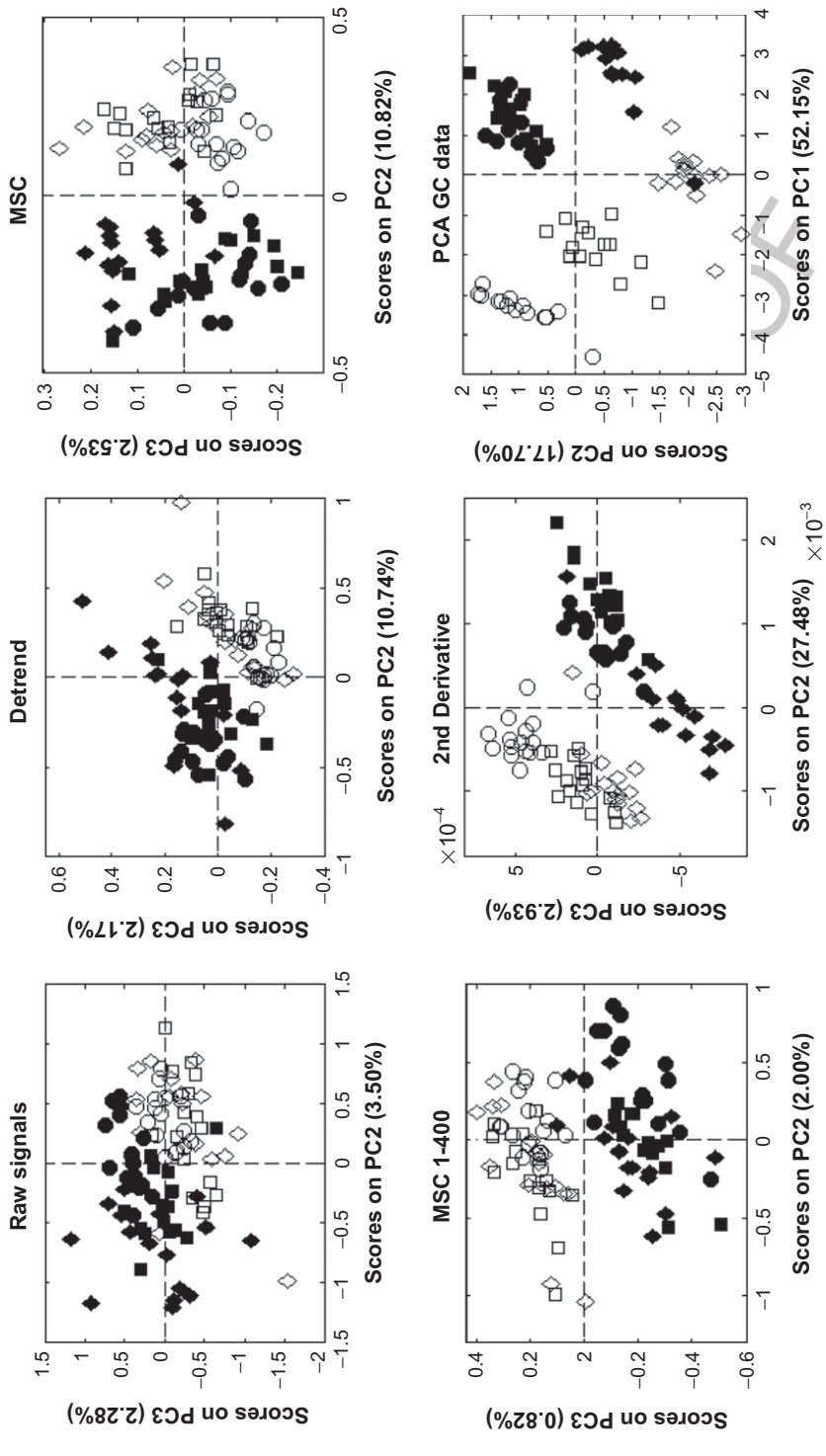
FIGURE 21—Cont'd

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:24 Page Number: 41



f0105 **FIGURE 21** Effect of different spectral pretreatments on the data set NIRbreadProcess on signal profiles (A) and on signal variance (B). The pretreatments applied are in the following order: none (top left), the mean signal for each leavening point (see legend) is reported; detrend (top middle); MSC (top right); MSC only first 400 signal points (bottom left); first derivative (bottom middle) and second derivative (bottom right). (For colour version of this figure, the reader is referred to the online version of this chapter.)

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:29 Page Number: 42



f0110 **FIGURE 22** Effect of different spectral pretreatments on data set NIRbreadProcess on data grouping as revealed by PC2 versus PC3 scatter plots. See text for a description of the meaning of each representation. The six process steps monitored are coded as follows: empty circles, S0; empty squares, S2; empty diamonds, S4; filled circles, D; filled squares, T; filled diamonds, L.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:31 Page Number: 43

B978-0-444-59528-7.00003-X, 00003

p0630 The FlourRheo data set consists of 269 samples of wheat flour employed in an industrial bread-making production, for which 14 variables related to flour rheology have been measured. This data set represents the database over 3 years (2007–2009) of incoming raw material (flour) delivered at an industrial bakery and is well suited as an example of a problem which is shared by most areas of the food industry, namely raw materials variability. Flour, for example, suffers a high degree of seasonal influence on its rheological properties, which also vary on the basis of wheat varieties and formulation: therefore, it is important to act in an explorative way to understand which are the relationships among variables that characterize a flour batch, so that it is possible to foretell which actions are to be taken on the production process in order to maintain the quality of the final product (e.g. bread) inside its target values [14,79].

p0635 The exploratory PCA (data have been autoscaled) clearly shows (Figure 23) the differences in terms of rheological properties of the flour deliveries considered in that period on the basis of wheat variety, mixture type and seasonal effect (symbols legend is reported in Figure 23C). The score plot of PC1 versus PC2 (Figure 23A) shows a differentiation of wheat flour deliveries mainly in terms of the presence of the different pure wheat varieties: samples obtained from a mixing of pure foreign wheat varieties (that is groups a and b, made of a mixture of WFor1 and WFor2) are located at negative values of PC2 and at values of PC1 negative or close to zero, together with mixtures belonging to group d, which is mainly composed of the aforementioned foreign wheat varieties (together with 30% of Italian wheat WIta1), whereas mixtures containing a higher percentage of Italian wheat varieties (WIta1 and WIta2) are located at positive values of PC2. Considering loadings (Figure 23B), it is possible to say that foreign wheat varieties have globally lower values in terms of some rheological parameters connected to flour ability to maintain optimal properties for longer leavening (PL value, ext135 and ext90) and mixing times (farstab), while reaching higher development (fardev) and showing a higher starch degradation (FN), which results in a higher contribution of sugars for yeast activity. In addition, the positioning at negative values of PC1 indicates a higher *W* value (that is a global index of resistance for the gluten network to gas retention in the leavening phase) for the foreign varieties. On the contrary, Italian wheat varieties show an opposite behaviour and appear generally richer in protein content; moreover, WIta1 (group f) appears different from WIta2 (group g), the latter being located at higher PC1 values. As most of the properties have a positive contribution in terms of loadings on PC1, this indicates that WIta2 presents higher values than WIta1, especially in terms of extensibility parameters (ext45, ext90 and ext135) and protein content, while having lower moisture content and W value. The mixture effect results are quite clear considering the disposition of groups of samples which have a different foreign versus Italian wheat variety ratio. Groups d, e and h move from 70:30 to 20:80 and 30:70, and are progressively positioned at more positive values of PC2 and

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:33 Page Number: 44

B978-0-444-59528-7.00003-X, 00003

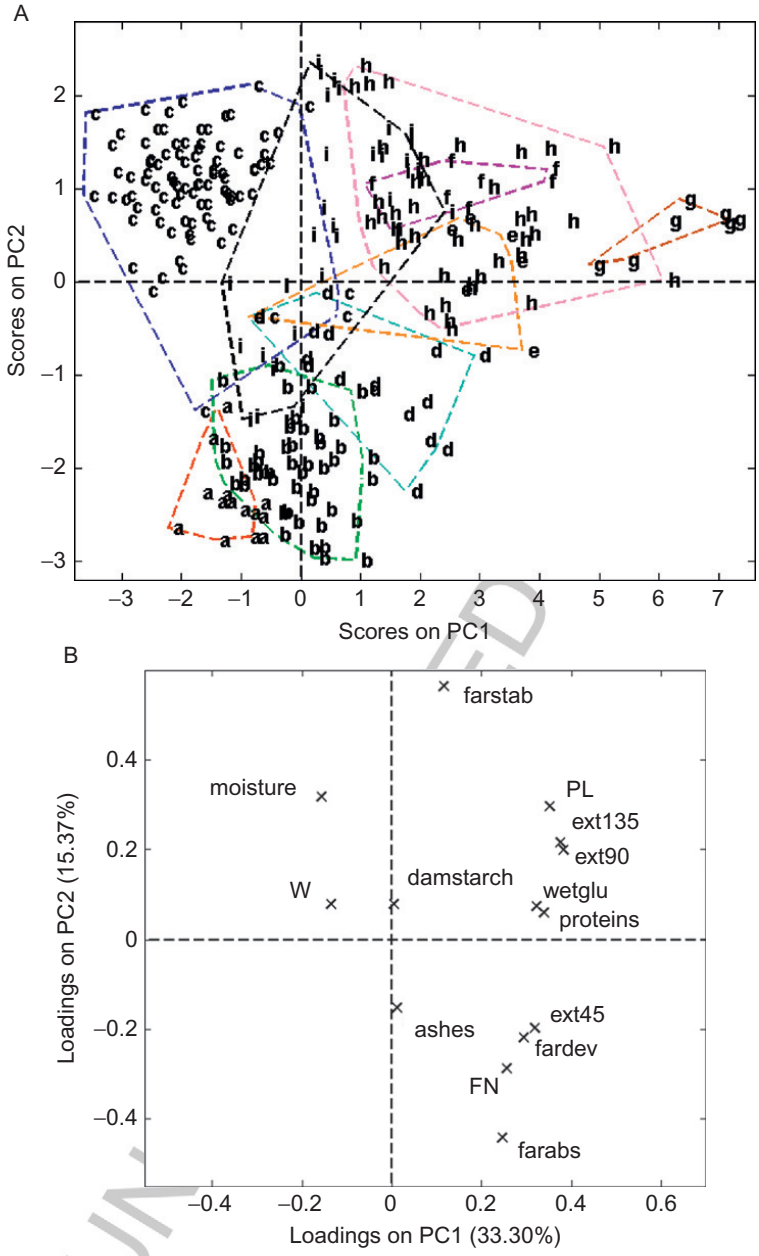


FIGURE 23—Cont'd

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:34 Page Number: 45

B978-0-444-59528-7.00003-X, 00003

C

Symbol	Year	Mixture composition
a	2007	70% WFor1 + 30% WFor2
b	2007	60% WFor1 + 40% WFor2
c	2008	75% WIta1 + 25% WFor1
d	2008	40% WFor2 + 30% WFor1 + 30% WIta1
e	2008	20% WFor1 + 80% WIta1
f	2008	100% WIta1
g	2008	100% WIta2
h	2008	70% WIta1 + 30% WFor2
i	2009	70% WIta1 + 30% WFor2

f0115 **FIGURE 23** Principal component analysis of the data set FlourRheo. (A) Scores plot and (B) loadings plot of PC1 versus PC2. The letters identify the mixtures reported in (C). (For colour version of this figure, the reader is referred to the online version of this chapter.)

PC1, moving from the area of foreign mixtures (quadrant III–IV) to the one of pure Italian varieties (quadrant I). Some peculiar behaviours appear when considering, for instance, group c: this group is characterized by a particular homogeneity in terms of inter-delivery variability when compared to other groups, and is positioned in quadrant II. Its mixture composition, having a foreign versus Italian wheat variety ratio of 25:75, does not differ much from the one of groups h and i, which is 30:70, which show a higher variability. In this case, a wheat variety effect might be present, as the foreign variety differs, but there can also be a significant seasonal effect. Group c belongs to the beginning of year 2008, while h and i were delivered between the end of 2008 and the beginning of 2009; this means that the Italian wheat variety also, which is the same, comes from two different harvests, which could have undergone different weather and growing conditions, thus developing different rheological properties. A seasonal effect may be seen also for groups a and b, which are very close to each other and homogeneous in terms of variability, both belonging to the year 2007. All these exploratory considerations can be the basis for a more detailed evaluation of similarities/differences of new flour deliveries with the ones present in the historical database, and for studies on the functional relationship of flour formulation on wheat flour rheological performance, thus leading to other steps of food analysis.

p0640 The NIRbreadProcess and GCbreadProcess concern 14 doughs from an industrial bread-making production sampled at six different points of the industrial process (S0, S2 and S4: beginning, middle and final point of the first leavening phase; D, T: beginning of the second leavening phase and L:

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:38 Page Number: 46

B978-0-444-59528-7.00003-X, 00003

end of the second leavening phase). In the first data set, the NIR spectrum was recorded in the 4350–9500 cm^{-1} region; in the second data set, the concentration of nine acids, sugars and other compounds related to dough leavening (succinic acid, malic acid, glycerol, fructose, inositol, sucrose, maltose, fumaric acid) was determined by means of GC [80].

p0645 In this context, the explorative analysis phase is crucial to check if the analytical techniques are able to collect information on the process. Figure 24A shows, as a biplot, that GC (data have been autoscaled prior to analysis) is able to capture information on sample composition which is, first of all, able to clearly differentiate the first leavening phase (negative values on PC1) from the second leavening phase (positive values on PC1), and that is indicative of the progression of dough leavening, as the trends from positive to negative values of PC2 suggest. This is connected to a general change, in terms of variables, from higher concentrations of sugars at the beginning of each leavening phase towards a higher concentration of some leavening products and the consumption of sugars at the final leavening points. As another technique was considered, that is NIR spectroscopy, it is also possible to compare the results obtained with both of them. As reported in Figure 24B, the PCA on the NIRbreadProcess data set (preprocessing: Savitsky–Golay second derivative, with second-order polynomial smoothing, and mean centring) indicates that the NIR signal can bear information, albeit in different components (in this case, PC2 and PC3), similar to the GC quantification of chemicals present in the leavening phase. Loadings (Figure 24C) illustrate which infrared vibrations are mainly involved in the samples distribution in the PC's space. The samples located at positive values of PC2, corresponding to most of the samples of the second leavening phase, are characterized by higher values for the spectral variables around 5200 cm^{-1}, several vibrational modes are absorbed in this region, such as the asymmetric NH stretching of the CONH2 group, the second overtone of the C═O stretching (in CO2R), the first stretching and the second bending overtones of water OH and the combination bands of starch OH. On the other hand, the PC3 loadings are mostly inherent to the NIR region below 4500 cm^{-1}, where contributing combination modes are found, for example, of CH and CC stretching and of CH and CH2 bending, which can be found in different constituents of flour, such as starch, lipids and proteins.

p0650 The rather broad nature of NIR signal can be accounted for by the fact that PC1, the main source of variability, is due to other effects connected to the production process which are not discussed here.

p0655 The exploratory analysis of these two data sets offers at least two different points for further analyses:

o0050 **1.** The NIRbreadProcess data set is shown as being a good representative of a process-monitoring approach in the food industry, where online analysers are employed on the production line.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:40 Page Number: 47

B978-0-444-59528-7.00003-X, 00003

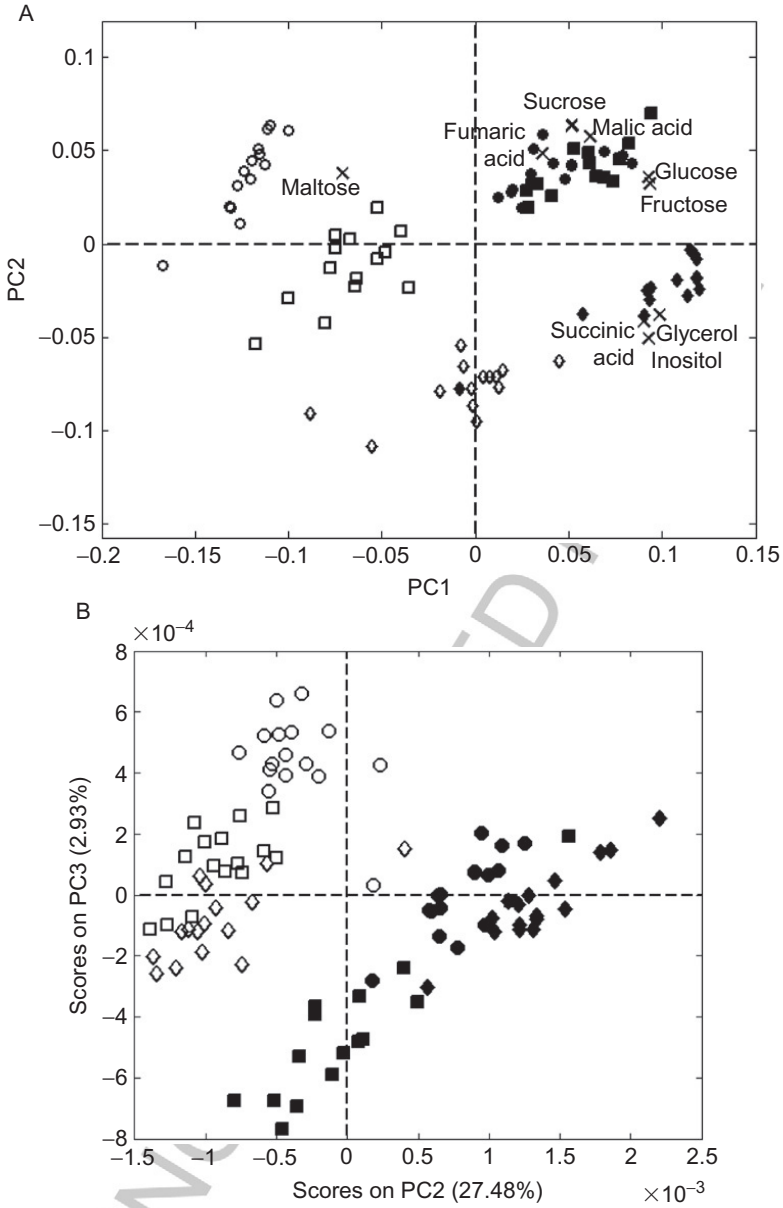
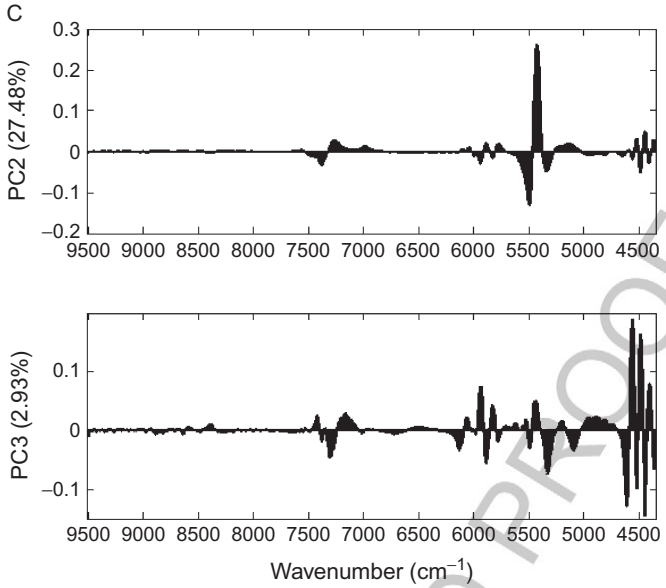


FIGURE 24—Cont'd

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:40 Page Number: 48

B978-0-444-59528-7.00003-X, 00003



f0120 **FIGURE 24** Principal component analysis of data sets. (A) GCbreadProcess: biplot of PC1 versus PC2, (B) NIRbreadProcess: scores plot of PC2 versus PC3 and (C) NIRbreadProcess: loadings plot of PC2 and PC3. The six process steps monitored are coded as follows: empty circles, S0; empty squares, S2; empty diamonds, S4; filled circles, D; filled squares, T; filled diamonds, L.

o0055 **2.** The fact that the two techniques offer similar views of the same process, that is the leavening progression, is a start for a feasibility study of indirect calibration models of properties of interest, that is the concentration of leavening products in dough from the NIR spectrum, thus obtaining important information on the process in a matter of seconds (the acquisition of the NIR signal) instead of more than 1 h (the GC run).

p0670 Finally, an example should be provided about a class of methods, which have also explorative purposes, which will be discussed with more detail and theoretical depth in Chapter XXX, that is multiway analysis methods [81]. These Au1 methods, among which parallel factor analysis (PARAFAC) will be shown here in action compared to PCA, are to some extent referred to as the conceptual (and mathematical) extension of PCA to arrays of order higher than two. They show their potentiality when the variability of a data set is related to different sources, or conditions, at which a full set of properties for each sample is measured [17–21]. An example, quite common in the food science analysis, is the excitation–emission fluorescence landscape, where, for each sample, an emission spectrum is recorded at each wavelength of the excitation signal.

p0675 The NIRdoughRising data set contains the Near Infrared Spectrum, in the 1376–2245 nm region, recorded at seven leavening times (at the beginning

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:44 Page Number: 49

B978-0-444-59528-7.00003-X, 00003

and every 10 min up to a 60-min rising phase) for doughs obtained from ten wheat flour mixtures characterized by a different formulation of four wheat varieties. As it is important to gain knowledge on the effect of wheat flour formulation on its rheological properties and on its performance in the leavening phase, this kind of experiment is useful in order to plan the best mixture to be used in an industrial production [82,83].

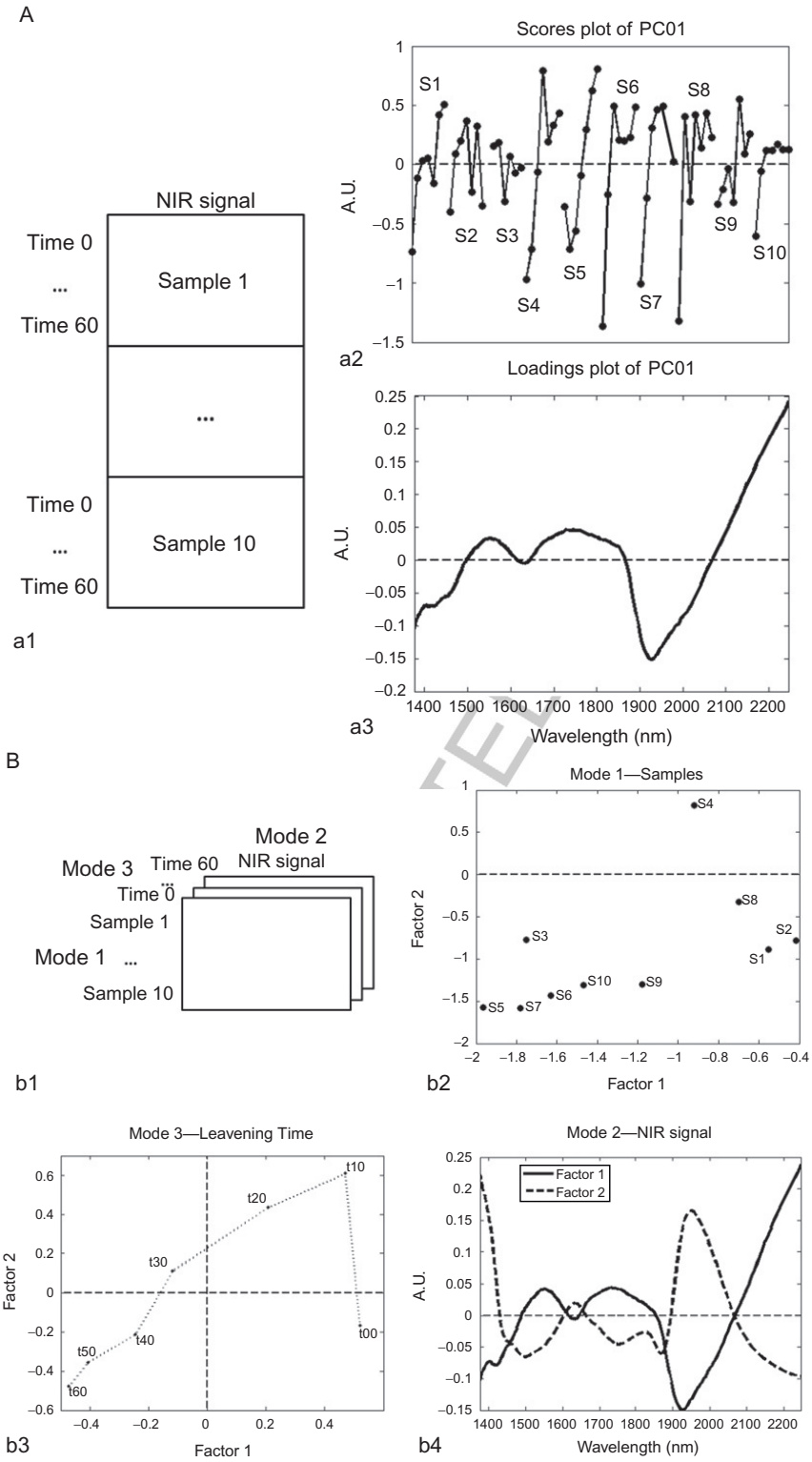
p0680 To analyze such a data set in an explorative way, one can chose from two approaches: to consider a two-way data set arranged as in Figure 25.a1, where time progression for each sample is 'unfolded' in different rows and NIR signal (preprocessed with smoothing and SNV) represents the source of variability; or in Figure 25.b1, where a three-way array is composed, having samples in Mode 1, NIR spectra in Mode 2 and leavening times in Mode 3. The 'unfolded' data matrix has been pretreated by centring separately each sample mixture with respect to its mean leavening time (see Section 3.1.3.1 for some remarks on this procedure), which corresponds to mean centring across Mode 3, the three-way array. PCA shows on PC1 scores (Figure 25.a2) that most flour samples have a similar trend from negative to positive values of PC1 (points are ordered according to leavening time), and that two main groups can be present, one which has higher variability (e.g. S4–S8) and another one with less variability with time (e.g. S2, S3 and S9). It is important to note that PCA confounds the variability between and within flour mixtures, considering them altogether. The advantage of using a PARAFAC model lies in the fact that these two sources of variability are explored separately and belong to Mode 1 and Mode 3, respectively. This allows obtaining clearer information on similarities and differences among samples, from Mode 1 scores plot (Figure 25.b2), where now two clusters appear and sample S4 is highlighted as peculiar, and from Mode 3 loadings plot (Figure 25.b3), which shows that a trend of evolution of the NIR spectrum with time exists. Mode 2 loadings (Figure 25.b4) and PCA loadings (Figure 25.a3) are quite similar, thus showing that the same information is captured by both approaches, but the visualization is particularly different.

s0065 3.1.5 Using PCs and PCA Models Beyond Data Exploration

p0685 In this chapter PCA is discussed as an EMDA tool; however, as discussed in several other chapters of this book PCA can be part of a modelling task, for example, in class modelling (SIMCA) or multivariate calibration (~~PCR~~) or Au2 in building reference normal operating condition models in multivariate statistical process control. PCA derivation does not differ in these contexts until assessing the right model dimensionality, that is, the number of significant PCs (chemical rank), became critical. Thus the choice of the number of PCs to be retained has to rely on validation, by using methods such as cross-validation, permutation tests and bootstrapping [58,84,85].

p0690 In these contexts, it is also assumed that with PCA a set of similar samples representative of a given category is modelled; hence, we also aim at

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:46 Page Number: 50



f0125 **FIGURE 25** Analysis of the data set NIRdoughRising. (A) Unfolded matrix (for PCA): (a1) data set scheme; (a2) scores of PC1; (a3) loadings of PC1. (B) Three-way array (for PARAFAC): (b1) data set scheme; (b2) loadings of F1 versus F2 for Mode 1; (b3) loadings of F1 versus F2 for Mode 3; (b4) loadings of F1 and F2 for Mode 2.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:46 Page Number: 51

B978-0-444-59528-7.00003-X, 00003

estimating 'population' parameters in PCA space, such as statistics for T^2 and Q and their respective confidence intervals. A discussion of inference in PCA may be found in Refs. [26,53,86].

s0070 3.1.6 Other Derivations of PCA

p0695 We have described PCA as a method to project/compress measurements collected in a J-dimensional space to an A-dimensional hyperplane ($A<J$) with the useful property of retaining salient samples/variables patterns unaltered in the reduced space.

p0700 However, only when the measurement errors are independent and follow a normal distribution (homoschedastic noise), PCA estimation of this subspace is optimal in a maximum likelihood sense [87]. On the contrary, if measurement error variances are non-uniform (heteroschedastic noise) and also not independent, the PCA projection may represent the samples incorrectly or non-optimally. Non-uniform measurement errors in a data set may arise from error sources inherent to a given type of instrumentation or experimental setting, for example, variations in noise across measurement channels in spectrophotometers, or due to the presence of missing information.

p0705 An effective visualization of clusters within the scores plots may be hindered by objects dominated by noisy measurements that are projected in such a way as to obscure the separation of object classes. To overcome this issue, a different derivation of PCA has been proposed, [87–90] namely maximum likelihood principal components analysis (MLPCA). Basically, it is a generalization of PCA to non-ideal error structures which, given reasonably accurate characterization of measurement noise, can improve the projections of the points into scores space. It does this by performing a separation of noise variance from other sources of variance in the data and by using maximum likelihood projection instead of orthogonal projection as standard PCA. The projections for the noisy measurements in PCA and MLPCA differ because PCA ignores noise direction and all projections are orthogonal, while MLPCA best projection geometrically corresponds to the point of nearest intersection of the measurement error ellipsoid with the subspace. MLPCA projection for a sample i is formulated as Au3

$$\mathbf{t}_i\mathbf{V}^{\mathrm{T}} = \mathbf{x}_i\sum\nolimits_i{}^{-1}\mathbf{V}_A\left(\mathbf{V}_A{}^{\mathrm{T}}\sum\nolimits_i{}^{-1}\mathbf{V}_A\right)^{-1}\mathbf{V}^{\mathrm{T}}{}_A \tag{23}$$

where $\mathbf{V}$ is the loading, $\mathbf{t}_i$ the sample score, A the number of components and $\mathbf{\Sigma}_i$ the error covariance matrix that is needed to project a sample. As in PCA, scores and loadings are orthonormal, and SVD decomposition method may be used, but the solution is not nested, that is, a model with two components cannot be obtained taking the first two PCs from a three-component model. Moreover, the minimization depends not only on the measurements but also on its associated error structure. The most common approaches to estimate

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:49 Page Number: 52

B978-0-444-59528-7.00003-X, 00003

measurement error covariance matrices generally fall into one of the following: experimental replication, theoretical modelling, or empirical modelling. By experimental replication a submatrix of replicates is obtained from which the error covariance can be estimated.

p0710 In Section 3.1.2 the capability of PCA to help identify multivariate outliers has been described as an advantage in the EMDA context; the other side of the medal is that PCA is sensitive to outliers: indeed, PCA direction is most influenced by outliers. In fact, outliers artificially increase the variance in an otherwise uninformative direction, and will drive the first PC in that direction. This constitutes a problem if we want to use the PCA model as a reduced informative view of our data, as we will end up with a distorted or non-optimal view of our samples relations (in fact, not only will the first PCs direction point to outliers but given the orthogonality constraint, the direction of subsequent PCs will also be influenced, i.e. distorted by the maximum variance direction that will represent the samples in the absence of outliers).

p0715 A solution in this case is to use robust models and robust statistics (such as median as measure of location and median of absolute deviation around the median as measure of spread) in estimating PCA parameters. The aim is to construct models and estimates clearly describing the majority of the data. Moreover, construction of robust models allows a proper identification of outlying observations. A review illustrating the basis of robust techniques in data analysis and chemometrics can be found in reference [91].

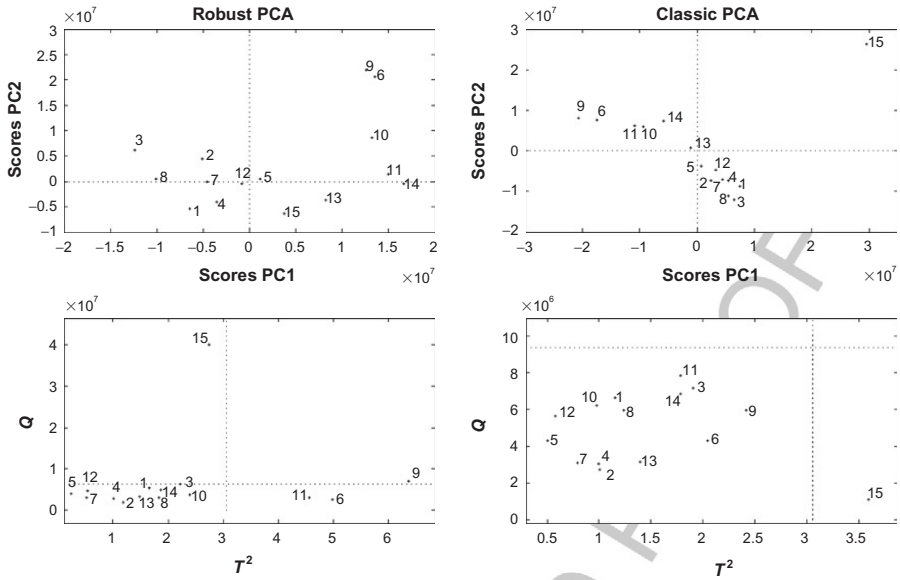
p0720 There are essentially two approaches for robust PCA: the first is based on PCA on a robust covariance matrix, which is rather straightforward as the PCs are the eigenvectors of the covariance matrix. Different robust estimators of covariance matrix may be adopted (MVT [92], MVE and MCD [93]) but the decomposition algorithm is the same. The second approach is based on projection pursuit (PP), by using a projection aimed at maximizing a robust measure of scale, that is, in a PP algorithm, the direction with maximum robust variance of the projected data is pursued; here different search algorithms are proposed.

p0725 A better visualization of data structure is demonstrated in robust PCA as shown in Figure 26.

p0730 The scores plots refer to PCA (robust PCA on the left side of the figure and classic PCA on the right side) of a set of extra virgin olive oil of Greek provenience whose volatile fraction has been characterized by headspace GC-MS. In both cases, data have been mean-centred before PCA. The outlying sample no. 15 strongly influences both PC1 and PC2 directions (Figure 26, top right), resulting in the fact that these PCs almost only describe the variability due to this sample. In the Q/T^2 plot, Figure 26, bottom right, sample no. 15 is seen as extreme in PC space (outside the Hotelling-T^2 limit) but not distant from the model (inside the Q-statistic limit), because the PC model is driven to describe this sample. By using a robust approach sample 15 is recognized as an outlier, outside the Q-statistic limit (Figure 26, bottom left), but PC1

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:50 Page Number: 53

B978-0-444-59528-7.00003-X, 00003



f0130 **FIGURE 26** Comparison of classical and robust PCA results. Scores and T^2/Q plots for a data set of extravirgin olive oil from Greece measured by headspace GC–MS. Left top: PC1 versus PC2 scores plot by Robust PCA; right top: PC1 versus PC2 scores plot by classic PCA; left bottom: T^2/Q plot for Robust PCA, dashed lines correspond to confidence limits at 95% for Hotelling-T^2 and Q statistics, respectively; right bottom: T^2/Q plot for classic PCA, dashed lines correspond to confidence limits at 95% for Hotelling-T^2 and Q statistics, respectively.

and PC2 directions are not influenced by this sample as seen in the corresponding robust PCA scores plot (Figure 26, top left).

p0735 Robust PCA is not so often used in EMDA; however, there are several freely available packages to do this, such as *robustbase* [94] and *rrcov* [95] in the R environment, TOMCAT from Walczaks' group [96] and LIBRA toolbox in MATLAB [97].

s0075 3.2 Other Projection Techniques

p0740 As one of the most relevant features of EMDA is its possibility of visualizing data sets in a graphical form which is easy to understand by the human brain for further reasoning, many methods that aim at finding low-dimensional representations of high-dimensional data sets have found applications in this context, PCA being perhaps the most commonly applied in many scientific fields, food analysis *in primis*. As described earlier, PCA's target is in finding a projection of data to a lower-dimensional space, which *maximizes the explained variance*. However, other methods can be easily introduced by considering a different criterion instead of the one of maximum variance. In the following sections, the most known will be briefly presented, together with examples and comparisons of the visualization results on some data sets.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:50 Page Number: 54

B978-0-444-59528-7.00003-X, 00003

s0080 ### 3.2.1 Projection Pursuit and Independent Component Analysis

p0745 The concept behind the PP approach was developed by Friedman and Tukey in 1974 and it is based on the identification of the most 'interesting' projection of data according to the highest degree of deviance from the normal distribution [98,99]. Independent component analysis (ICA) [100–102] is strictly connected to PP, as it considers as a criterion the maximum statistical independence of the estimated components. This can be obtained through a wide (and still debated) number of possible definitions of independence, among which the maximization of non-Gaussianity is probably the most commonly employed. The non-Gaussianity-based algorithms, such as fastICA [102,103], are based on the central limit theorem and operate by maximizing a quantity called 'negentropy' $\Delta S(x)$, that is the difference of the entropy of a normally distributed random variable $S(x_G)$ and the entropy of the variable under consideration $S(x)$, which can be defined as

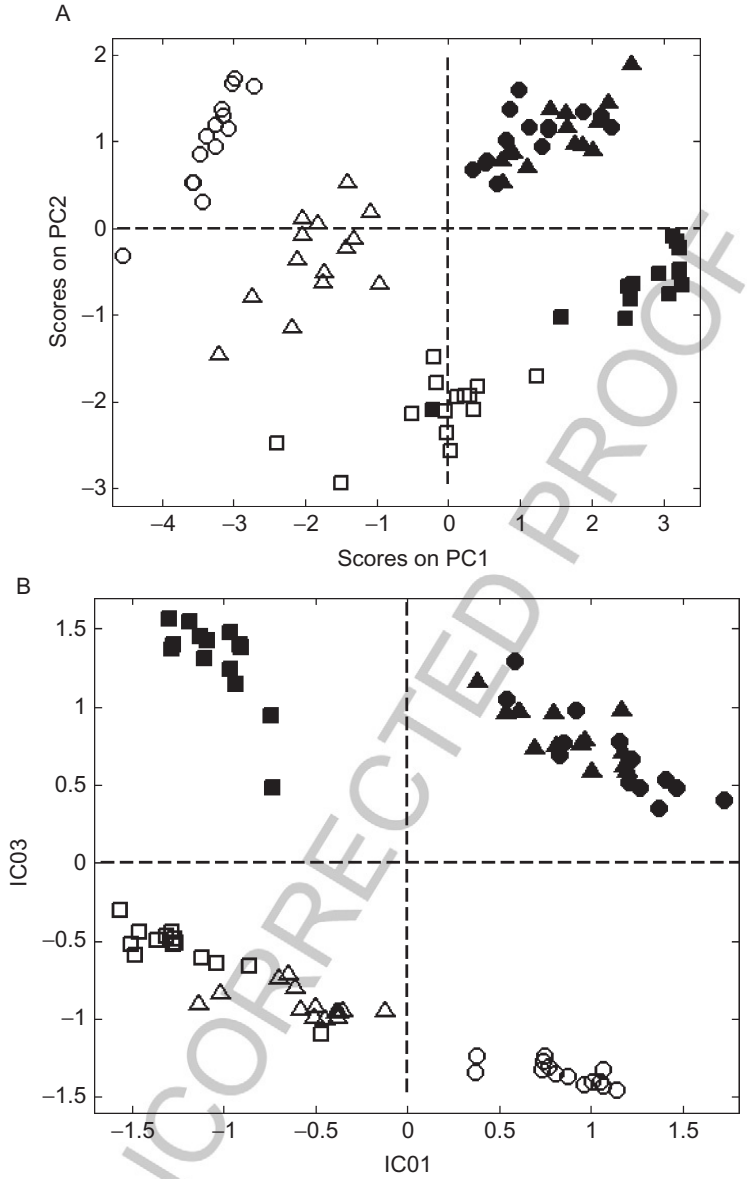
$$S(x) = -\int f(x) \log f(x) \mathrm{d}x \tag{24}$$

p0750 The most important aspect of this approach is that ICs are not as affected by high variability normally distributed variables as PCs are, thus making it possible to distinguish directions of differentiation for data where relevant information is hidden in small, non-normal differences. This means that IC scatter plots can highlight sample clustering or trends different than PCA, and that PCA visualization of object differences can be still present in a IC scatter plot, although in different components. The ICA method, being strictly related to the blind source separation methodology [104,105], is widely used in the signals and communication fields, working on the assumption that the data matrix can be resolved into a set of 'base' mutually statistically independent source signals and a 'mixing' matrix which contains the proportions by which the base signals are overlapped to obtain each original signal. The method finds some (albeit not very common) applications in food science [106–108], and presents some differences and potential advantages with respect to PCA, especially when related to the deconvolution of instrumental signals (finding independent sources, rather than orthogonal loadings as in PCA, reduces the number of mathematical artefacts on the signal) [105]. One significant issue is to find out which factors are significant and which are not, thus the individuation of validation procedures, which can be complicated by the fact that two consecutive models may differ in the order and signs of similarly indexed ICs [109,110].

p0755 Figure 27 shows the scatter plots of PC1 versus PC2 for a PCA model (Figure 27A) of the GCbreadProcess data set (preprocess: autoscale) and of IC1 versus IC3 for a PP–ICA model (preprocess: autoscale) (Figure 27B), and a comparison of both. The information, that is the differentiation of the six points of the process in five clusters (points D and T being overlapped),

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:53 Page Number: 55

B978-0-444-59528-7.00003-X, 00003



f0135 **FIGURE 27** Comparison of (A) principal component analysis and (B) projection pursuit–independent component analysis for the GCbreadProcess data set. The six bread-making phases are coded thus: empty symbols: circle S0, triangle S2, square S4; full symbols: circles D, triangles T, squares L.

is perhaps clearer in PP–ICA, where one component, IC1, indicates the difference between the beginning (positive values, points S0, D and T) and the end of leavening (negative values, points S2, S4 and L), while the other, IC3, distinguishes the first leavening (negative values, points S0, S2 and S4) from the second one (positive values, points D, T and L).

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:53 Page Number: 56

B978-0-444-59528-7.00003-X, 00003

s0085 ### 3.2.2 Multidimensional Scaling and Principal Coordinate Analysis

p0760 The methods discussed so far are based on the projection of a full set of data to a new space of lower dimensionality, that is samples which have been characterized by a given number of variables. In some circumstances, however, only item–item similarities/dissimilarities matrixes are available or preferable, thus the explorative evaluation of data must be conducted with a different approach. Multidimensional scaling (MDS) methods [111–114] work on item–item similarity matrixes by assigning to each of the items a location in an N-dimensional space, usually with N small enough so that 2D and 3D visualization of data disposition is possible. The goal is to reconstruct a low-dimensional map of samples that leads to the best approximation of the same similarity matrix as the original data. Depending on the kind of input matrix and the criterion used to define which approximation is the best, different MDS algorithms and approaches are possible.

p0765 Metrical MDS operates on an input matrix of dissimilarities, or distances, between pairs of samples, giving as a result a matrix of coordinates whose configuration minimizes a loss function. This method presents an optimization phase, which can be performed with a variety of loss functions to be considered, and other possible variations of the methods concern the input distance matrix, which can be calculated according to different weights and criteria. When the Euclidean distance is considered, the classical MDS, also known as the principal coordinates analysis, consists in performing a PCA on the double-centred distance matrix and then rotating the solution so that the stress criterion S is minimized:

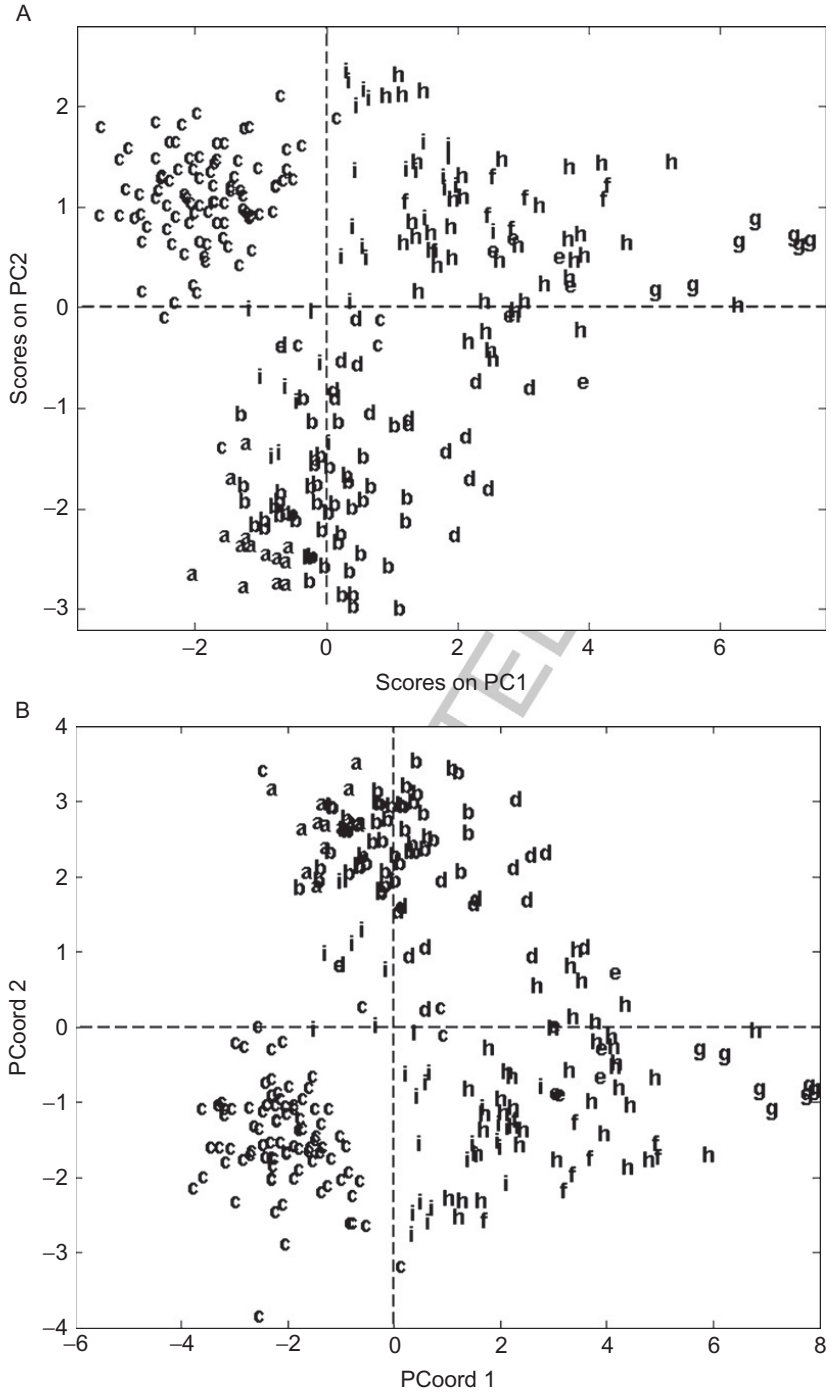
$$S = \sum_{k<i} (d_{ik} - e_{ik})^2 \tag{25}$$

where e_{ik} corresponds to the input distances and d_{ik} are the distances between objects x_i and x_k in the projection space. The results of the principal coordinates analysis of the Euclidean distances of samples are substantially analogous to those of PCA of the original data set. The comparison is shown in Figure 28, where the PCA of FlourRheo data set (see Section 3.1.4 for more information on the data) has very similar results to the principal coordinates analysis of the distance matrix for the same samples according to the Euclidean distance (Figure 28B).

p0770 When a non-parametric, monotonic relationship is searched between the distance matrix and the distance between objects in the projection space, non-metric MDS is introduced. Usually the approaches differ in the stress criterion which is chosen to be minimized, and no analytical solution is available, so that other methods must be considered, such as an iterative, gradient descent optimization in Sammon's mapping, repeating the mapping several times starting from different sets and with different parameters to avoid local

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:58 Page Number: 57

B978-0-444-59528-7.00003-X, 00003



f0140 **FIGURE 28** Comparison of (A) principal component analysis and (B) classical multidimensional scaling–principal coordinate analysis for the FlourRheo data set. See Figure 23 for the legend.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:01:58 Page Number: 58

B978-0-444-59528-7.00003-X, 00003

optima, or an alternation of search for a good configuration in low-dimensional space, and an appropriate non-monotonic transformation to the original space, as in Kruskal–Wallis mapping.

p0775 The use of MDS in food science is not particularly common, and it is especially used in biological studies, psychic–economic studies on food perception and sales, and sensory analysis [115–118]. In all those situations which belong more to analytical food science it is very rare to see this method applied, perhaps because of the fact that, in such a field, it is quite hard to reason in terms of distances among samples rather than variables, the interest being more focused both on the visualization of sample similarities and dissimilarities, and on the interpretation of which variables are responsible. When the full distance matrix (or the similarity/dissimilarity one) is used, the variable influence is lost, which causes a drawback both in terms of interpretation and in terms of lack of a mapping operator to use on new objects in order to project them into the lower-dimensional point configuration.

s0090 3.2.3 A Non-Linear Approach: Self-Organizing Kohonen's Maps

p0780 Most of the methods discussed so far are based on a projection to latent variables which is obtained as a linear combination of the original ones according to a given criterion, or on the description of the relationships between samples by means of distances which work in a metric space. However, it is possible to come across real cases in which the dissimilarities among samples are caused by clusters characterized by different densities and distributions, and/or non-linear patterns, that is situations in which a linear, Gaussian-based model shows its limits. In these cases it is possible to consider non-linear approaches, such as the aforementioned modifications of MDS, or artificial neural network algorithms. The latter family will be discussed in more detail elsewhere in this book, together with a more detailed coverage of one of these methods, self-organizing maps (SOMs), which presents many analogies in terms of data reduction and visualization when compared to the other techniques such as PCA and fits well in an EMDA context.

p0785 SOMs are a type of artificial neural network first introduced by Teuvo Kohonen, thus often referred to as Kohonen maps [119–121]. Without entering too much into detail, as another chapter of this book (Chapter XXX) is Au1 dedicated to the applications and potentialities of non-linear methods, SOMs are based on an unsupervised learning method to train the network in order to obtain a low-dimensional (typically 2D) representation of the input space of the samples. After the training phase, where the map is built using input 'examples', the network can be used in a mapping perspective, automatically classifying new input samples. An SOM is made up of a number of components, called nodes (or neurons), which are each associated to a weight vector of the same dimensions of the data variables, and disposed in the map space according to a geometry which characterizes the network, the most common

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:00 Page Number: 59

B978-0-444-59528-7.00003-X, 00003

arrangements for a 2D map being a hexagonal or rectangular grid. The mapping from a higher dimensional input to a lower dimensional network is done by finding the node with the smallest distance (according to the chosen metric) between the weight vector and the sample vector; in other words, each node represents an 'archetype' sample, to which all the training samples are compared. The most similar node (the 'winning unit') is then rotated towards the object by replacing the weight vector it carries with an average of the old values and the ones of the sample, weighted by a learning rate α, which decreases as the training iterations progress. In addition, all the other nodes in the neighbourhood of the winning unit are updated as long as that node changes, which results in the fact that, at the end of training, the neighbouring units are generally more similar than units far away.

p0790 Figure 29 shows the comparison of the results of PCA and SOM on the FlourRheo data set. As far as the SOM is concerned, the map was obtained with the R package *som* [122]. The network was an 8×8 rectangular grid, and the autoscaled data were treated for 500 iterations, under an initial learning rate $\alpha = 0.1$. As one may note, PCA scores (Figure 29A) show that clusters exist, but are characterized by different densities (i.e. internal group variability) rather than actual separation on the basis of one or more linear combination of flour properties (which concur in composing the PCs), thus the separation is quite difficult to obtain. On the contrary, the self-assembling map produces an almost perfect separation of the objects on the nodes of the net, as shown in Figure 29.b1. Interpretability is granted by the weights, reported in Figure 29.b2, whose importance for each node, and hence for the separation in each class, is indicated on a grey scale (the lighter, the more positive in sign; the darker, the more negative in sign).

s0095 4 CLUSTERING TECHNIQUES

p0795 Cluster analysis methods represent a family of EMDA tools alternative or complementary to the projection to latent variables tool discussed so far. The main target of cluster analysis is to find groups within a given data set, based on the principle for which similar objects are represented by close points in the space of the variables which describe them. The possible methods differ either in how groups are defined or in the algorithm used to create the groups. Generally speaking, group definition is based on within-group measures (e.g. high similarity between observations) or alternatively on between-group measures (e.g. maximum distance between objects), while clustering algorithms are based on different ways to define proximity, either similarities or dissimilarities. The most intuitive way to define the similarity level of samples (the concept of dissimilarity is complementary, that is its value increases the more the objects are different, while similarity increases the more the objects are similar) is based on the conversion of the $N \times M$ data

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:00 Page Number: 60

B978-0-444-59528-7.00003-X, 00003

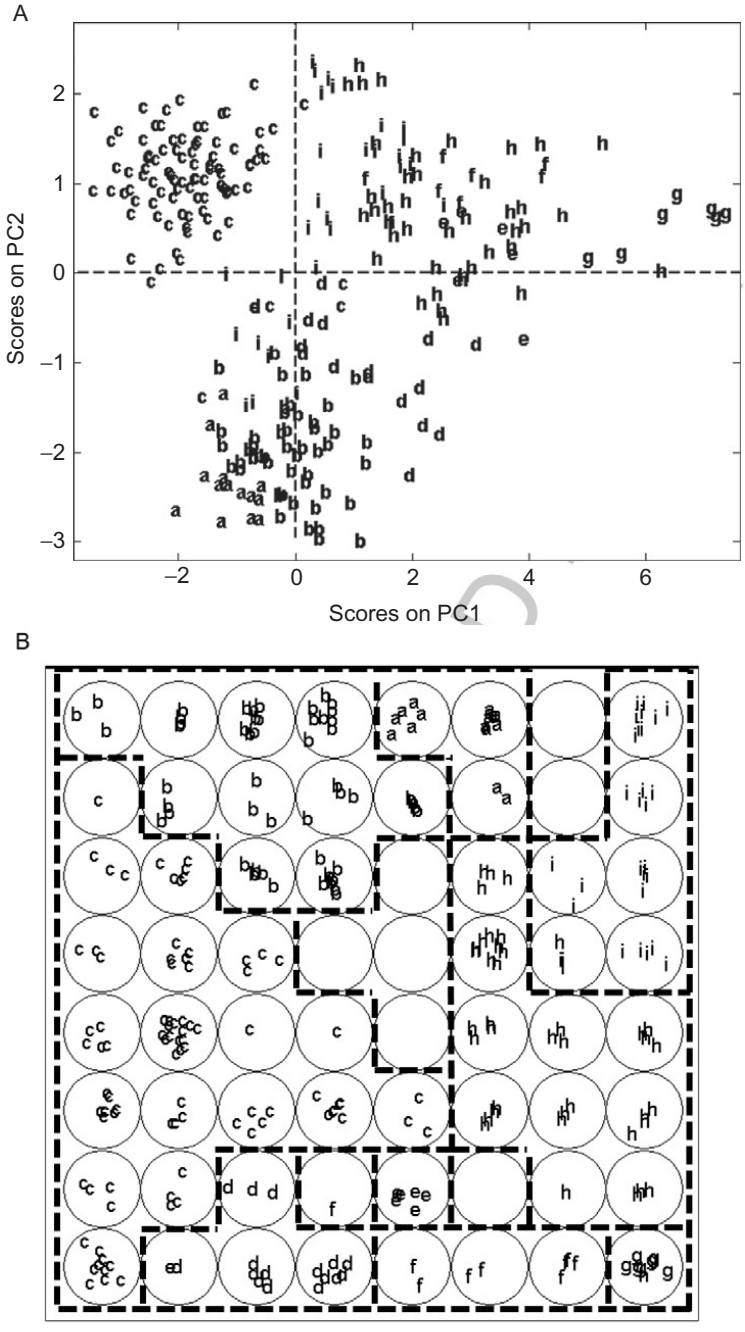
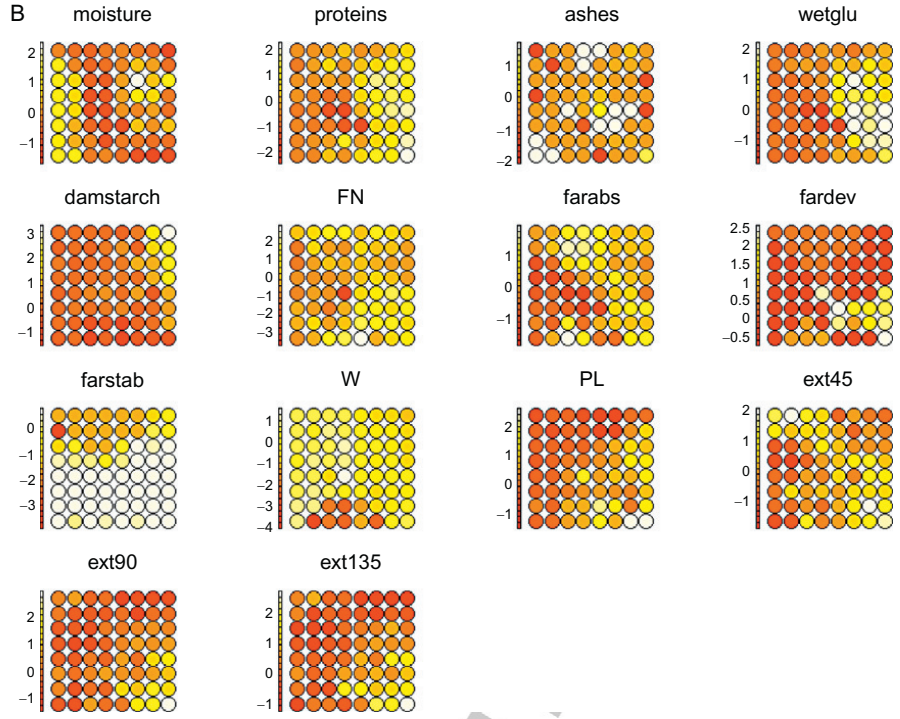


FIGURE 29—Cont'd

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:03 Page Number: 61



f0145 **FIGURE 29** Comparison of (A) principal component analysis and (B) self-organizing Kohonen's map for the FlourRheo data set. See Figure 23 for the legend. (b1) Distribution of the samples of the data set on the different nodes of the map. Dashed lines encompass neighbour nodes, where samples of the same category are grouped; and (b2) weight profiles for each of the variables on the different nodes. (For colour version of this figure, the reader is referred to the online version of this chapter.)

matrix in an $N \times N$ matrix of distances **D** obtained by defining a metric, such as the Euclidean distance:

$$d_{ij} = \sqrt{\sum_{m=1}^{M} \left(x_{im} - x_{jm}\right)^2} \tag{26}$$

where the sum is extended over the M variables which characterize each pair of objects i and j. It is clear that $d_{ij} = 0$ when $i = j$, and $d_{ij} > 0$ when $i \neq j$, which leads to the definition of a similarity matrix **S**, whose elements are

$$s_{ij} = 1 - \frac{d_{ij}}{d_{\max}} \tag{27}$$

p0800 Similarity ranges in the interval [0, 1] and assumes higher values the more similar the two objects i and j are. Of course, several different distance measurements can be implemented to evaluate similarity among objects, and also

B978-0-444-59528-7.00003-X, 00003

different similarity criteria can be established in the algorithm. Moreover, in many cases, it can be interesting to cluster variables together, instead of samples. In this case, a common technique to relate variables can be the use of their correlation coefficient as a proximity index on which to base one of the clustering methods which can be used for objects, which will be briefly described in this section.

p0805 As far as clustering algorithms are concerned, the wide choice of methods is related to the fact that clusters themselves can have very different characteristics in terms of shape, dimension and density, and each different cluster analysis approach is more oriented towards detecting a particular type of cluster rather than others, for example they work better when objects form round, dense clusters, rather than having elongated, overlapping distributions. However, although model-based clustering is possible, it goes beyond the exploratory purpose, as it requires quite a lot of *a priori* knowledge on the system: the best approach is then evaluating the outcome of methods suitable for different situations, and obtaining from their results information on the kind and degree of clustering which is present in the data.

p0810 As said earlier in this chapter, grouping of objects can be explored in different ways; to start with, let us consider visual inspection of scatter plots. When the number of variables, hence scatter plots, to consider grows over three, cluster analysis methods can offer a simplification advantage as well as projection to latent variables methods do. Moreover, projection methods can also benefit from cluster analysis methods, as it is possible to operate on the new, lower-dimensional data set, for example, on the score values of a PCA, with a clustering method, in order to further enhance the grouping of objects. This is particularly useful when the number of retained components is quite high, thus requiring the use of several scores scatter plots. Using cluster analysis on the scores themselves allows obtaining a direct inspection of clusters in only one graph: however, as in the more general case, no information can be obtained about which variables (here, PCs) are responsible for the formation of clusters.

p0815 Clustering algorithms can be divided into two main families [123,124]: partitioning and hierarchical methods. Partitioning [125] aims to segment a large data set of heterogeneous objects into k clusters, where k is either known *a priori* or hypothesized in an explorative way (k-clustering) or 'discovered' by the algorithm in an iterative way. A method representative of this family is MacQueen's k-means [126,127]. On the contrary, hierarchical clustering operates on a nested decomposition at various levels of similarity/dissimilarity, working either according to a bottom-up or top-down approach. The bottom-up approach leads to agglomerative clustering methods [128,129], which begin with each data as a distinct cluster and progress by merging clusters on the basis of their similarity, up to a stopping criterion (such as a threshold in similarity, or even the fact that all objects have been finally linked). Some representative methods of this family are single linkage, average

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:09 Page Number: 63

B978-0-444-59528-7.00003-X, 00003

linkage and complete linkage [130,131]. The top-down approach is at the basis of divisive clustering methods, where, in the beginning, all data are in a single cluster and are continuously separated until the stopping criterion is reached. Each method aims to define clusters, whose position in the M-dimensional space is defined by a centroid, the vector of the means of the variables computed over the elements which belong to the cluster. The instrument which is used to visualize the clustering result is called a dendrogram, which reports in a graphic way the degree of similarity at which each object and cluster is linked.

p0820 One of the most intuitive ways to describe how cluster analysis works in practice is by referring to the agglomerative hierarchical cluster analysis (HCA) method. Beside the common preliminary steps already discussed, that is definition of the metric (Euclidean, Mahalanobis, Manhattan distance, etc.) and calculation of the distance matrix and the corresponding similarity matrix, the analysis continues according to a recursive procedure such as

o0060 **1.** The two most similar objects are identified (i.e. those which have the highest similarity degree)

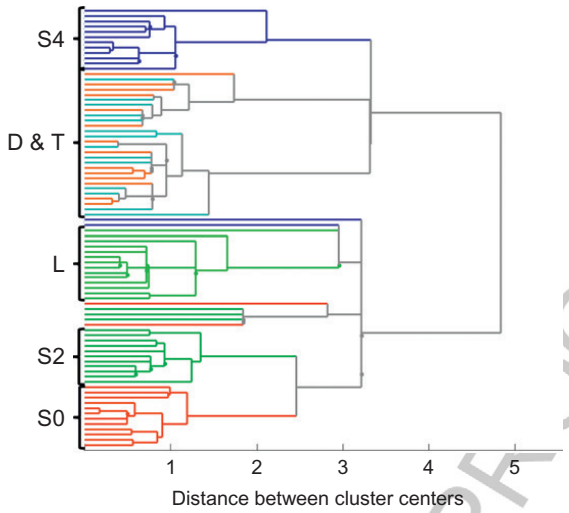
o0065 **2.** The two objects found at point 1 are linked in a cluster

o0070 **3.** A calculation of the similarity index of the new cluster versus all the other objects is performed. The similarity index calculation criterion differs according to the chosen clustering method, but the operation has the common result of substituting in the similarity matrix the rows and columns related to the two objects which have just been linked with a new row and column that report the similarity index of the new cluster with all the remaining objects.

p0840 The procedure is repeated by moving to the next pair of most similar objects: it is important to note whether in the beginning the comparison is done between objects, whereas in the following steps it is done by comparing clusters, according to one of the several similarity criteria which can be considered. For example, the centroid linkage criterion consists in substituting the objects which form the new cluster with the centroid of the cluster, so that the updated similarity matrix contains the distances between the centroids of the new clusters. The result of the procedure is represented as a dendrogram, as reported in Figure 30. Here, the objects are reported on the y-axis, while on the x-axis the similarity or the distance is reported. The lines which depart from each object are connected according to the degree of similarity at which the linkage between objects or clusters happens, so that it is possible to visualize in a fast way which level of similarity intercourses among the samples. In this example, the GCbreadProcess data set has been considered (preprocess: autoscale) in an agglomerative HCA with the centroid linkage criterion. It is possible to observe, when compared to other analyses of the same data set (e.g. boxplot, Figure 2, or PCA, Figure 24), that the subdivision in clusters is already present, as well as the common similarity of process phases D

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:09 Page Number: 64

B978-0-444-59528-7.00003-X, 00003



f0150 **FIGURE 30** HCA of the GCbreadProcess data set: representation of the dendrogram. (For colour version of this figure, the reader is referred to the online version of this chapter.)

and T. In this case, HCA represents a fast check on the presence of clusters, and possibly outliers (it is manifest that some samples connect to the main classes at higher distances, although belonging to that given process phase). However, information on variable influence is lost, which means that this graph should be analysed, for instance, together with a boxplot, to understand which variables are responsible for the separation into clusters, although it might not be able to explore the relationships among them, as it would be possible in PCA.

s0100 5 REMARKS

p0845 In this chapter, we have tried to offer a view, and recommendations for use, of the available tools to look at data, however complex and burdened by unsystematic variability they can be. Food data analysis is rich of real challenges, in this context, as we have shown with real application data. The focus has been twofold: furnishing a guide tour through data exploration and furnishing salient references to deepen the knowledge of specific aspects and less-known/applied techniques.

p0850 Explorative data analysis and especially EMDA offers an integrated set of methods to furnish us computer-aided 'eyes' to have a global perception of the high-dimensional world, and to do what we, as researchers, appreciate better: unravel relations, understand/connect patterns, formulate hypothesis and progress further. In other words, they offer us the possibility of handling systems complexity beyond reductionism.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:12 Page Number: 65

B978-0-444-59528-7.00003-X, 00003

REFERENCES

[1] Tukey JW. Exploratory data analysis. Lebanon, IN, USA: Addison-Wesley; 1977.
[2] Tukey JW. Sunset salvo. Am Stat 1986;40:72–6.
[3] Scott D. Multivariate density estimation: theory, practice, and visualization. New York: John Wiley and Sons; 1992.
[4] Shimazaki H, Shinomoto S. A method for selecting the bin size of a time histogram. Neural Comput 2007;19:1503–27.
[5] Chambers J, Cleveland W, Kleiner B, Tukey P. Graphical methods for data analysis. Boston: Wadsworth; 1983.
[6] Massart DL, Smeyers-Verbeke J, Capron X, Schlesier K. Visual presentation of data by means of box plots. LC–GC Europe 2005;18:215–8.
[7] McGill R, Tukey JW, Larsen WA. Variations of box plots. Am Stat 1978;32:12–6.
[8] Munck L, Nørgaard L, Engelsen SB, Bro R, Andersson CA. Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. Chemometr Intell Lab 1998;44:31–60.
[9] Perrota N, Trelea IC, Baudrita C, Trystram G, Bourgine P. Modelling and analysis of complex food systems: state of the art and new trends. Trends Food Sci Technol 2011;22:304–14.
[10] Gonzalvez A, Armenta S, De la Guardia M. Trace-element composition and stable-isotope ratio for discrimination of foods with protected designation of origin. Trends Anal Chem 2009;28:1295–311.
[11] Reid LM, O'Donnell CP, Downey G. Recent technological advances for the determination of food authenticity. Trends Food Sci Technol 2006;17:344–53.
[12] Cozzolino D. Recent trends on the use of infrared spectroscopy to trace and authenticate natural and agricultural food products. Appl Spectrosc Rev 2012;47:518–30.
[13] Bevilacqua M, Bucci R, Magrì AD, Magrì AL, Marini F. Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: a case study. Anal Chim Acta 2012;717:39–51.
[14] Li Vigni M, Durante C, Foca G, Marchetti A, Ulrici A, Cocchi M. Near infrared spectroscopy and multivariate analysis methods for monitoring flour performance in an industrial bread-making process. Anal Chim Acta 2009;642:69–76.
[15] Belton P, Capozzi F. Special issue: magnetic resonance in food: dealing with complex systems. Magn Reson Chem 2001:49:S1–S134.
[16] Ritota M, Marini F, Sequi P, Valentini M. Metabolomic characterization of Italian sweet pepper (*Capsicum annum* L.) by means of HRMAS-NMR spectroscopy and multivariate analysis. J Agric Food Chem 2010;58:9675–84.
[17] Callejón RM, Amigo JM, Pairo E, Garmón S, Ocaña JA, Morales ML. Classification of sherry vinegars by combining multidimensional fluorescence, PARAFAC and different classification approaches. Talanta 2012;88:456–62.
[18] Christensen J, Nørgaard L, Bro R, Engelsen SB. Multivariate autofluorescence of intact food systems. Chem Rev 2006;106:1979–94.
[19] Cordella CBY, Tekye T, Rutledge DN, Leardi R. A multiway chemometric and kinetic study for evaluating the thermal stability of edible oils by ^{1}H NMR analysis: comparison of methods. Talanta 2012;88:358–68.
[20] Cocchi M, Durante C, Grandi M, Manzini D, Marchetti A. Three-way principal component analysis of the volatile fraction by HS-SPME/GC of aceto balsamico tradizionale of Modena. Talanta 2008;74:547–54.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:14 Page Number: 66

B978-0-444-59528-7.00003-X, 00003

[21] Pereira AC, Reis MS, Saraiva PM, Marques JC. Madeira wine ageing prediction based on different analytical techniques: UV–vis, GC–MS, HPLC-DAD. Chemometr Intell Lab 2011;105:43–55.
[22] Baldwin EA, Bai J, Plotto A, Dea S. Electronic noses and tongues: applications for the food and pharmaceutical industries. Sensors 2011;11:4744–66.
[23] Pereira AC, Reis MS, Saraiva PM. Quality control of food products using image analysis and multivariate statistical tools. Ind Eng Chem Res 2009;48:988–98.
[24] Montalbán JM, De Juan A, Ferrer A. Multivariate image analysis: a review with applications. Chemometr Intell Lab 2011;107:1–23.
[25] Elmasry G, Kamruzzaman M, Sun DW, Allen P. Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: a review. Crit Rev Food Sci Nutr 2012;52:999–1023.
[26] Joliffe IT. Principal component analysis. 2nd ed. New York: Spinger-Verlag; 2002.
[27] Jackson JE. A user's guide to principal components. New York: John Wiley and Sons; 1991.
[28] Massart DL. Handbook of chemometrics and qualimetrics: part A. In: Massart DL, Vandeginste BGM, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J, editors. Data handling in science and technology series, part A, vol. 20. Amsterdam: Elsevier; 1998. p. 519–56 [chapter 17]. Au5
[29] Esbensen KH, Geladi P. Principal component analysis: concept, geometrical interpretation, mathematical background, algorithms, history, practice. In: Brown SD, Tauler R, Walczak B, editors. Comprehensive chemometrics: chemical and biochemical data analysis, vol. 2. Amsterdam: Elsevier Ltd.; 2009. p. 211–27 [chapter 2.13].
[30] Varmuza K, Filzmoser P. Introduction to multivariate statistics analysis in chemometrics. Boca Raton, FL: CRC Press Taylor & Francis Group; 2009.
[31] Wehrens R. Chemometrics with R: multivariate data analysis in the natural sciences and life sciences. Heidelberg, Dordrecht, London, New York: Springer; 2011.
[32] Burns DA, Ciurzak EW, editors. Handbook of near infrared analysis. New York: CRC Press Taylor & Francio Group; 2008. p. 151–88.
[33] Beebe KR, Pell RJ. ~~Seascholtz,~~ chemometrics: a practical guide. New York: John Wiley and Sons; 1998. Au5
[34] Davies AMC, Fearn T. Back to basics: the principles of principal component analysis. Spectrosc Eur 2004;16:20–3.
[35] Davies AMC. Back to basics: application of principal component analysis. Spectrosc Eur 2005;17:30–1.
[36] Wold S, Esbensen KH, Geladi P. Principal component analysis. Chemometr Intell Lab 1987;2:37–52.
[37] Smilde A, Bro R, Geladi P. Models for two-way one-block data analysis: component models. In: Multi-way analysis with applications, multiway analysis in the chemical sciences. John Wiley & Sons; 2004. p. 35–45. Au6
[38] Wu W, Massart DL, De Jong S. The kernel PCA algorithms for wide data. Part I: theory and algorithms. Chemometr Intell Lab 1997;36:165–72.
[39] Wold H. Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. In: Gani J, editor. Perspectives in probability and statistics. Sheffield, England: Applied Probability Trust; 1975.
[40] Wold H. Nonlinear estimation by iterative least square procedures. In: David FN, editor. Research papers in statistics, festschrift for J. Neyman. New York: Wiley; 1966. p. 411–44.
[41] Eckart C, Young G. The approximation of one matrix by another of lower rank. Psychometrika 1936;1:211–8.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:16 Page Number: 67

B978-0-444-59528-7.00003-X, 00003

[42] Golub GH, Reinsch C. Singular value decomposition and least squares solutions. Numer Math 1970;14:403–20.
[43] Gabriel KR. The biplot graphic display with application to principal component analysis. Biometrika 1971;58:453–67.
[44] Krooneberg PM. Applied multiway data analysis. Hoboken, NJ: John Wiley & Sons Inc.; 2008.
[45] Geladi P, Manley M, Lestander T. Scatter plotting in multivariate data analysis. J Chemometr 2003;17:503–11.
[46] Mardia K, Kent J, Bibby J. Multivariate analysis. In: Probability and mathematical statistics series. New York: Academic Press; 1979. Au7
[47] Cattel RB. The scree test for the number of factors. Multivar Behav Res 1996;1:245–76.
[48] Bro R, Kjeldahl K, Smilde AK, Kiers HAL. Cross-validation of component models: a critical look at current methods. Anal Bioanal Chem 2008;390:1241–51.
[49] Camacho J, Ferrer A. Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: theoretical aspects. J Chemometr 2012;26:361–73.
[50] Hotelling H. Multivariate quality control. In: Eisenhart C, Hastay M, Wallis WA, editors. Techniques of statistical analysis. New York: MacGraw-Hill; 1947. p. 111–84.
[51] Jackson JE, Mudholkar GS. Control procedures for residuale associated with principal component analysis. Technometrics 1979;21:341–9.
[52] Tracy ND, Young JC, Mason RL. Multivariate control charts for individual observations. J Qual Technol 1992;24:88–95.
[53] Ferrer A. Multivariate statistical process control based on principal component analysis (MSPC-PCA): some reflections and a case study in an autobody assembly process. Qual Eng 2007;19:311–25.
[54] Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. Technometrics 1995;37:41–59.
[55] Kourti T, MacGregor JF. Multivariate SPC methods for process and product monitoring. J Qual Technol 1996;28:409–28.
[56] Westerhuis JA, Gurden SP, Smilde AK. Generalized contribution plots in multivariate statistical process monitoring. Chemometr Intell Lab 2000;51:95–114.
[57] Conlin AK, Martin EB, Morris AJ. Confidence limits for contribution plots. J Chemometr 2000;14:725–36.
[58] Bro R, Smilde AK. Centering and scaling in component analysis. J Chemometr 2003;17:16–33.
[59] Cocchi M, Durante C, Grandi M, Lambertini P, Manzini D, Marchetti A. Simultaneous determination of sugars and organic acids in aged vinegars and chemometric data analysis. Talanta 2006;69:1166–75.
[60] Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. Scaling. In: Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS). Umea: Umetrics; 1999. p. 213–25. Au7
[61] Smilde AK, van der Werf MJ, Bijlsma S, van der Werff-van der Vat B, Jellema RH. Fusion of mass spectrometry-based metabolomics data. Anal Chem 2005;77:6729–36.
[62] Keun HC, Ebbels TMD, Antti H, Bollard ME, Beckonert O, Holmes E, et al. Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. Anal Chim Acta 2003;490:265–76.
[63] Wold S, Johansson E, Cocchi M. PLS: partial least squares projections to latent structures. In: Kubinyi Hugo, editor. 3D QSAR in drug design: theory, methods and applications. Leiden: ESCOM Science Publishers; 1983. p. 523–50.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:16 Page Number: 68

B978-0-444-59528-7.00003-X, 00003

[64] Eilers PHC. Parametric time warping. Anal Chem 2004;76:404–11.
[65] Forshed J, Schuppe-Koistinen I, Jacobsson SP. Peak alignment of NMR signals by means of a genetic algorithm. Anal Chim Acta 2003;487:189–99.
[66] Savorani F, Tomasi G, Engelseni SB. COSHIFT, a versatile tool for the rapid alignment of 1D NMR spectra. J Magn Reson 2010;202:190–202.
[67] Tomasi G, Savorani F, Engelsenicoshift SB. An effective tool for the alignment of chromatographic data. J Chromatogr A 2011;1218:7832–40.
[68] Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem 1964;36:1267.
[69] Van Den Bogaert B. Finding frequencies in signals: the Fourier transform & when frequencies change in time: towards the wavelet transform. In: Walczak B, editor. Wavelet in chemistry. Elsevier Science B.V.; 2000. p. 33–56. Au6
[70] Davies AMC. Back to basics: spectral pre-treatments—derivatives. Spectrosc Eur 2007;19:32–3.
[71] Rinnan A, Van Der Berg F, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. Trends Anal Chem 2009;10:1201–22.
[72] Kohler A, Zimonja M, Segtnan V, Martens H. Standard normal variate, multiplicative signal correction and extended multiplicative signal correction preprocessing in biospectroscopy. In: Brown SD, Tauler R, Walczak B, editors. Comprehensive chemometrics: chemical and biochemical data analysis. vol. 2. Amsterdam: Elsevier Ltd.; 2009. p. 211–27 [chapter 2.09].
[73] Davies AMC. Something has happened to my data; potential problems with standard normal variate and multiplicative scatter correction pre-treatments. Spectrosc Eur 2009;21:16–9.
[74] Afseth NK, Kohler A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. Chemometr Intell Lab 2012;117:92–9.
[75] Bylesjo M, Rantalainen M. Model based preprocessing and background elimination: OSC, OPLS, and O2PLS. In: Brown SD, Tauler R, Walczak B, editors. Comprehensive chemometrics: chemical and biochemical data analysis. vol. 2. Amsterdam: Elsevier Ltd.; 2009. p. 129–37 [chapter 2.08].
[76] Norgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl Spectrosc 2000;54:413–9.
[77] Xiaobo Z, Jiewen Z, Povey MJW, Holmes M, Hanpin M. Variables selection methods in near-infrared spectroscopy. Anal Chim Acta 2010;667:14–32.
[78] Niazi A, Leardi R. Genetic algorithms in chemometrics. J Chemometr 2012;26:345–51.
[79] Li Vigni M, Baschieri C, Foca G, Marchetti A, Ulrici A, Cocchi M. Monitoring flour performance in bread making. In: Preedy VR, Watson RR, Patel VB, editors. Flour and breads and their fortification in health and disease prevention. London: Academic Press, Elsevier; 2011. p. 15–25.
[80] Li Vigni M. Wheat flour and industrial bread-making: a multivariate approach to quality and process monitoring. Doctoral thesis, University of Modena and Reggio Emilia, Italy; 2010. p. 102–11.
[81] Smilde A, Bro R, Geladi P. Multi-way analysis with applications. In: Multiway analysis in the chemical sciences. John Wiley & Sons, Ltd.; 2004. Au8
[82] Li Vigni M, Durante C, Foca G, Ulrici A, Møller Jespersen BP, Bro R, et al. Wheat flour formulation by mixture design and multivariate study of its technological properties. J Chemometr 2010;24:523–33.
[83] Li Vigni M, Cocchi M. Near infrared spectroscopy and multivariate analysis to evaluate wheat flour doughs leavening and bread properties. Anal Chim Acta 2013;764:17–23.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:19 Page Number: 69

B978-0-444-59528-7.00003-X, 00003

[84] Efron B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika 1981;68:589–99.
[85] Lunneborg CE. Data analysis by resampling. USA: Duxbury Press; 2000.
[86] Ferrer A. Statistical control of measures and processes. In: Brown SD, Tauler R, Walczak B, editors. Comprehensive chemometrics: chemical and biochemical data analysis. Amsterdam: Elsevier Ltd.; 2009.
[87] Wentzell PD. Other topics in soft-modeling: maximum likelihood-based soft-modeling methods. In: Brown SD, Tauler R, Walczak B, editors. Comprehensive chemometrics: chemical and biochemical data analysis. Amsterdam: Elsevier Ltd.; 2009.
[88] Wentzell PD, Hou S. Exploratory data analysis with noisy measurements. J Chemometr 2012;26:264–81.
[89] Wentzell PD, Andrews DT, Hamilton DC, Faber K, Kowalski BR. Maximum likelihood principal component analysis. J Chemometr 1997;11:339–66.
[90] Wentzell PD, Lohnes MT. Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations. Chemometr Intell Lab 1999;45:65–85.
[91] Daszykowski M, Kaczmarek K, Vander Heyden Y, Walczak B. Robust statistics in data analysis—a review basic concepts. Chemometr Intell Lab 2007;85:203–19.
[92] Devlin JS, Gnanadesikan R, Kettering JR. Robust estimation of dispersion matrix and principal components. J Am Stat Assoc 1981;76:354–62.
[93] Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. New York: John Wiley & Sons Inc.; 1987.
[94] Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, et al. robustbase: Basic Robust Statistics. 76:45–54.
[95] Todorov V, Filzmoser P. An object-oriented framework for robust multivariate analysis. J Stat Softw 2009;32:1–47.
[96] Daszykowski M, Serneels S, Kaczmarek K, Van Espen P, Croux C, Walczak B. TOMCAT: a MATLAB toolbox for multivariate calibration techniques. Chemometr Intell Lab 2007;85:269–77.
[97] Verboven S, Hubert M. LIBRA: a MATLAB library for robust analysis. Chemometr Intell Lab 2005;75:127–36.
[98] Friedman JH, Tukey JW. A projection pursuit algorithm for exploratory data analysis. IEEE Trans Comput 1974;C-23:881–90.
[99] Jones MC, Sibson R. What is projection pursuit? J Roy Stat Soc Ser A (Gen) 1987;150:1–37.
[100] Comon P. Independent component analysis: a new concept? Signal Process 1994;36:287–314.
[101] Lee TW. Independent component analysis: theory and applications. Boston, MA: Kluwer Academic Publishers; 1998.
[102] Hyvärinen A, Karhunen J, Oja E. Independent component analysis. New York: Wiley; 2001.
[103] Hyvärinen A, Oja E. Independent component analysis: algorithms and application. Neural Netw 2000;13:411–30.
[104] Comon P, Jutten C. Handbook of blind source separation, independent component analysis and applications. Oxford, UK: Academic Press; 2010.
[105] Bugli C, Lambert P. Comparison between principal component analysis and independent component analysis in electroencephalograms modelling. Biometrical J 2006;48:1–16.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:20 Page Number: 70

B978-0-444-59528-7.00003-X, 00003

[106] Aguilera T, Lozano J, Paredes JA, Alvarez FJ, Suarez JI. Electronic nose based on independent component analysis combined with partial least squares and artificial neural networks for wine prediction. Sensors 2012;12:8055–72.
[107] Ammari F, Cordella CBY, Boughanmi N, Rutledge DN. Independent components analysis applied to 3D-front-face fluorescence spectra of edible oils to study the antioxidant effect of *Nigella sativa* L. extract on the thermal stability of heated oils. Chemometr Intell Lab 2012;113:32–42.
[108] Westad F. Independent component analysis and regression applied on sensory data. J Chemometr 2005;19:171–9.
[109] Bouveresse DJR, Moya-González A, Ammari F, Rutledge DN. Two novel methods for the determination of the number of components in independent components analysis models. Chemometr Intell Lab 2012;112:24–32.
[110] Westad F, Kermit M. Cross validation and uncertainty estimates in independent component analysis. Anal Chim Acta 2003;490:341–54.
[111] Cox TF, Cox MAA. Multidimensional scaling. ~~London:~~ Chapman and Hall; 2001. Au10
[112] Borg I, Groenen PJF. Modern multidimensional scaling. 2nd ed. New York: Springer; 2005. Au10
[113] Sammon JW. A nonlinear mapping for data structure analysis. IEEE Trans Comput 1969;18:401–9.
[114] Kruskal JB, Wish M. Multidimensional scaling. Beverly Hills, CA: Sage Publications; 1978.
[115] Gunden C, Thomas T. Assessing consumer attitudes towards fresh fruit and vegetable attributes. J Food Agric Environ 2012;10:85–8.
[116] Ballester J, Patris B, Symoneaux R, Valentin D. Conceptual vs. perceptual wine spaces: does expertise matter? Food Qual Prefer 2008;19:267–76.
[117] Lee SJ, Noble AC. Use of partial least squares regression and multidimensional scaling on aroma models of California Chardonnay wines. Am J Enol Viticult 2006;57:363–70.
[118] Taguchi Y, Oono Y. Relational patterns of gene expression via non-metric multidimensional scaling analysis. Bioinformatics 2005;21:730–40.
[119] Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern 1982;43:59–69.
[120] Kohonen T. Self-organizing maps. 3rd ed. Springer; 2001. Au6
[121] Príncipe JC, Miikkulainen R, editors. Advances in self-organizing maps. Springer: Berlin; 2009.
[122] Yan J. Package 'som', version 0.3-5; 15 February 2012. http://cran.r-project.org/web/packages/som/.
[123] Todeschini R. Introduzione alla chemiometria. Napoli: EdiSES; 1998.
[124] Lee I, Yang J. Common clustering algorithms. In: Brown SD, Tauler R, Walczak B, editors. Comprehensive chemometrics: chemical and biochemical data analysis. vol. 2. Amsterdam: Elsevier Ltd.; 2009. p. 211–27 [chapter 2.27].
[125] Berry MJA, Linoff GS. Data mining techniques for marketing, sales and customer support. New York: John Wiley & Sons; 1997.
[126] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Le Cam L, Neyman J, editors. 5th Berkeley symposium on mathematical statistics and probability. vol. 1. 1967. p. 281–97. Au11
[127] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min Knowl Disc 1998;2:2283–304.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:20 Page Number: 71

B978-0-444-59528-7.00003-X, 00003

[128] Lance GN, Williams WT. A general theory of classificatory sorting strategies: II. Clustering systems. Comput J 1967;10:271–7.
[129] Gower JC. A comparison of some methods of cluster analysis. Biometrics 1967;23:623–8.
[130] Sneath PHA. The application of computers to taxonomy. J Gen Microbiol 1957;17:201–26.
[131] Ward J. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963;58:236–44.

Comp. by: pdjeapradaban Stage: Proof Chapter No.: 3 Title Name: DHST
Date:18/4/13 Time:12:02:22 Page Number: 72

B978-0-444-59528-7.00003-X, 00003

Non-Print Items

Abstract

In the food research and production field, system complexity is increasing and several new challenges are emerging every day. This implies an urgent necessity to extract information and obtain models capable of inferring the underlying relationships that link all the variability sources which characterize food or its production process (e.g. compositional profile, processing conditions) to very general end properties of foodstuff, such as the healthiness, the consumer perception, the link to a territory and the effect of the production chain itself on food.

This makes a 'deductive' theory-driven research approach inefficient, as it is often difficult to formulate hypotheses. Explorative multivariate data analysis methods, together with the most recent analytical instrumentation, offer the possibility to come back to an 'inductive' data-driven attitude with a minimum of *a priori* hypotheses, instead helping in formulating new ones from the direct observation of data.

The aim of this chapter is to offer the reader an overview of the most significant tools that can be used in a preliminary, exploratory phase, ranging from the most classical descriptive statistics methods, to multivariate analysis methods, with particular attention to projection methods. For all techniques, examples are given so that the main advantage of these techniques, which is a direct, graphical representation of data and their characteristics, can be immediately experienced by the reader.

Keywords: Descriptive statistics; Projection techniques; Principal component analysis; Clustering techniques; Multivariate explorarory analysis Au4