10 The exploitation of chemometric methods in the analysis of spectroscopic data: application to olive oils

A. JONES, A. D. SHAW, G. J. SALTER, G. BIANCHI and D. B. KELL

10.1 Introduction

Multivariate analysis is the term used to describe the analysis of data where numerous observations or variables are obtained for each object studied (Afifi and Clark, 1996). The identity or value of any one object sample will be reflected in some or all of the variables measured to a greater or lesser degree. With spectroscopic data it is not generally possible to identify or quantify an object from one variable (Mark, 1991). This may, however, be achieved by disentangling the complicated interrelationships between a number of the variables by means of multivariate statistical methods (Martens and Næs, 1989).

Classification problems are those where the aim is to identify objects, for example the region of origin of an olive oil. Quantification problems are those where the aim is to predict the magnitude of a quantity, for example the amount of an adulterant in an olive oil. Multivariate methods may be applied equally to either.

In recent years, more powerful computers and widely available statistical software have led to a tremendous increase in the use of multivariate data analysis. With the power of modern computers, most efforts have focused on analysing the whole spectrum of data (Brereton and Elbergali, 1994). This approach relies on statistical software to produce optimum results from the data fed into it; much of these data could contain little information, and therefore be of no value to the model.

Recent research has shown that judicious variable selection can improve statistical predictions of models (Baroni *et al.*, 1992; Brereton, 1995; Brereton and Elbergali, 1994; Broadhurst *et al.*, 1997; Brown, 1993; Cruciani and Watson, 1994; Defalguerolles and Jmel, 1993; Hazen, Arnold and Small, 1994; Heikka, Minkkinen and Taavitsainen, 1994; Kubinyi, 1994a, b, 1996; Lindgren *et al.*, 1995; Norinder, 1996; Shaw *et al.*, 1996, 1997; Sreerama and Woody, 1994); statistical theory, in particular the parsimony principle

(Flury and Riedwyl, 1988; Seasholtz and Kowalski, 1993), supports these results.

Olive oil is an ideal candidate for multivariate analysis. For economic reasons, the labelling of olive oils is frequently falsified (Collins, 1993; Firestone and Reina, 1987; Firestone, Carson and Reina, 1988; Firestone *et al.*, 1985; Li-Chan, 1994; Simpkins and Harrison, 1995b; Zamora, Navarro and Hidalgo, 1994), so there is a need for easy and cheap methods for identification. This chapter concentrates on the application of multivariate methods to nuclear magnetic resonance (NMR) and pyrolysis mass spectrometry (PyMS) data. It provides a brief introduction to principal components analysis (PCA), principal components regression (PCR), partial least squares regression (PLS) and the use of artificial neural networks (ANNs), then moves on to variable selection and its application to olive oil data.

10.2 Olive oil

10.2.1 Economics

The value of olive oil produced annually is around US$2.5 billion (Kiritsakis, 1991), other olive products amounting to around US$300 million. A total of 9.4 million tonnes of olive fruit are produced per annum from 805 million olive trees worldwide, occupying some 24 million acres of land. Some 98% of these trees are in the Mediterranean area. Of the 60 million tonnes of seed oil consumed worldwide every year, 2 million tonnes are olive oil (Anon., 1994).

Almost 25% of the farming income in the Mediterranean basin as a whole comes from olive products, Spain and Italy being far and away the largest producers, with Greece (with around half the production of Spain and Italy) coming third. In 1987, Italy and Spain contributed about 65% of world olive production (Salunkhe *et al.*, 1991).

Olive production often follows a two-year cycle, a good crop one year being followed by a poor or medium crop the next year. This is probably the biggest problem facing the olive industry (Kiritsakis, 1991).

10.2.2 Chemistry

The olive fruit is a drupe, that is to say it contains a stone, pulp and an outer skin (like a plum). The chemical composition of the fruit (Bianchi, Giansarte and Lazzari, 1996) is approximately as given in Table 10.1. Table olives generally have a lower oil content (around 10%–14%) than olives used for oil production.

The main fatty acids contained in olive oil, which are attached (esterified) to the glycerol backbone in one of the three locations α, β or α' are shown in Table 10.2.

Table 10.1 Chemical composition of the olive fruit

Component	Percentage
Water	48
Oil	21
Mono- and disaccharides	3
Polysaccharides	27
Waxes, triterpenes and phenols	1
Other minor components	Trace
Total	100

Table 10.2 The main fatty acids contained in olive oil

Acid	Abbreviation	Percentage range	Structure
Saturated			
Palmitic acid	C16:0	7.5–20	$CH_3-(CH_2)_{14}-COOH$
Stearic acid	C18:0	0.5–5	$CH_3-(CH_2)_{16}-COOH$
Lignoceric acid	C24:0	1.0 (max.)	$CH_3-(CH_2)_{22}-COOH$
Monounsaturated			
Palmitoleic acid	C16:1Δ9	0.3–3.5	$CH_3-(CH_2)_5HC=CH-(CH_2)_7-COOH$
Oleic acid	C18:1Δ9	56–83	$CH_3-(CH_2)_7-HC=CH-(CH_2)_7-COOH$
Eicosenoic acid	C20:1Δ11	trace	$CH_3-(CH_2)_7-HC=CH-(CH_2)_9-COOH$
Polyunsaturated			
Linoleic acid	C18:2Δ9, 12	3.5–20	$CH_3-(CH_2)_4-HC=CH-CH_2-HC=CH-(CH_2)_7-COOH$
Linolenic acid	C18:3Δ9, 12, 15	0.0–1.5	$CH_3-CH_2-(HC=CH-CH_2)_3-(CH_2)_6-COOH$
Arachidonic acid	C20:4Δ5, 8, 11, 14	0.8 (max.)	$CH_3-(CH_2)_4-(HC=CH-CH_2)_4-(CH_2)_2-COOH$

Other saponifiable constituents include phosphatides. Minor constituents, together called the 'unsaponifiable fraction', include hydrocarbons, terpenes, fatty alcohols, wax, phenols and amino acids. In addition, there will be a small amount of free fatty acids (FFAs), the amount being dependent on the grade of oil.

Virgin olive oil is the oil extracted by purely mechanical means from sound, ripe fruits of the olive tree (*Olea europaea* L.). Extra virgin olive oil is absolutely perfect in flavour and odour, and has a maximum free fatty acid content in terms of oleic acid of 1 g per 100 g (EC, 1991; Goodacre, Kell and Bianchi, 1993; Kiritsakis, 1991).

Compared with other edible oils, olive oil contains a low percentage of saturated fatty acids (that is, fatty acids with no double bonds in the carbon chain) at around 16% (mainly palmitic, 16:0). It contains a high percentage of monounsaturated fatty acids, around 70% (mainly oleic, 18:1) (MAFF, 1995; Mottram, 1979) and around 15% polyunsaturated fatty acids (mainly linoleic, 18:2).

Virgin olive oil generally conserves for a longer time than do most other vegetable oils (the maximum duration of optimal usage is often around 18 months). It is suggested that this is a result of the combined effect of a high monounsaturate content and some of the minor constituents of the oil, which act as antioxidants. Perrin (1992) identifies phenolic compounds as the main antioxidants; Kiritsakis and Dugan (1985) additionally mention carotene, whilst noting that chlorophyll has the opposite effect.

Garcia *et al.* (1996) note the effects of storage temperature of the fruits before oil extraction on the quality of olive oil. Prolonged storage at the wrong temperature (which, at least in Spain, is typically in great heaps in the open air) can, they point out, increase the amount of free oleic acid; this affects the grade of the oil. This is a problem in Spain, as there are insufficient mills to process all the olives at the peak of the harvesting season. The problem may be overcome by storage in cool buildings (Kiritsakis, 1991).

Other factors affecting the chemistry of the oil are the extraction method used (Aparicio, Navarro and Ferreiro, 1991; Rade *et al.*, 1995; Ranalli and Martinelli, 1994), storage conditions (Garcia *et al.*, 1996; Kiritsakis, 1984; Rade *et al.*, 1995), orography (e.g. distance from sea, altitude) (Aparicio, Ferreiro and Alonso, 1994; Armanino, Leardi and Lanteri, 1989) and the time of harvest (Boschelle *et al.*, 1994; Haumann, 1996; Tsimidou and Karakostas, 1993).

There are various accepted classifications of olive oil (Kiritsakis, 1991). The European Union rules governing olive oil classification are very comprehensive, covering in 83 pages not only the oil characteristics but also the methods of analysis to be used, right down to the selection of tasters (EC, 1991).

10.2.3 Health aspects

Much has been made in recent years of the so-called 'Mediterranean diet' (Gussow, 1995; Trichopoulou *et al.*, 1995b; Tsimidou, 1995), of which olive oil is a basic ingredient. Olive oil has a fine aroma and a pleasant taste, which is generally agreed to be at its best in extra virgin olive oils, and is considered to have many nutritional and health benefits (Kiritsakis, 1991). It is almost the only vegetable oil to be consumed as it is, that is without raffination (excluding the little-consumed nut and sesame oils) (Perrin, 1992).

There are many varied claims and suggested reasons as to the health benefits. There is very strong evidence that olive oil consumption reduces the risk of death due to circulatory system diseases (Fraser, 1994; Kafatos and Comas, 1991). Visioli and Galli (Galli, Petroni and Visioli, 1994; Visioli and Galli, 1994, 1995) suggest that this is at least partially because of the presence of natural antioxidants (including the bitter-tasting glycosidic compound oleuropein) and micronutrients preventing low-density lipo-

protein (LDL) cholesterol from oxidizing [oxidized LDL particles are particularly atherogenic (Fraser, 1994)] and so retarding the formation of atherosclerotic lesions (MAFF, 1995). They say that these antioxidants and micronutrients are present in large amounts in extra virgin olive oil (lesser amounts are found in other grades of olive oil). They imply that these properties may be more important than the high monounsaturated/saturated fatty acid ratio, and even suggest (Visioli, Vinceri and Galli, 1995) that the wastewater used for washing the olive paste during oil production could be utilized as it too is high in antioxidants.

A diet relatively high in monounsaturated fatty acids (MUFAs) does in any case reduce the levels of the undesirable LDLs in the body (Bosaeus *et al.*, 1992; MAFF, 1995; Shepherd and Packard, 1992); indeed the reduction is as great as that of a low-fat, high-carbohydrate diet (Kafatos and Comas, 1991). There is also evidence that a high MUFA diet increases the beneficial high-density lipoprotein (HDL) cholesterol, in contrast to polyunsaturated fats, which decrease both LDL and HDL levels (Kafatos and Comas, 1991), although this finding is not universally accepted (Haumann, 1996).

It has also been suggested that increased olive oil consumption helps prevent the onset of rheumatoid arthritis, and reduces its severity (Linos *et al.*, 1991).

The importance of monounsaturated fats has been recognized only in recent years; not so long ago they were completely overlooked with regard to blood cholesterol levels in favour of polyunsaturated fats (Mottram, 1979, pp. 52–3).

Martin-Moreno *et al.* (1994) also note that olive oils contain a 'generous amount of antioxidants' and speculate that 'diets high in monounsaturated fats presumably yield tissue structures that are less susceptible to antioxidative damage than would be the case in high polyunsaturated diets' (p. 778). They identify an inverse correlation between breast cancer and olive oil intake, as do Trichopoulou and co-workers (Trichopoulou, 1995; Trichopoulou *et al.*, 1995a, b), who also claim that margarine consumption increases this risk. Trichopoulou *et al.* (1995c) suggest that olive oil consumption is one of the factors in the traditional Greek diet that is a cause of the longevity of those elderly people in a study group who followed that diet. Greece has the highest consumption of olive oil in the world, at 20.8 kg per person per year (Kiritsakis and Markakis, 1991), or 57 g a day (at 500 calories, around 20%–25% of the energy requirement of an adult).

Murphy (1995, p. 302) notes that 'over the next decade and beyond, we face the prospect of being able to engineer most major oil crops to produce the fatty acid composition of our choice'. So, if monounsaturates in olive oils were the only reason for the health benefits claimed, this selling point might not have much of a future.

10.2.4 Analysis

As a consequence of the benefits mentioned, and because of the amount of labour and land needed to produce a given amount of oil, olive oil commands a much higher price than do most other edible oils. This in turn means that there is a great temptation to adulterate the oil with a cheaper oil, such as olive pomace oil, corn oil, sunflower oil, or even lard or castor oil (Firestone and Reina, 1987; Firestone, Carson and Reina, 1988; Firestone *et al.*, 1985; Zamora, Navarro and Hidalgo, 1994). Rapeseed oil, which can have a similar quantity of oleic acid (Shahidi, 1990), and high oleic sunflower oil, are also popular choices as adulterants. In addition, it is claimed that many oils purported to be extra virgin had been processed in order to reduce the acidity level and so gain this classification. Firestone *et al.* (1985) reported on a US survey in which 4 out of 5 virgin olive oils were correctly labelled, compared with only 3 out of 20 olive oils. In 1988 they followed up the 1985 report (Firestone, Carson and Reina, 1988), noting some improvement. This time, although only 17 out of 31 virgin olive oils were correctly labelled, so were 15 out of 26 olive oils; over 40% were incorrectly labelled. A British Broadcasting Corporation (BBC) radio investigation into olive oils in May 1994 suggested similar figures for the British market.

The necessity to be able to detect adulterations in oils in general was highlighted in May 1981, when 20000 people became ill and 350 died in Spain after consuming oils containing 'refined' aniline denatured rapeseed oil (Aldridge, 1992).

One problem faced today is that, as detection methods become more sophisticated so too do the methods of the adulterators. Some methods, which rely on the detection of compounds which do not appear in the genuine product, are useless if the adulterator knows the technique and therefore removes the offending compound (Aparicio, Alonso and Morales, 1996).

Grob *et al.* (1994a) report that extra virgin oils can be distinguished by the presence of a substantial quantity of volatile components (i.e. they have not been deodorized). If none of these volatiles is present, the oil has been treated. 'Pure' oils, being a blend, are more difficult to distinguish. Grob *et al.* (1994b) were able to detect adulteration of olive oils down to 10% (even lower for most oils) by using LC–GC–FID (liquid chromatography–gas chromatography–flame ionization detection) by direct analysis of these minor components. They do note, however, that strong raffination made adulteration difficult to detect.

Historically, and indeed to the present day, panels of trained assessors have been used to differentiate olive oils (Aparicio and Morales, 1995; Aparicio, Gutierrez and Morales, 1992; Lyon and Watson, 1994; Morales *et al.*, 1995). Chemometric methods applied to the results of such panels do indeed give good results, but it is slow and restricted to the sensory

characteristics. Problems arising from different panels describing the same sensory attribute with different terms (which is quite understandable when panels are from different cultures and use different languages) may be overcome by means of sensory wheels, a technique which explores the relationships between sensory attributes (Aparicio and Morales, 1995; Boskou, 1996). Also, in the present scientific climate, where the culprits' methods are becoming ever more sophisticated, it is not possible to use taste panels to detect reliably the adulteration or misclassification of origin. Such a method cannot, according to Peri and Rastelli (1994), be used as a legal tool for evaluating quality or origin, although it still is (EC, 1991).

Goodacre, Kell and Bianchi (1993) were successful in detecting adulteration of extra virgin olive oil by using ANNs and PyMS. Extensions to this work are described below in the section on pyrolysis mass spectrometry.

The Institute of Food Research (IFR, 1994) used Fourier transform infrared (FTIR) spectroscopy and NMR (not stated, but presumably ^{13}C) for the identification of oils. Both methods successfully differentiated olive oils of 'differing botanical origin' and could also discriminate extra virgin and other grades of olive oil. Fatty acid composition was found to be the main factor for discriminating the origin, and 'other trace analytes' were used to distinguish the grade of oil. Other workers have also successfully applied FTIR to the analysis of olive oils and other edible oils (Ismail *et al.*, 1993; Lai, Kemsley and Wilson, 1994, 1995; van de Voort, 1994; van de Voort, Ismail and Sedman, 1995; van de Voort *et al.*, 1994a, b).

Despite the IFR results, Zamora, Navarro and Hidalgo (1994) were also able to distinguish between different grades of oil by means of ^{13}C NMR alone, which relies largely on the fatty acid composition. Forina and Tiscornia (1982) agree that fatty acid content is important in the geographical classification of olive oils.

Vlahov (1996), at the Istituto Sperimentale per la Elaiotecnica, has used ^{13}C NMR to determine the quantities of diacylglycerols (DGs), both 1,2-DG and 1,3-DG, in olive varieties Grossa di Cassano, Nebbio, Coratina, Leccino, Dritta and Caroleo. It was found that the later-ripening olive varieties, Coratina and Nebbio, had significantly lower DG totals than had the other varieties. The ratio of 1,2-DG and 1,3-DG also varied significantly between varieties. She concludes that the results may represent preliminary new parameters for future analysis. No monoacylglyceride signals were found in the samples analysed. Other preliminary work by Valhov and Angelo (1996) has suggested that the position of fatty acids within the triglycerides may be important for discrimination (Forina and Tiscornia, 1982; Zupan and Gasteiger, 1993).

Very recently, high-field ^{1}H NMR has been used with success by one group for regional and variety discrimination of virgin olive oil (Segre *et al.*, 1996) and oil quality (Sacchi, 1996). They suggest that proton NMR at very high field (they were using 600 MHz) can be a more powerful technique

than ^{13}C NMR for quality control of virgin olive oils. Their results for region and variety are promising but suggest that further work is yet required. Their variety results were obtained with oils from Umbria only, rather than oils from various regions.

Gigliotti, Daghetta and Sidoli (1994) used high-pressure liquid chromatography (HPLC) for geographical characterization of olive oils. Using ratios of OOO, SOO, LLL and OOL (O = oleic, S = stearic and L = linoleic acid) and ECN (equivalent carbon number) they were able to distinguish Moroccan, Tunisian, Greek, Spanish and Sicilian oils. Their results also show that many bought oils do not fit into any of their categories, suggesting the possibility of adulteration. Perrin (1992) notes that Tunisian olive oil has a much lower monounsaturate/polyunsaturate ratio than do oils of other countries (less than 3:1 compared with 4:1–10:1 generally).

Boschelle *et al.* (1994) applied chemometric methods to chemophysical data to identify the cultivar of olives from the Gulf of Trieste area. They were successful in distinguishing the local variety Bianchera from those newly introduced into the region.

Tsimidou and Karakostas (1993) used data on the percentage of five fatty acids (palmitic, palmitoleic, stearic, oleic and linoleic – data on minor acids were not available) to classify Greek olive oils by region. Their results show that year of harvest is more influential in PCA than is variety or region of origin. In contrast to the results shown later in this chapter (Section 10.4), they found more difficulty in distinguishing variety than region.

Zupan and Gasteiger (1993) used Kohonen ANNs to discriminate Italian olive oils by region, suggesting that this is much better than PCA for mapping onto a two-dimensional plane. The network inputs were the fatty acid content of 572 olive oils from nine different areas of Italy (north and south Apulia, Calabria, Sicily, inner Sardinia, coastal Sardinia, east and west Liguria and Umbria). The analysis used the percentage of eight fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, arachidic, linolenic and eicosenoic) in the oil, determined previously by unstated methods (presumably GC). Although not able to identify any of the regions with 100% accuracy, the Kohonen nets were able to predict correctly up to 302 oils from a test set of 322 (although 51 of these 302 were only correctly assigned by using a K nearest-neighbour decision for empty space hits). Forina and Tiscornia (1982), using the same data set, were able to predict oils with 94.5% accuracy by using a K nearest-neighbour decision projected onto the hyperspace of the training set variables. The test and training sets were randomly selected and were repeated ten times.

Garcia and López (1993) and Aparicio, Alonso and Morales (1994b) were able to distinguish between Italian, Spanish and Portuguese olive oils by using an expert system, SEXIA (Aparicio and Alonso, 1994). For the Italian regions, Garcia and López were able to distinguish 99% of their Sardinian

samples, 91.3% of the northern Italian samples, but only 77.4% of the southern Italian samples. They note that there is evidently greater disparity between oils from the south. These figures would appear to be in line with our results shown later in this chapter (Section 10.4.12), where Toscana region oils were much easier to predict than southern oils.

Guinda, Lanzón and Albi (1996) were able to distinguish five varieties of Spanish olive oils by examining the hydrocarbon fraction (excluding squalene) of 50 oils from six provinces of Spain. They did not use chemometric techniques, but a simple decision tree (e.g. *if Percentage First Fraction C25 ≥ 28 then variety = Empeltre else . . .*).

Sato (1994) showed that near infra-red spectroscopy can be used with PCA to discriminate many vegetable oils from each other, including olive oil. Schwaiger and Vojir (1994) had similar success with GC analysis and PCA, the first two principal components separating olive oil well from the other oils.

Although the Raman effect was discovered in 1928 by Sir Chandrasekhara Venkata Raman, it has not until recently been applied to food adulteration problems (Baeten *et al.*, 1996; Li-Chan, 1994; Ozaki *et al.*, 1992; Sadeghi-Jorabchi *et al.*, 1990, 1991). Baeten *et al.* (1996) used FT-Raman which, they claim, produces fluorescence-free spectra, using a 1.064 μm laser. They were able to detect adulteration with soybean, corn and olive pomace with 100% accuracy down to 1% adulterant. In fact 780 nm excitation in a confocal instrument (Williams, 1994; Williams *et al.*, 1994) produces excellent dispersive Raman spectra from olive oils in a wholly non-invasive fashion (N. Kaderbhai and the authors, unpublished observations). Baeten *et al.* (1996) comment that at present liquid and gas chromatography is the most accurate technique to determine adulteration, and it is this method that is the European Union adulteration standard (EC, 1991), but that FT-Raman has the potential for detecting adulterants beyond the limits of liquid and gas chromatography.

Not surprisingly, climate has been shown to affect the chemical composition of olive oil (Aparicio, Ferreiro and Alonso, 1994). Variations in chemical composition between regions are presumably largely explained by this factor (although there are probably other factors).

Simpkins and Harrison (1995a) summarize many of the methods used for detection of authenticity in olive oils and many other food products. They note that most new applications they reviewed relied on advanced statistical procedures for data analysis.

Li-Chan (1994) reviews current developments in the detection of adulteration of olive oil, pointing out that multivariate spectroscopic methods of analysis derive their power from the simultaneous use of multiple variables in the spectrum. The effect of interference in some variables can then be reduced by the calibration method used (PCR, PLS, etc.). This type of approach has previously been used for crude petrochemical oil, as described by Kvalheim *et al.* (1985) and Brekke *et al.* (1990), with very promising results.

Aparicio, Alonso and Morales (1996) suggest that the future lies in spectroscopic (probably, they say, FT-Raman) and chromatographic techniques, coupled with mathematical algorithms.

10.3 Data acquisition methods

10.3.1 Nuclear magnetic resonance

According to Yoder and Schaeffer (1987), NMR spectroscopy is probably the most powerful tool available to the chemist for the probing of the structure of molecules. It can rapidly produce data from which structural formulas and even some three-dimensional aspects of the structure of molecules can be deduced. Its use is in the detection of nuclear-spin reorientation in an applied magnetic field (Campbell and Dwek, 1984). Indeed Harris (1986) claims that it is arguably the single most important tool for obtaining detailed information on chemical systems at the molecular level. It is certainly recognized as a valuable technique for analysing food products (Belton, 1995).

Nuclei of certain isotopes (e.g. ^{1}H,^{13}C,^{19}F) possess intrinsic angular momentum or spin (they are said to be spin-active and have the ability to resonate). A nucleus which is spinning also possesses an associated magnetic moment μ. When placed in a strong magnetic field, these nuclei can absorb electromagnetic radiation in the radio frequency range. An NMR spectrometer picks up and displays the precise frequency at which resonance takes place (Williams, 1986).

Carbon forms the backbone of all organic molecules, and Carbon-13 (^{13}C) is the only magnetic carbon isotope (Wehrli, Marchand and Wehrli, 1988). From the point of view of the organic chemist, it is fortunate that such an isotope exists, forming some 1.1% by weight of naturally occurring carbon (Stryer, 1981).

In the field of olive oils, NMR has been applied before (Anon., 1994; Bianchi *et al.*, 1993, 1994a; Gussoni *et al.*, 1993; IFR, 1994; Zamora, Navarro and Hidalgo, 1994), demonstrating the potential of this technique for analysing olive oils. Brekke *et al.* (1990) and Kvalheim *et al.* (1985) also show that a combination of NMR and PCA can be used to distinguish between different North Sea crude oils.

Theory. The magnitude of the spin (angular momentum) of the nucleus is $h[I(I+1)]^{1/2}$, where I is the nuclear spin quantum number and h is the reduced Planck's constant $h/2\pi$. I may have only integral or half-integral values ($0, \frac{1}{2}, 1, 1\frac{1}{2}), \ldots, 6$, in units of $h/2\pi$) (Friebolin, 1993), the value being dependent upon the isotope. For ^{13}C and ^{1}H, $I = \frac{1}{2}$. When $I = 0$, the nucleus has no angular momentum. Transitions between nuclear spin energy levels give rise to the resonance phenomenon of NMR.

In a nucleus containing an even number of both protons and neutrons, $I = 0$. This includes the common atoms ^{12}C and ^{16}O. This leads to considerable simplification of the spectra of organic molecules. The reason for this is that nucleons with opposite spin can pair (though neutrons can pair only with neutrons, and protons with protons), just as electrons pair. If the numbers of neutrons and/or protons is odd, then the spin is non-zero, though the actual value depends upon orbital-type internucleon interactions (Akitt, 1983).

Because it is difficult to know to sufficient accuracy the value of the magnetic field applied (Harris, 1986), a standard of known resonance frequency is usually used. The most convenient standard for ^{13}C and proton NMR is tetramethylsilane (TMS) because it contains four equivalent carbon atoms, and the resultant strong signal means that only a small amount (1%–5%) need be added; it gives rise to sharp signals, is chemically inert and is soluble in most organic materials (Kemp, 1986). The TMS peak is taken as 0 in the δ scale and increases in a downfield direction.

The resonance frequency ν_0 of an isolated nucleus is:

$$\nu_0 = \frac{\gamma B_0}{2\pi} \tag{10.1}$$

where B_0 is the strength of the steady magnetic field and γ is a constant (the gyromagnetic ratio) (Stryer, 1981).

Depending on the chemical environment (bondings, etc.) the precise resonance frequency of any one ^{13}C atom will be shifted by a few parts per million (ppm) from that of the standard. This is the chemical shift. The formula for calculating the chemical shift δ, then, is

$$\delta = \frac{\nu_{\text{sample}} - \nu_{\text{ref}}}{\nu_{\text{ref}}} \cdot 10^6 = \frac{\Delta\nu_{\text{sample}}(\text{Hz})}{\nu_{\text{ref}}(\text{MHz})} \tag{10.2}$$

(adapted from Brevard and Grainger, 1981; Friebolin, 1993).

The nucleus of an atom is surrounded by electrons. The electron cloud shields the nucleus to some extent from the applied magnetic field. The amount of shielding depends on the nature of the electrons around the nucleus and is given by the formula

$$B_{\text{eff}} = B_0 - \sigma B_0 = (1 - \sigma)B_0 \tag{10.3}$$

where σ is the shielding constant, B_{eff} is the magnetic field at the nucleus and B_0 is the applied magnetic field, thus altering the chemical shift. Then, the chemical shift of a given atom varies as a function of its chemical environment, thus allowing the description of the chemical environment from the chemical shift.

For most studies, the same deuterated solvents are employed in both ^{13}C and proton NMR. A major advantage of ^{13}C spectroscopy over proton NMR is the larger chemical shift range that for most organic substances is ~200ppm in comparison with ~10ppm for proton NMR.

The chemical shift of carbon depends on the hybridization of carbon and its structural environment, and from high to low field is sp^3, sp and sp^2. This trend is observed in the typical olive oil ^{13}C spectrum. The chemical shift of methyl and methylene sp^3 carbon is in the range 10–90 ppm; the range of olefinic sp^2 carbons is 127–130 ppm; the carbonyl carbons appear in the range 170–175 ppm. The glycerol carbon shifts are found in the 60–70 ppm region (Vlahov, 1996). The high-resolution ^{13}C NMR spectra of olive oil, usually dissolved in $CDCl_3$, are readily obtained running 250–300 scans. Spectra consist of 40–45 signals. Most of the signals are assigned according to the chemical shift of standard compounds and literature data. A typical spectrum for olive oil is shown in Fig. 10.1. The principal assignments are given in Fig. 10.2.

10.3.2 Pyrolysis mass spectrometry

Pyrolysis mass spectrometry (PyMS) is a technique that (via Curie-point pyrolysis) thermally degrades a sample of interest at a known temperature in an inert atmosphere or a vacuum. It causes molecules to cleave at their weakest points to produce smaller, volatile fragments called pyrolysate (Irwin, 1982). The mass spectrometer can then be used to separate the

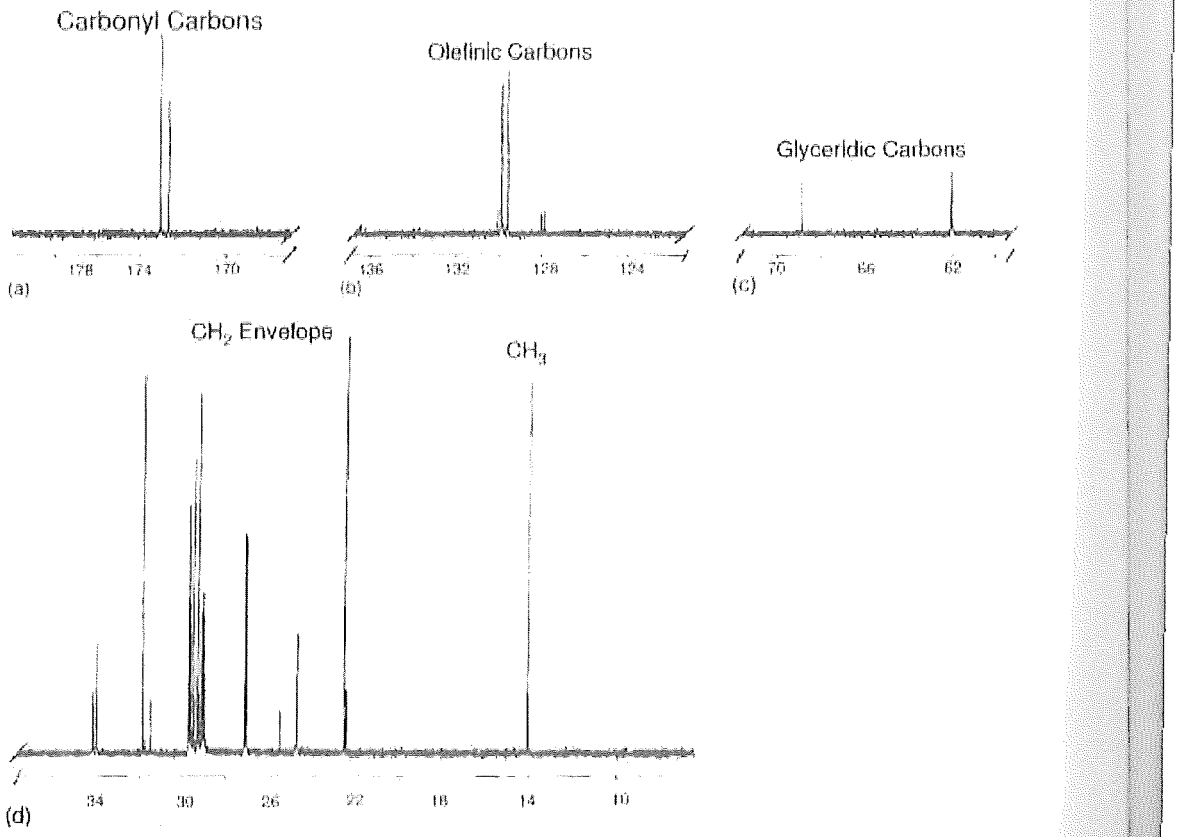

Figure 10.1 ^{13}C NMR spectrum of extra virgin olive oil obtained from the Dritta cultivar: (a) carbonyl region; (b) olefinic region; (c) glyceridic region; (d) CH_2 envelope and methyl region.

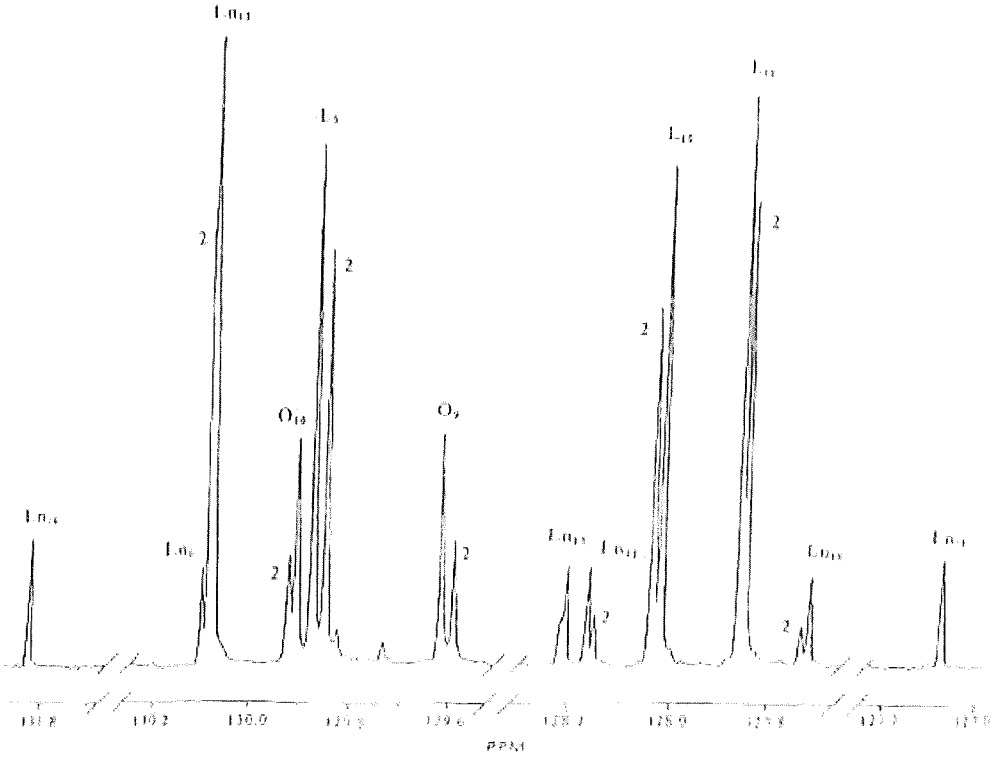

Figure 10.2 ^{13}C NMR spectrum of a seed oil: the expanded olefinic region. Ln = linolenic acid; L = linoleic acid; O = oleic acid.

components of the pyrolysate on the basis of their mass-to-charge ratio (m/z) to produce a pyrolysis mass spectrum (Meuzelaar, Haverkamp and Hileman, 1982), which can then be used as a 'chemical profile' or fingerprint of the sample material analysed. The spectrum obtained may then be used as an input to some analysis tool such as Neural Nets, PCA, PCR, PLS or other statistical or computational techniques (Michie, Spiegelhalter and Taylor, 1994; Weiss and Kulikowski, 1991), allowing the sample to be categorized. Much work has been done in this field already, including the use of bacterial strains with both standard back-propagation ANNs (Goodacre, Neal and Kell, 1994) and Kohonen ANNs (Goodacre *et al.*, 1996), complex binary and tertiary mixtures (Goodacre, Neal and Kell, 1994), casamino acids and glycogen (Goodacre, 1993; Neal, 1994), microbial fermentations (Goodacre, 1994b; Goodacre and Kell, 1996; Goodacre *et al.*, 1995) and the adulteration of foodstuffs such as olive oils (Goodacre, Kell and Bianchi, 1992, 1993) and orange juice (Goodacre, Hammond and Kell, 1997). It has also been successfully applied to other discrimination problems, by Aylott *et al.* (1994) with PCA and CVA for classification of some of the 500 brands of Scotch whisky, and by Kajioka and Tang (1984) to distinguish between six species of the family *Legionellaceae*. The latter led to the identification of a new strain of the bacterium.

In Curie-point pyrolysis the material is placed on an iron–nickel alloy foil which is heated to the Curie point of the foil (this is 530°C for 50:50 Fe–Ni foils). For a given type of foil, the Curie-point temperature is constant, therefore this type of pyrolysis is very reproducible. The foil holding the sample is rapidly heated to its Curie point by passing a radio-frequency current for 3 s (in the case of the Horizon 200-X instrument used at Aberystwyth) through a coil surrounding the foil. The foil takes around 0.5 s to reach this point; at this temperature the material on the foil is thermally

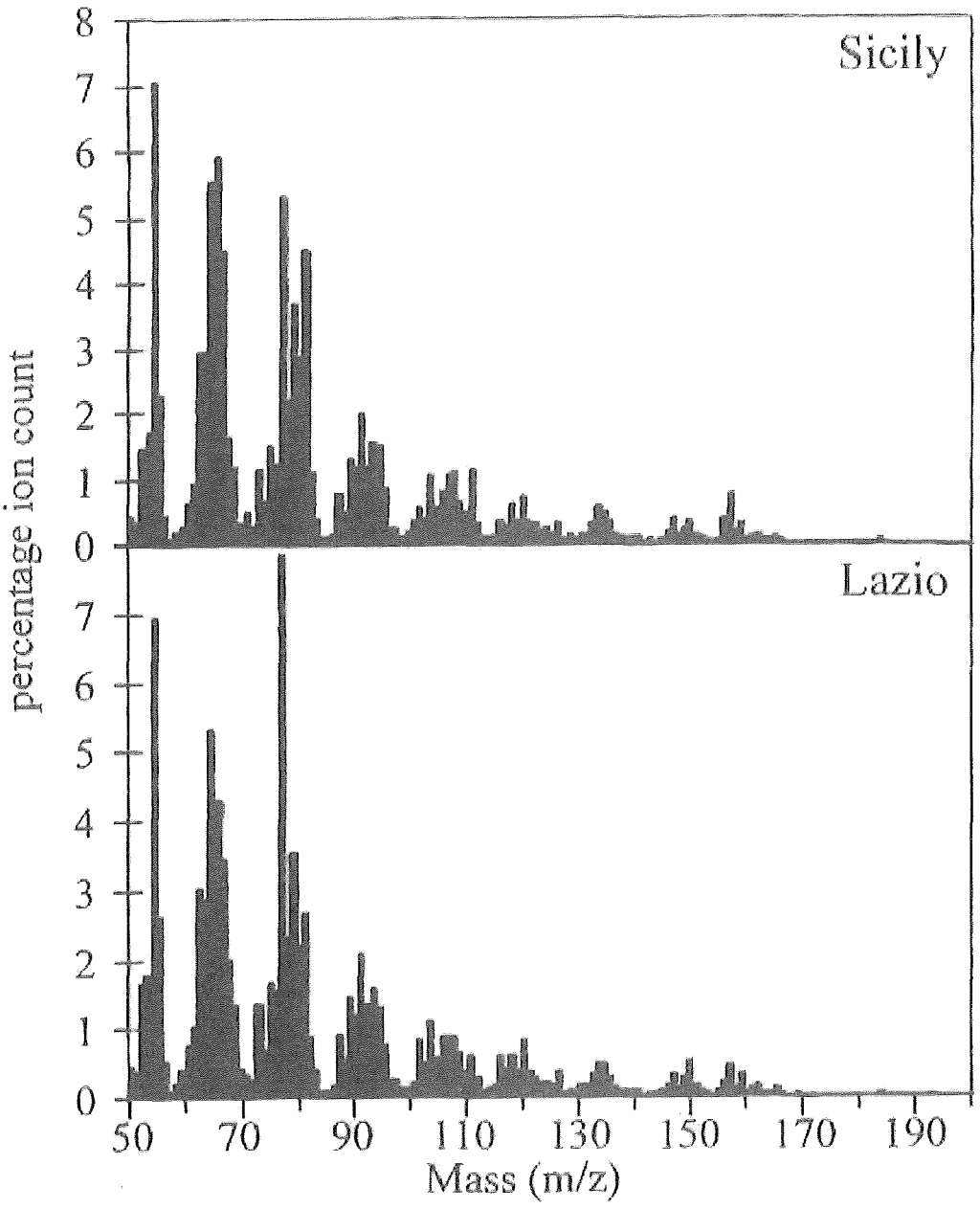

Figure 10.3 Two typical pyrolysis mass spectrometry spectra of extra virgin olive oils from Sicily and Lazio.

degraded into its pyrolysate. The pyrolysate is then separated into its components by the mass spectrometer part of the equipment, producing a pyrolysis mass spectrum (Fig. 10.3). Goodacre (1994a) and Goodacre and Kell (1996) give a very thorough analysis of the PyMS technique and equipment, along with a brief history. More detail is given in the books by Irwin (1982) and Meuzelaar, Haverkamp and Hileman (1982).

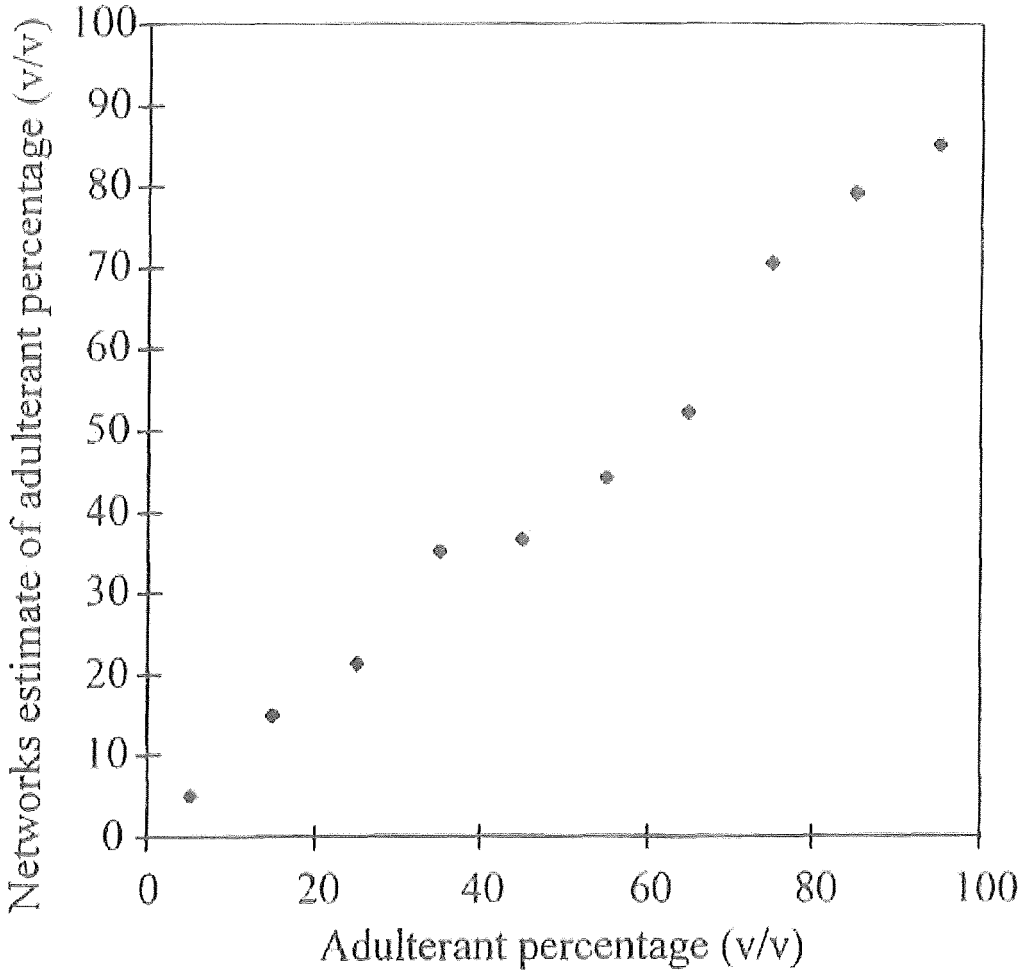

Figure 10.4 A typical prediction curve for adulteration of extra virgin olive oil. Adulteration series using from 0%–100% adulterant were prepared using husk oil, with samples every 5% [a total of 21 samples (in triplicate)]. 1.5 μl of sample was analysed by pyrolysis mass spectrometry and the spectra collected over the *m/z* range 51–200. Data were normalized as a percentage of total ion count to remove the most direct influence of sample size *per se*. Normalized spectra were sorted according to the level of adulteration, with triplicates being kept together. Half of these samples (every other one) were used to train a neural net, the remaining half being used as a test set. The network architecture is composed of an input layer of 150 nodes, one node for each mass within the spectrum of each oil, a hidden layer of 8 nodes and a single output node. Headroom was maintained at ±10% for each input. A sigmoid (logistic) transfer function was used, as the data relationship is suspected to be non-linear. The supervised learning algorithm was standard back propagation (that updates weights as each pattern is presented) with a learning rate of 0.2 and a momentum of 0.8. Patterns were presented randomly with noise of 0.05 added. All data going into and out of the net were scaled automatically (each column individually) by the software.

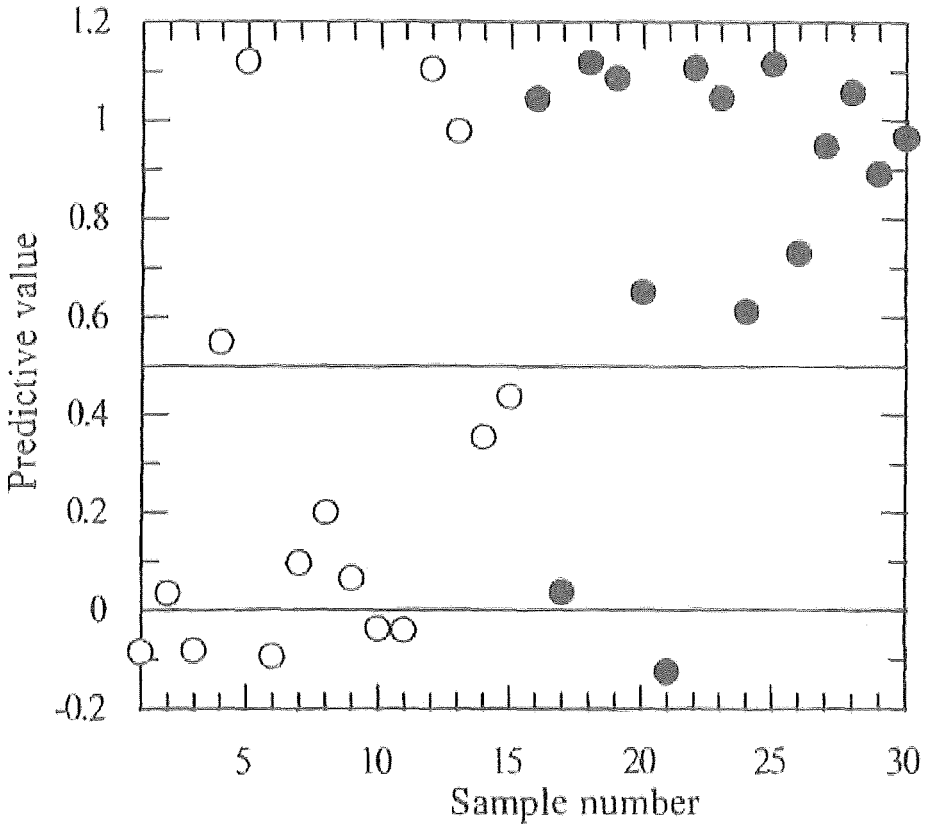

Figure 10.5 Neural net prediction of virgin olive oil adulterated at the 1% level by high oleic sunflower oil. (a) Prediction with variable selection, 60 samples, 30 adulterated by 1% vol./vol. high oleic sunflower oil and 30 unadulterated, were run through the pyrolysis mass spectrometer and the spectra collected. Data were normalized to the total ion count of the samples and principal components analysis (PCA) performed. The first factor from the PCA analysis was then used to select the variables (58) which were fed into an artificial neural net using a stochastic back-propagation algorithm; 15 adulterated and 15 unadulterated samples, encoded, respectively, as 1 and 0, were used to train the net (58 nodes on the input layer, 4 nodes in the hidden layer and a single output node), and 30 samples were used to test the model. The results for the unseen test set are shown, and were obtained after 29 000 epochs with an error on the training set of 0.01. The horizontal line at 0.5 is an aid to interpretation. All samples with a value close to 1 (i.e. those above the line) are the predicted adulterated samples (actual adulterated samples being filled circles); all those with a value close to 0 (i.e. those below the line) are the predicted unadulterated samples (actual unadulterated samples being open circles); 11 out of 15 (73%) unadulterated samples and 13 out of 15 (87%) adulterated samples being predicted correctly.

One of the great advantages of PyMS is that it is relatively cheap compared with other methods of analysis. Many samples can be run through the PyMS machine in a short time (typically less than 2 min each) at a cost of less than £1 sterling per sample.

The atomic masses up to 50 are discarded since they include very common compounds such as methane (CH_4, 16 amu), ammonia (NH_3, 17 amu), water (H_2O, 18 amu), methanol (CH_3OH, 32 amu) and hydrogen sulphide (H_2S, 34 amu), which are likely to be present in large quantities in any pyrolysate. Fragments with an *m/z* ratio of over 200 are rarely analytically important for bacterial discrimination so these are also discarded (Good-

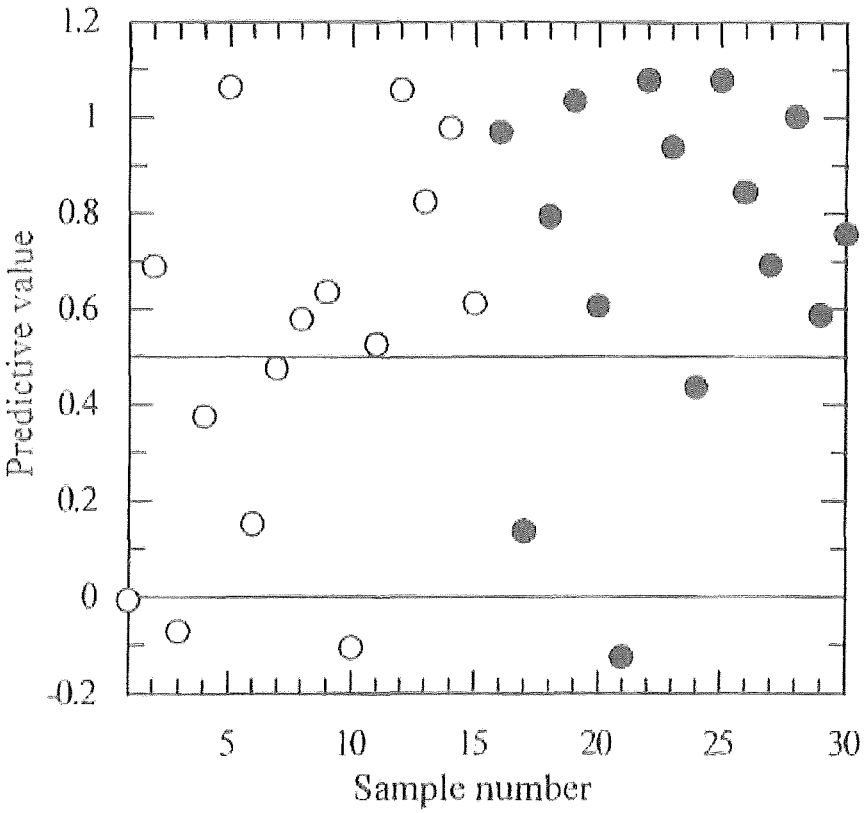

Figure 10.5(b) Prediction with no variable selection. The experimental rationale was as for part (a), except that the artificial neural network was run on all 150 variables, that is, no variable selection was used. Optimization in this case resulted in having 8 nodes in the hidden layer; the outputs were obtained after 10 000 epochs with an error on the calibration set of 0.01; 6 out of 15 (40%) unadulterated samples and 12 out of 15 (80%) adulterated samples were predicted correctly. It may be seen that very poor separation was achieved – the two data bases do not line up on the predictive value of 0 or 1 but form a loose cloud in the middle area around 0.5, indicating that the net was unable to separate the two groups clearly.

acre, 1994a). It is assumed that these fragments are also unimportant for other organic compounds, although no reference has been found to support or contradict this belief.

The exploitation of PyMS for assessing the adulteration of olive oil (Goodacre, Kell and Bianchi, 1992, 1993) has been continued by Bianchi, Giansante and Lazzari (1996) and by Salter *et al.* (1997) who have shown that the principle is generally and quantitatively applicable to a wide range of potential adulterants, including olive oil, hazelnut oil, husk oil, corn oil, peanut oil, maize oil, soya oil, sunflower oil, high oleic acid sunflower oil and grape stone, along with rapeseed/soya and palm/peanut/sunflower oil mixes (Fig. 10.4). Predictions may be improved upon in some cases by the use of variable selection, and in many cases adulterants may be detected when present at less than 2% (vol./vol. Figs 10.4–10.6). Bianchi *et al.* (1994b) suggest that adulteration by lower grades of olive oil may be detected by the presence of long chain esters.

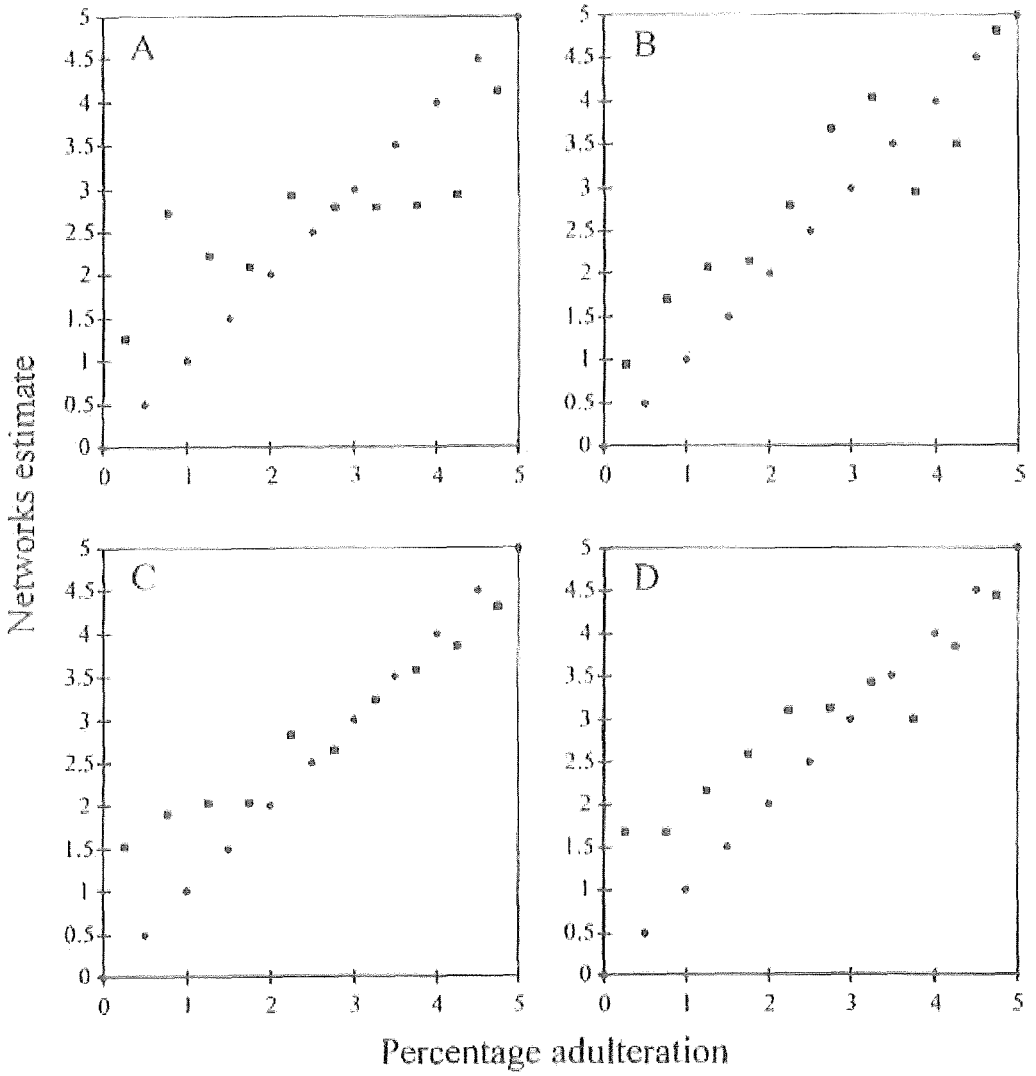

Figure 10.6 Adulteration series from 0%–100% were prepared using husk oil, with samples every 5% [a total of 21 samples (in triplicate)]; 1.5 µl of sample was analysed by pyrolysis mass spectrometry in triplicate over the *m/z* range 51–200. Data were normalized as a percentage of total ion count to remove the most direct influence of sample size *per se*. Normalized spectra were sorted according to level of adulteration, with triplicates being kept together. Half of these samples (every other one) were used to train a neural net, the other half being used as a test set. The network architecture is composed of an input layer corresponding to the number of variables used, one node for each mass used within the spectrum of each oil. The optimum number of nodes within the hidden layer varied, whilst a single output was used that represented the network's estimation of the adulteration as a numeric value. Headroom was maintained at ±10% (for each input) for all networks. A sigmoid (logistic) transfer function was used (unless stated otherwise) as the data relationship was suspected to be non-linear. The supervised learning algorithm was standard back propagation (that updates weights as each pattern is presented) with a learning rate of 0.2 and a momentum rate of 0.8. Patterns were presented randomly with noise of 0.05 added. All data going into and out of the net were scaled automatically (each column individually) by the software. (A) Prediction using no variable selection. All 150 masses are used, with 8 nodes in the hidden layer. The network was run over 3622 epochs to a training error of 0.025 and a test set RMSEP of 1.25. After this point overtraining occurred. (B) Mutual information was used as a method of variable selection. All 150 variables were assessed using mutual information (Battiti, 1994). The three variables that

10.4 Multivariate methods

For the purposes of this chapter, it is convenient to consider the complete set of measured quantities (e.g. the different m/z ratios of PyMS ion counts) for a given sample as being a single n-dimensional position vector, $x^{\mathrm{T}} = (x_1\, x_2 \ldots x_n)$ with x_i being the measured value of quantity i. A set of m such vectors is represented as a matrix, $\mathbf{X}$, with rows, x_i^{T}. The value stored at row i, column j of $\mathbf{X}$ is denoted X_{ij}.

A scores matrix is one in which the columns contain the values of transformed variables. An example of a scores matrix is that containing principal component values. A loadings, or weightings, matrix is one by which a data matrix may be multiplied to obtain a scores matrix.

10.4.1 Principal components analysis

PCA was devised by Hotelling (1933) as an aid to the interpretation of academic test results. The reasoning behind PCA is that a set of test results often shows correlations between the different types of test. These results are presumably reflecting some 'mental factor'. Hotelling derived PCA as a method for extracting a set of uncorrelated variables as linear combinations of the original variables. The 'mental factors' are contained in the new data set. Hotelling named the new variables the 'components' and suggested that, if the components were arranged in decreasing order of variance, then the first few, the 'principal components', should be those which characterized the 'mental factors'. Later components would reflect less important effects, with some possibly reflecting measurement errors. Tests were suggested which would allow later components to be discarded as noise, so leaving a lower-dimensional data set, with the most easily interpreted effects appearing first. Figure 10.7 demonstrates how this can be useful. Even a three-dimensional cloud of points can be difficult to interpret if it is viewed from a poor vantage point. By rotating the cloud in a manner which is in some way

---

were clearly the most important were fed into an artificial neural network (ANN) having 4 nodes in the hidden layer. After some 363 483 epochs a training error of 0.082 and a test RMSEP of 1.1 was recorded. (C) The w value (text, Section 10.4.11) was used as a method of variable selection. All masses were ranked according to the w value and the best 50 selected as an input for an ANN that had 8 nodes in the hidden layer. After some 9357 epochs a training error of 0.05 and test RMSEP of 0.99 was achieved. (D) Principal components analysis was used as the method for variable selection. All masses were ranked according to the score of the first principal component, the best 50 variables being selected as the input for an ANN. The ANN consisted of an input layer of 50 nodes, a hidden layer of 8 nodes and a single output node. In this case a linear transfer function was found to be optimal, perhaps reflecting the linear nature of the variable selection. After some 233 epochs a training error of 0.1 and a test RMSEP of 1.14 for the illustrated prediction was achieved.

optimal, one can obtain a much clearer view of any structure within the data. PCA helps to achieve this aim.

Hotelling's derivation of PCA was made by assuming that the original n variables in $\mathbf{X}$ were normally distributed. The overall distribution of the n-dimensional vector set of samples in $\mathbf{X}$ is therefore a multivariate normal distribution and will appear ellipsoidal if plotted as a scatter plot in n-dimensional space. The method used was to derive a formula for the optimally fitting ellipsoid, using least squares methods for the optimization.

An alternative derivation may be used which gives the same results as Hotelling's but which does not require the assumption of normality of the variables in $\mathbf{X}$, as follows.

Suppose one has a cluster of n dimensional points, x_i. The requirement is to find a set of axes for the cluster such that when the x_i are transformed linearly to the new axes the transformed variables are uncorrelated, that is one needs to find n orthogonal unit vectors, p_i such that the $\mathbf{X}\mathbf{p}_i$ and $\mathbf{X}\mathbf{p}_j$ are uncorrelated for $i \neq j$. Assuming that the variables in $\mathbf{X}$ have been centred to have zero mean, and scaled to unit variance, one has $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ as the correlation matrix of $\mathbf{X}$.

One must therefore find an orthonormal matrix, $\mathbf{P}$, such that $\mathbf{S}' = [\mathbf{X}\mathbf{P}]^{\mathrm{T}}[\mathbf{X}\mathbf{P}]$ has non-zero values on the diagonal only, that is:

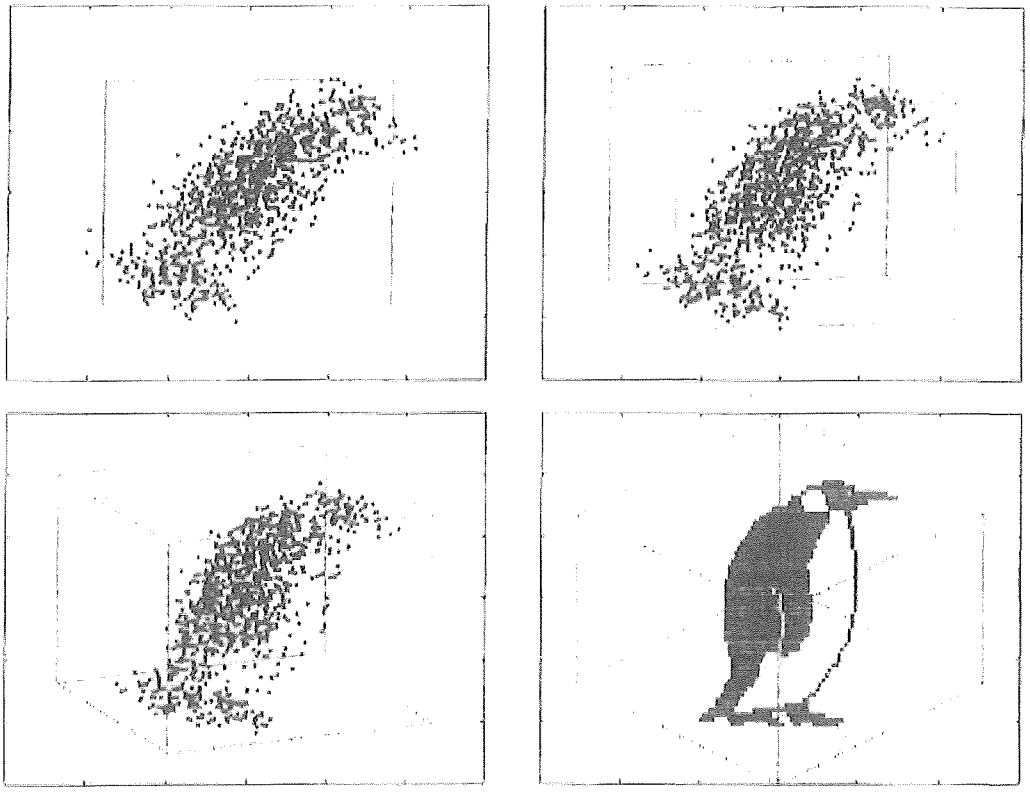

Figure 10.7 Example of how principal components analysis can help to extract useful information

$$\mathbf{S}' = \begin{bmatrix} a_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & a_2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & & \ddots & & \\ 0 & 0 & 0 & \dots & 0 & a_n \end{bmatrix} = \text{diag}(a_1, \dots, a_n), \tag{10.4}$$

$$\mathbf{S}' = [\mathbf{XP}]^{\mathrm{T}}[\mathbf{XP}] = \mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XP} = \mathbf{P}^{\mathrm{T}}\mathbf{SP} = \text{diag}(a_1, \dots, a_n). \tag{10.5}$$

S is a symmetric matrix [since correlation (X_i, X_j) = correlation (X_j, X_i)] of real values. Therefore, from the theory of eigenanalysis, it follows that P is the matrix whose columns are the orthogonal eigenvectors of S and that $a_1, \dots, a_n$ are the corresponding eigenvalues. For a more detailed discussion of eigensystem theory see, for example, Morris (1982).

The problem of finding the components of X is therefore that of deriving the eigensystem of S, the correlation matrix of X.

Next, consider a component, T_i. The sample values of T_i, $\mathbf{t}_i$ are derived from X by $\mathbf{t}_i = \mathbf{X}\mathbf{p}_i$, where $\mathbf{p}_i$ is an eigenvector. Let λ_i be the corresponding eigenvalue. Then

$$\begin{aligned} \text{var}(T_i) &= \mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i \\ &= (\mathbf{X}\mathbf{p}_i)^{\mathrm{T}}(\mathbf{X}\mathbf{p}_i) \\ &= \mathbf{p}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p}_i \\ &= \mathbf{p}_i^{\mathrm{T}}\mathbf{S}\mathbf{p}_i \\ &= \mathbf{p}_i^{\mathrm{T}}\lambda_i\mathbf{p}_i \\ &= \lambda_i\mathbf{p}_i^{\mathrm{T}}\mathbf{p}_i \\ &= \lambda_i \end{aligned} \tag{10.6}$$

So, to extract the components in decreasing order of variance, one need only extract the eigenvectors in decreasing order of eigenvalue.

Finally, consider

$$\begin{aligned} \sum_{i=1}^{n} \text{var}(T_i) &= \sum_{i=1}^{n} \text{var}\left(\sum_{j=1}^{n} P_{ji}X_i\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} P_{ji}P_{ki}S_{jk} \end{aligned} \tag{10.7}$$

However, P is orthonormal, so its columns (and rows) must be linearly independent and the dot product of any two columns (rows) must be 0 if they are not the same, or 1 if they are. That is,

$$\sum_{i=1}^{n} P_{ji}P_{ki} = \delta_{jk}, \quad \text{where } \delta_{jk} = 1 \text{ if} j = k, 0 \text{ otherwise;} \tag{10.8}$$

$$\sum_{i=1}^{n} \mathrm{var}(T_i) = \sum_{k=1}^{n} \sum_{j=1}^{n} S_{jk}\delta_{jk}$$
$$= \sum_{i=1}^{n} S_{ii} \qquad (10.9)$$
$$= \sum_{i=1}^{n} \mathrm{var}(X_i) \quad .$$

So, the proportion of the variance in $\mathbf{X}$ explained by component i is therefore $\lambda_i \left[\sum_{i=1}^{n} \mathrm{var}(X_i)\right]^{-1}$, or λ_i/n for X_i scaled to variance 1.

By using the above, one may then extract principal components (by calculating eigenvectors) in decreasing order of variance (by observing the corresponding eigenvalues) and know the proportion of the variance in $\mathbf{X}$ explained by any component (by observing the eigenvalue size).

Extracting principal components. The mathematical literature details a number of methods for determining the eigensystem of a matrix. In the case of principal component extraction, the matrix $X^{\mathrm{T}}X$ is square and symmetric, as

$$(X^{\mathrm{T}}X)_{ij} = \sum_{k=1}^{n} X_{ik}^{\mathrm{T}} X_{kj} = \sum_{k=1}^{n} X_{jk}^{\mathrm{T}} X_{ki} = (X^{\mathrm{T}}X)_{ji}, \qquad (10.10)$$

which makes a number of methods available.

Jacobi's method. Jacobi's method allows all n eigenvectors and eigenvalues to be extracted at the same time. It works by applying a sequence of transformations to reduce $X^{\mathrm{T}}X$ to diagonal form. The product of the transformation matrices is a matrix having the eigenvectors as columns, and the diagonal elements of the reduced matrix are the eigenvalues. The off-diagonal elements of the matrix are zeroed by applying a plane rotation matrix to each off-diagonal position (i, j). A full description of this method may be found in Gourlay and Watson (1973).

Unfortunately, Jacobi's method is computationally intensive, requiring of the order of $6n^3$ arithmetic operations to calculate the eigensystem.

The power method. The power method may be used to determine the largest eigenvalue and its corresponding eigenvector. The method uses the iterative scheme $z_k = Az_{k-1}, k = 1, 2, 3, \ldots$ where A is the $n \times n$ matrix and z_0 is initialized to a first estimate of the eigenvector. Then z_k tends to the eigenvector and $(z_k)_i/(z_{k-1})_i$ to the eigenvalue as $k \to \infty$ [where $(z)_i$ indicates element i of the vector, z]. Each iteration uses order n^2 operations, with the convergence rate being proportional to the ratio of the first two eigenvalues, ratios close to unity giving slow convergence. The method fails for largest eigenvalues of equal modulus, with the eigenvector oscillating in sign.

Subsequent eigenvalues can be found by transforming the matrix, A, such that the remaining eigensystem is retained, but the influence of z is removed. This process is known as deflation and may be achieved in a number of ways. Again, see Gourlay and Watson (1973) for further details.

Simultaneous iteration. Simultaneous iteration (Clint and Jennings, 1970) allows the simultaneous extraction of the k largest eigenvalues and their corresponding eigenvectors. The method works by reducing the problem to that of calculating the eigensystem of a reduced matrix of size $k \times k$. This eigensystem is used to improve an estimate of the eigenvectors of the larger matrix. The process is iterated until the required accuracy is obtained. If $k \ll n$, then a great saving in time may be obtained.

The non-linear iterative partial least squares method. The previous methods required the covariance matrix, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ to be calculated prior to extraction of the eigensystem. The non-linear iterative partial least squares (NIPALS) algorithm (Wold, 1966, 1975) sidesteps this requirement by noting that component weights may be derived from X and the corresponding components by linear regression. The components are, by definition, obtained from X and the weights vector. Thus, an iterative process can be defined:

- make an initial estimate for the scores vector, t_0;
- repeat the following:
  - Regress X on t_{n-1} to obtain an estimate of the loadings vector, p_n:

$$p_n^{\mathrm{T}} = (t_{n-1}^{\mathrm{T}} t_{n-1})^{-1} t_{n-1}^{\mathrm{T}} \mathbf{X};$$

  - normalize p_n to have unit length;
  - project the X matrix to form a new scores vector, $t = Xp_n$;

  until p_n converges.

This process generates the first principal component loadings and scores for X. Their effect may then be removed from X by projecting t back into the coordinate system used for X and subtracting:

$$\mathbf{X}' = \mathbf{X} - tp^{\mathrm{T}}$$

The iteration can be repeated on $\mathbf{X}'$ to obtain the second and subsequent components. This algorithm is of order n^2 for each component extracted, depending on the convergence rate.

10.4.2 Predictive models

The methods discussed thus far allow the exploration of a multidimensional data set, and may allow clusters to be separated, but they do not result in predictive models.

To build a system capable of predicting characteristics of interest from measured values, we need to form a model which relates the measurements to the physical effect of interest. This model can then be applied to future measurements to predict the effect. In mathematical terms, the objective is to find the multidimensional function, f, such that $y = f(x)$, where x is a new multivariate reading and y is the effect to be predicted. For the present, we shall assume that f is a linear function, for simplicity. Artificial neural networks (ANNs), a non-linear method, will be discussed later.

In the linear case, then, we have y being a simple weighted sum of the variables x: $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$. In vector notation, this may be expressed $y = x \cdot b + b_0$, b_0 being a constant offset. Finally, if we augment x with an entry, $x_0 = 1$, then b may be augmented with b_0 to give $y = x \cdot b$.

If we wish to predict a number of values of y, we may express this as a single matrix multiplication, $y = \mathbf{X}b$, where the x vectors have now been placed as rows of $\mathbf{X}$.

For model formation, then, the problem is that of choosing the vector, b to give good modelling of the values of y. In the following discussions, let $\mathbf{X}_c$ and y_c be a set of measured calibration values for the variables x and y. A number of methods for choosing an estimate for b, $\hat{b}$ from $\mathbf{X}_c$ and y_c will now be considered.

10.4.3 Multiple linear regression

Multiple linear regression (MLR) provides the most 'obvious' course of action for estimating b, that of minimizing errors in the model. A relationship, $y_c = \mathbf{X}_c b + e$ is assumed, with e containing all unmodelled variation in y_c. For an arbitrary value of b, e will be made up of two parts: the systematic error due to poor choice of b, and the random error of measurement in $\mathbf{X}_c$ and y_c. For MLR, the random measurement errors are assumed to have zero mean (i.e. the true value) and equal variance. Under this assumption, it is obvious that the best model we can choose is that which reduces the length of e to a minimum. In practice, it is simpler to minimize the squared length of e, or $(y_c - \mathbf{X}_c b)^{\mathrm{T}}(y_c - \mathbf{X}_c b)$. This method is known as least squares regression and is familiar in the single dimensional case as finding the 'line of best fit'. The least squares estimator for b is given by $\hat{b} = (\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c)^{-1}\mathbf{X}_c^{\mathrm{T}}y_c$ (for a derivation, see Wonnacott and Wonnacott, 1981).

Note that this formula involves the inversion of $\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c$, the covariance matrix of $\mathbf{X}$. If this inverse does not exist, then $\hat{b}$ cannot be calculated by means of MLR. Singularity of $\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c$ corresponds to the linear dependence of a subset of the $\mathbf{X}$ variables. So, if $\mathbf{X}_c$ contains such relationships, MLR cannot be applied. In practice, it is rare to have exact linear dependence, since the measurement errors will tend to preclude this. However, near linearity will tend to make the inverse numerically unstable and subject to large errors, so much the same effect occurs. Another formulation of the

MLR equation can be obtained by expressing it in terms of the matrix of eigenvectors, P, and the diagonal matrix of eigenvalues, diag(λ_i), as follows:

$$\mathbf{P}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{P} = \mathrm{diag}(\lambda_i). \tag{10.11}$$

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \mathbf{P}\,\mathrm{diag}\left(\frac{1}{\lambda_i}\right)\mathbf{P}^{\mathrm{T}} \tag{10.12}$$

$$\hat{b} = \mathbf{P}\,\mathrm{diag}\left(\frac{1}{\lambda_i}\right)\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}y. \tag{10.13}$$

If X contains (almost) collinear variables, then λ_i will be (close to) zero for some i, giving an eigenvalue matrix containing large values which are susceptible to small alterations in the calibration data. In our application, collinearity of the X variables is to be expected, since harmonic responses at similar frequencies and amplitudes are likely to be similar.

MLR may be modified to reduce the problems of collinearity through the use of variable selection. Rather than calculating the model relating y_i to $\mathbf{X}_i$, we relate y_i to $\mathbf{X}'_i$, a matrix containing a subset of the original X variables. The subset is chosen to eliminate collinearity and to improve the model by removing irrelevant variables (a discussion of overfitting is given in Section 10.4.7). Three common methods are used for subset selection: Forward selection (FS), backward elimination (BE) and stepwise multiple linear regression (SMLR).

Forward selection. Here the model is built up by adding variables in X, one at a time, to the model. The variables are added by choosing those which improve the model most at each step according to some statistical test. Variables which are collinear with a variable already added will contribute little to the quality of a subsequent model. Likewise irrelevant variables contribute little. Such variables will not, therefore, be added. Variables are added until a 'stopping' criterion is reached, for example when the improvement afforded by addition of the next variable drops below a threshold value.

Backward elimination. Here, all the variables are used to form an initial model. Variables are then selected for deletion, once again by using a statistical test and threshold to determine when to stop. In the case of BE, exactly collinear variables must be eliminated before the initial model is formed.

Stepwise multiple linear regression. This is a modified form of forward selection. The model starts out including only one variable, and more variables are subsequently added. But at each stage a BE-style test is also applied. If a variable is added, but becomes less important as a result of subsequent additions, SMLR will allow its removal from the model.

A similar method, 'best of all subsets', regression considers models formed from all possible subsets of the X variables. For each subset the linear regression is formed and tested for its quality. The best of these is selected as the final regression set. Although this method is guaranteed to do at least as well as the stepwise methods, the problem of combinatorial explosion rears its head: for n variables in X the number of possible subsets is 2^n. Even for moderate values of n, the number of regressions required can be too large to be practical. Near-infra-red spectroscopic methods, for example, typically have hundreds of variables. Heuristics may be used to select a feasible subset of the original variables on which to run 'best of all subsets' regression, but one is still left with the problem of selecting this subset.

10.4.4 Ridge regression

Ridge regression (RR) (Hoerl and Kennard, 1970a, b) is an attempt to solve the problems of collinearity in MLR by modifying the regression equation. Rather than having $\boldsymbol{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$, as for MLR, RR forms the family of regressors $\boldsymbol{b}^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + k)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$, where $0 \leq k \leq 1$. In eigenvector and eigenvalue form, this is stated as $\hat{\boldsymbol{b}}^* = \mathbf{P}\,\mathrm{diag}\,(1/(\lambda_i + k))\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$, where the λ_i and $\mathbf{P}$ are eigenvalues and eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. By varying k we can prevent division by small values and so make the model more stable against perturbations of the calibration set. The new regressor, $\boldsymbol{b}^*$ will be biased, tending to centre on a value removed from the optimal $\boldsymbol{b}$, but will have a smaller mean square error. For a well-chosen value of k, we can make $\boldsymbol{b}^*$ lie close to the true $\boldsymbol{b}$ and have better stability.

RR introduces the concept of the 'ridge trace'. This is a graph showing the weightings, b_i^* as k varies from 0 to 1. The ridge trace highlights those variables which are unstable, as these will show great variation over the range of the trace. These variables are then candidates for deletion. The ridge trace can be used to guide the choice of k. A value can be chosen where the individual traces have 'flattened' out.

10.4.5 Principal components regression

Principal component regression (PCR) combines PCA and MLR to give another method for removing the effects of collinearity (Massy, 1965). PCA generates components which are orthogonal and in decreasing order of eigenvalue (variance). Hence, to remove variables with small eigenvalues, one needs merely to select a threshold eigenvalue and discard all components below this threshold. This corresponds to selecting the first n components. A simple MLR model is then formed on the remaining components.

For modelling, then, we have $\boldsymbol{b} = (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\boldsymbol{y}$, where $\mathbf{T} = \mathbf{XP}$, $\mathbf{P}$ being the matrix of retained component weightings. For prediction, we must first transform the new X data by using P; $\boldsymbol{b}$ is then used to make

the prediction. So, $\hat{y} = (\mathbf{XP})\hat{b}$. If we combine $\mathbf{P}$ and $\hat{b}$ to give a new $\hat{b}$, $\hat{b}_{\mathrm{PCR}}$, we may predict values by using the standard prediction $\hat{y} = \mathbf{X}\hat{b}_{\mathrm{PCR}}$, with $\hat{b}_{\mathrm{PCR}} = \mathbf{P}\hat{b}$.

If no components are excluded, then we have

$$\begin{aligned}
\hat{b}_{\mathrm{PCR}} &= \mathbf{P}[(\mathbf{XP})^{\mathrm{T}}(\mathbf{XP})]^{-1}(\mathbf{XP})^{\mathrm{T}}y \\
&= \mathbf{P}(\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XP})^{-1}\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}y \\
&= \mathbf{PP}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{PP}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}y \\
&= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{PP}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}y \\
&= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}y \\
&= \hat{b}_{\mathrm{MLR}}
\end{aligned} \tag{10.14}$$

Hence, PCR and MLR provide exactly the same estimator when no principal components are excluded.

PCR resembles RR when considered from an eigenvalue perspective. Where RR uses the gambit of offsetting eigenvalues to prevent them being small, PCR merely deletes them. RR gives a biased estimation through application of the offset, whereas the PCR estimator is unbiased. A good coverage of PCA and PCR applications is given by Jolliffe (1986).

Although PCR allows us to delete noise effects, we are still left with the possibility of irrelevant systematic effects being retained. Jolliffe (1982) and Davies (1995) both note the substantial inertia amongst users of PCR to the application of selection criteria to the generated principal components.

10.4.6 Latent variables

Latent variables are a useful concept when dealing with high dimensional sets containing collinearity. Consider, for PCA, we have $\mathbf{T} = \mathbf{XP}$, where $\mathbf{P}$ is the eigenvector matrix and $\mathbf{T}$ is the component scores matrix. We have already seen that only a few of the components may be significant so define $\mathbf{P}'$ to contain only the first few eigenvectors. Given $\mathbf{T}$ and $\mathbf{P}'$, we can estimate $\mathbf{X}$ by $\hat{\mathbf{X}} = \mathbf{P}^{\mathrm{T}}\mathbf{T}$. Viewing the model in this form, we have a small number of variables (the components) which are responsible for the $\mathbf{X}$ observations. These underlying variables are termed 'latent variables'.

The latent variables may be considered as causal effects for our observations $\mathbf{X}$. We wish to model the data $\mathbf{Y}$ from our observations. Hence it makes sense to form the model for $\mathbf{Y}$ on the causes rather than the effects. Given the data in $\mathbf{X}$ we can estimate the causes and from these form the model.

Figure 10.8 shows the relationship between traditional methods such as MLR and latent-variable-based methods. Using the traditional approach, the effects of any t_i are spread throughout the variables in $\mathbf{X}$. We have to select one of these to represent each t_i. The value of x_i selected is unlikely

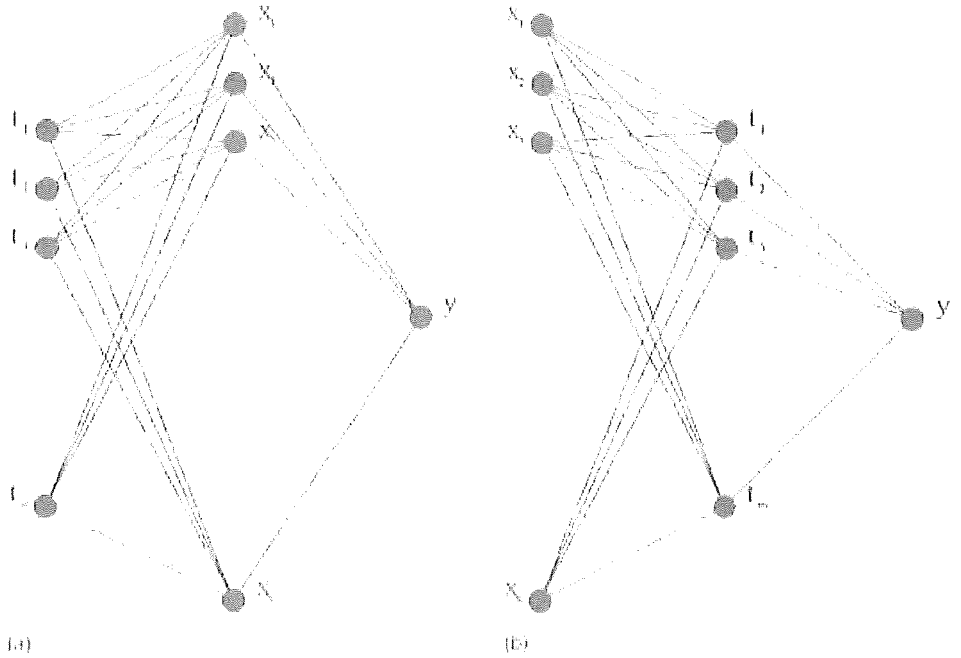

Figure 10.8 Causal relationships in modelling: (a) traditional multiple linear regression; (b) latent-variable-based modelling

to represent the variations of t_i alone, so the selection process is likely to introduce irrelevant effects which are inseparable from the relevant. By choosing a single value, x_i, to represent t_i, we are also making the model susceptible to noise. An error in measuring the single x_i will strongly affect the prediction, since we are relying entirely on x_i for the contribution of t_i.

Now consider the latent variable method. In this case, we are using all the x_i to model each t_i. An error in measuring a single x_i will not cause a major error in t_i, since we are (weighted) averaging over a number of x_i. At the second stage, we are predicting y by using a number of independent variables, so the effects of irrelevant latent variables can be completely removed.

Latent variable methods have another advantage when the number of variables in $\mathbf{X}$ is high. For simple MLR, we have to invert $(\mathbf{X}^{\mathrm{T}}\mathbf{X})$, which, as noted, means that the columns of $\mathbf{X}$ must be linearly independent. If $\mathbf{X}$ contains more columns than rows, this is not possible. Hence, for simple MLR we must have more samples than variables for the model to be stable. For spectroscopic applications, where hundreds of variables can be present, this is often not feasible. In these cases, then, we are forced to take fewer samples and immediately discard a number of variables, losing information in the process. The use of latent variables avoids this. If m samples of n variables are taken, with $m < n$, only m latent variables will have non-zero variance. All of the original variability in the data can therefore be retained for investigation.

10.4.7 Validation

When forming a model from calibration data, we encounter a problem. We are calibrating $\mathbf{X}_c$ against y_c. Since we have no extra information, the only method is to form a model which explains the y_c readings in terms of the X_i readings. It is possible to form a model which succeeds in relating X_c to y_c, but has poor predictive power.

As an example, consider the following extreme model:

$$\hat{y} = \begin{cases} y_m & \text{if } x = x_m, \text{for some } i; \\ 0 & \text{otherwise} \end{cases} \tag{10.15}$$

This model will predict y_c perfectly from $\mathbf{X}_c$ for any calibration set, but is unlikely to supply correct answers in the future. This situation, called overfitting, can occur very easily. Without reference to an external set of test data, we have no way of knowing whether we are losing predictive power. The model can be improved to the point where, rather than modelling real effects, it is modelling noise. At this point the data have been overfitted.

Conversely, if we are too cautious and stop improving the model too soon, we will underfit the calibration data by not taking relevant effects into consideration. In this case, when we come to using the model for prediction, the unconsidered effects may change and push the model away from correct answers.

So, how can we tell the difference between an underfitted, good, or overfitted model when we have no external reference points? This problem is the basis of model validation.

In order to assess the utility of a model, we need to obtain a value which gives a measure of its ability to predict future values. A commonly used measure is the mean square error of prediction (MSEP) and its root (RMSEP) (Allen, 1971). This is simply the mean squared difference between the predicted values of Y and the true values of Y for a given model. Obviously, better models will yield better MSEP values, so the aim of model formation is to minimize the MSEP.

When forming a model, we wish to minimize the influence of random variations in the data and maximize the influence of the underlying causal effects. It is therefore important to use many (X, Y) pairs for training, as the standard error of a sample set is inversely proportional to the number of samples. Conversely, for estimation of the MSEP we need many (X, Y) test pairs, for exactly the same reasons. When forming a model, we usually have a limited number of (X, Y) pairs to use. There is therefore a trade-off between using samples for model formation and MSEP estimation. A number of methods have been suggested for addressing this problem and these will now be discussed.

Self-prediction. It may be tempting to use self-prediction to estimate the MSEP for a model. In self-prediction the whole calibration set is used to

form the model. The MSEP is then estimated with the same calibration set. This method makes good use of the available observations because as many as possible are used for both modelling and validation. However, this is not a sensible method to use in general. Consider our trivial model above. It predicts every value in the calibration set perfectly and therefore has an estimated MSEP of zero for self-prediction validation. The true MSEP will be much greater and the model is in fact useless in all but contrived circumstances.

Self-prediction does, however, have the advantage that its MSEP estimate gives a lower bound on the true MSEP of the model. It is often instructive, when using latent variable models, to view a graph of the MSEP for self-prediction versus the number of latent variables. The 'true' MSEP graph will usually tend to show a local minimum as we move from underfitting to overfitting. In the case of self-prediction, the MSEP graph often shows

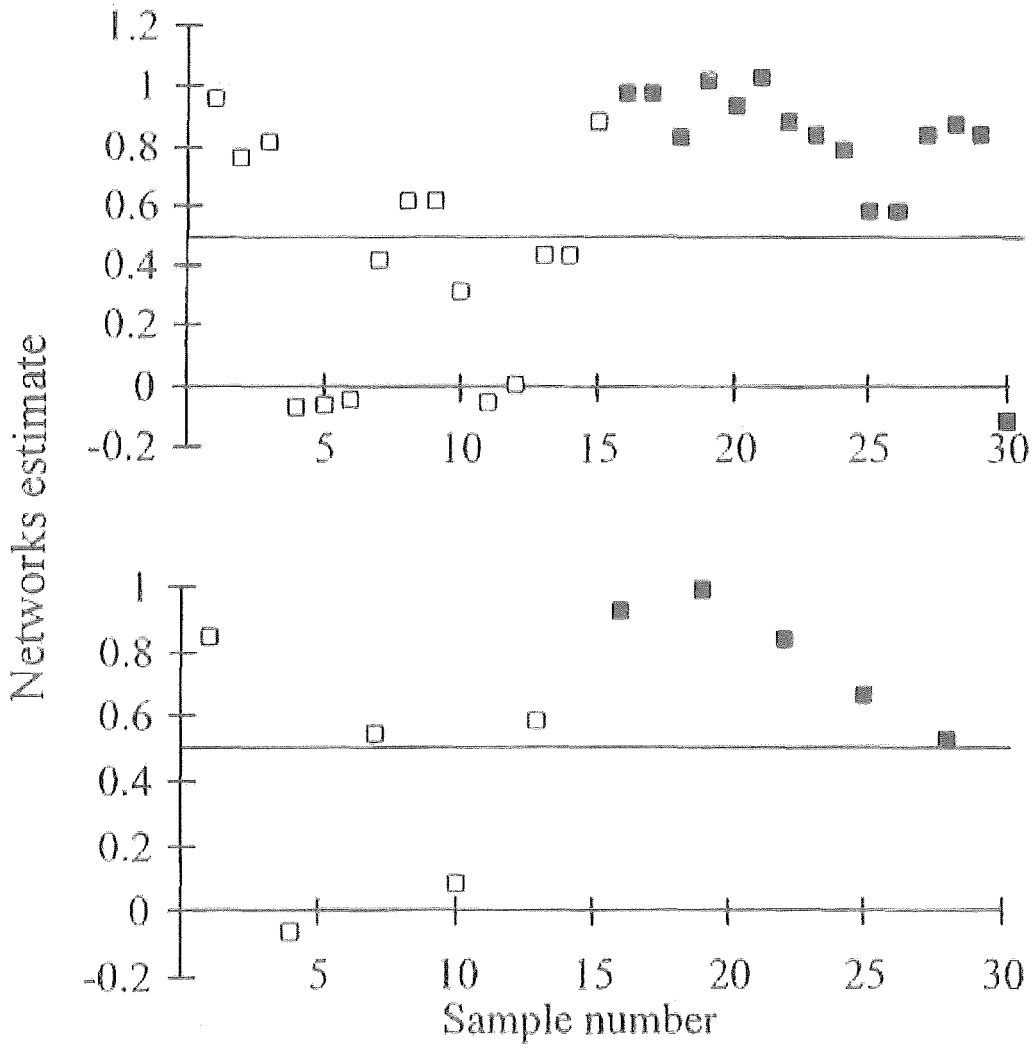

Figure 10.9(a)

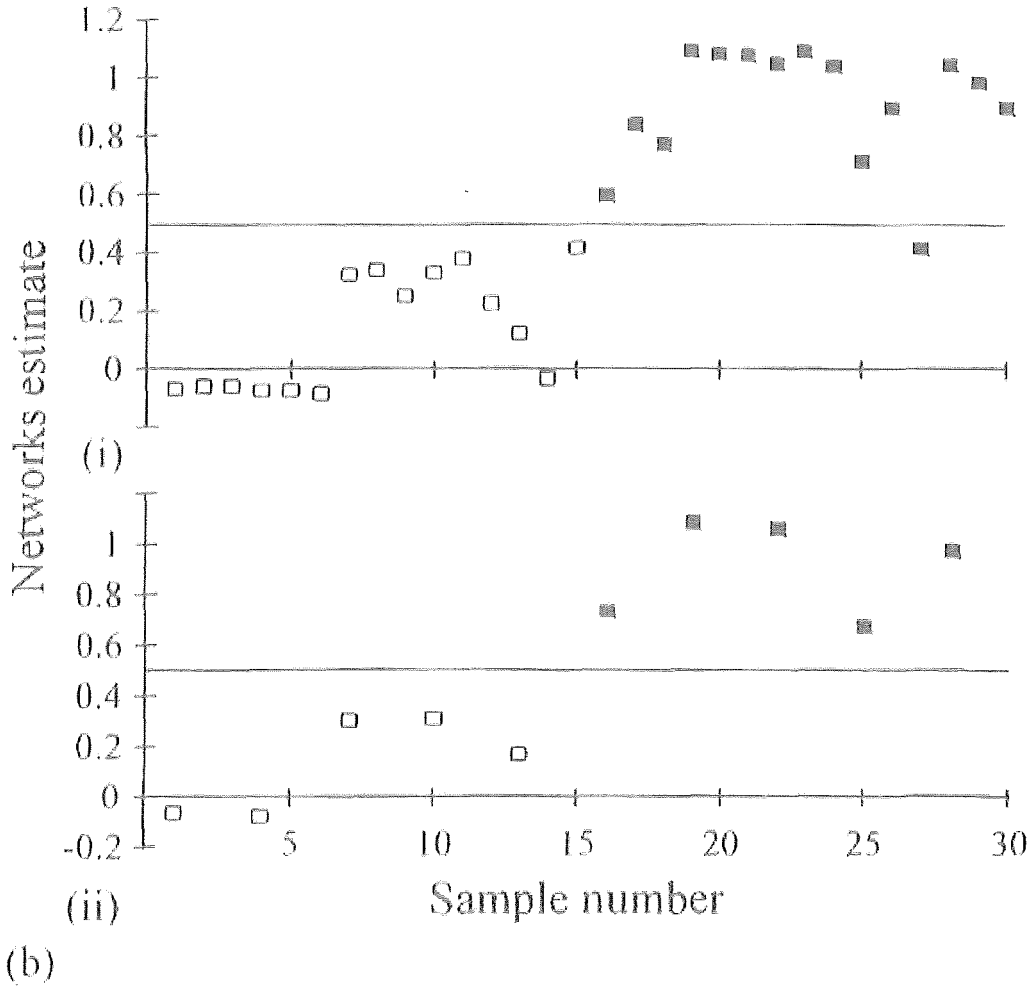

Figure 10.9 Prediction of the region of origin of extra virgin olive oils by means of artificial neural networks (ANNs). (a) ANNs trained on randomly split data (the output was taken after 27 000 epochs, with a training error of 0.009): (i) all samples in triplicate form; (ii) averaged output for each triplicate form. (b) ANNs trained on data split by means of the Duplex partitioning method (the output was taken after 8000 epochs, with a training error of 0.0196): (i) all samples in triplicate form; (ii) averaged output for each triplicate form. In each training method a three-layered ANN (150 input nodes, 6 hidden nodes and a single output node) was trained with 33 objects (15 oil samples from Lazio, coded 0, and 18 samples from Sicily, coded 1) and interrogated periodically by using 30 test objects [15 Lazio (□) and 15 Sicily (■)] previously unseen by the training net. The horizontal line corresponds to a network estimate of 0.5.

similar reductions in MSEP until the 'true' MSEP local minimum, tending then to trail off slowly to zero as the number of latent variables increases. A sudden change of slope can often be seen in the self-prediction MSEP graph, this being useful in estimating the number of latent variables to use.

So, if there is no alternative to self-prediction, it may be used with great care to make an attempt to select the number of factors, but its use is in general 'dangerous' and to be avoided.

Training and test sets. If more data can be obtained easily, then the validity of a model can be tested by using a new set of observations to estimate the

MSEP directly. The calibration set is used merely for forming the model, with the new data being used for validation.

Unfortunately this method depends on the availability of these extra observations. The aim of multivariate calibration is often to allow prediction of values of Y which are difficult or expensive to measure. When this is the case, time or expense can preclude the acquisition of large amounts of test data. In fact it is often the case that a single set of observations is presented with which to calibrate and validate the model. This complicates matters somewhat.

The simplistic approach in this situation is to partition the set of observations into two sets and to use one for calibration and the other for validation. Each set must be chosen to contain a representative spread of values of X and Y, otherwise the model will be formed or validated from a subset of the available range, leading to erroneous predictions or estimations of validity. Snee (1977) suggests a number of methods by which a suitable partition of the data may be achieved. Figure 10.9 shows the effect of different partitioning regimes.

The trade-off between calibration and prediction set sizes mentioned above is especially important in this situation. A model based on only half the observations is quite likely to contain an unrepresentative set of calibration objects. Increasing the number of observations for calibration will reduce the chances of choosing a poor calibration set while reducing the quality of the MSEP estimate.

Full cross-validation. Ideally, what is needed is a hybrid of self-prediction and training and test set validation. We need to use all observations in both model formation and validation without encountering the problems of self-prediction. Full cross-validation (Geisser, 1975; Stone, 1974) attempts to do just this. The term 'cross-validation' is often applied to the partitioning form of training and test set validation and it is from this that full cross validation was developed (from here on we will use the term 'cross-validation' to refer to full cross-validation).

When the observations are partitioned into two equal sets we can form the model from either set and validate using the other. An improved idea is to form two models, one from each set, and in each case use the remaining set for validation. We then have two estimates of MSEP and two models to try. If the partition splits the data well then we would expect these models to place similar significance on each variable and thereby have similar properties and MSEPs. A model formed with use of the full data set should also have a similar structure and similar (or better) MSEP.

We can therefore use the following steps to estimate the MSEP of a model where the full data set is used:

1. partition the observations into two sets, S_1 and S_2;
2. form a model from S_1, use S_2 to estimate its MSEP (MSEP1);
3. form a model from S_2, use S_1 to estimate its MSEP (MSEP2);
4. form a model from the full set of observations and average MSEP1 and MSEP2 to estimate its MSEP.

The problem of choosing a 'good' partition still remains, but we now have a model based on all the observations, and an MSEP also based on all observations, but no self-prediction.

Consideration of the above steps leads to a method for reducing the partition-selection problem. We need to increase the proportion of observations used in the individual models. Instead of partitioning into two sets, the set can be split into n sets of approximately equal size. For each set, S_i, a model is formed from all observations not in S_i. The MSEP for this submodel is then estimated by using S_i as the test set. All n estimated MSEP values are then averaged to give an estimate of the MSEP for the full model. Again, we have the full observation set being used to create and validate the model, but in this case the larger calibration partition size reduces the chance of selecting a 'bad' spread of observations. For $n = 2$, this procedure equates to that detailed above. The limiting case occurs when each S_i contains a single observation. This is called 'leave-one-out' validation. This limit should give a good MSEP estimate as each submodel differs from the full model only in the effect of the single omitted observation. The choice of n is a trade-off between modelling time for n models and MSEP estimate quality.

Cross-validation is a good method for estimating the optimal number of components in latent variable methods because in these cases we expect the model to move from underfitted to optimal to overfitted through addition of components. A graph of MSEP versus number of components will therefore show a minimum at the point where the model is optimally fit (Wold, 1978).

Leverage correction. Leverage is a concept applied to observations used in a calibration model. It is a value between $1/n$ and 1, where n is the number of observations. The leverage of an object indicates its importance to the structure of the model. Observations having high leverage are important to the model in that they contribute greatly to its structure.

Consider the case when a model is fitted increasingly closely to its calibration set. Any observations which contain high noise levels (i.e. outliers) will start to gain in leverage as they skew the model to fit them. For such an overfitted model, then, there will be an increasing number of observations having high leverage.

This consideration gives rise to the idea of using leverage to estimate the validity of a model. Indeed it has been shown (Martens and Næs, 1989) that the residuals (and hence MSEP) for cross-validation, $f_{i(CV)}$, can be estimated from the leverage of observations in an MLR model by means of the relations

$$\hat{f}_{i(\mathrm{CV})} = \frac{\hat{f}_i}{1 - h_i}, \tag{10.16}$$

$$h_i = \frac{1}{n} + \hat{t}_i^{\mathrm{T}} (\hat{\mathbf{T}}^{\mathrm{T}} \hat{\mathbf{T}})^{-1} \hat{t}_i, \tag{10.17}$$

where h_i is the leverage of observation i and $\hat{f}_i$ its Y residual. This scheme works as expected in that if an observation is an outlier and the model is forced to fit it then its leverage h_i will be close to unity and hence the estimated cross-validation residual large. The MSEP estimate,

$$\frac{1}{n} \sum_{i=1}^{n} \hat{f}_i^2 \ (\mathrm{CV}),$$

will therefore also be large. Notice also that a high leverage value for an observation in a non-overfitted model is an indicator that that observation may be a significantly bad outlier and is worthy of investigation.

For cross-validation the $\hat{t}_i$ will be different for each model. But, as noted, for similar models, these $\hat{t}_i$ should be similar. Hence the residuals, $\hat{f}_i$ should also be similar for each submodel used. We may therefore simply use the residuals for the full calibration model and avoid the multiple calibration of cross-validation.

The process for leverage correction validation is therefore:

1. model on full calibration set and retain the Y residuals;
2. calculate the leverage for each observation from the model scores;
3. estimate the MSEP from residuals and leverages.

For calibration methods other than MLR, the cross-validation relation [equation (10.16)] does not hold, but Martens and Næs (1989) have suggested adjusted leverage-correction MSEP, $E_{\mathrm{MSEP}}^{\mathrm{LC}}$, estimators for PCR and PLSR:

$$E_{\mathrm{MSEP}}^{\mathrm{LC}} = \frac{1}{F_{\mathrm{df}}} \sum \hat{f}_i^2 \left[1 - \sum_{a=1}^{A} \left(\frac{\hat{t}_{ia}^2}{\hat{\mathbf{t}}_a^{\mathrm{T}} \hat{\mathbf{t}}_a} \right) \right]^{-1}, \tag{10.18}$$

where A is the number of factors, $\hat{\mathbf{t}}_a$ the PCA (or PLS) scores vector for factor a and $\hat{f}_i$ the residuals; F_{df} is the number of degrees of freedom. Martens and Næs recommend that a good value for F_{df} is $n - 1 - A$, this value reducing the underestimation of MSEP that is inherent in leverage correction.

10.4.8 Partial least squares regression

Consider the NIPALS algorithm for PCA discussed earlier (Section 10.4.1). It was stated that the method uses the fact that the PCA loadings are regression coefficients for scores, and vice versa; that is, for X, an $r \times c$

data matrix, with $p(c \times 1)$ and $t(r \times 1)$ being the corresponding loadings and scores vectors for the first component. We have $t = Xp$ by definition. This may be written in two ways to give rise to regression equations:

$$X = tp^{\mathrm{T}}, \tag{10.19}$$

or

$$X^{\mathrm{T}} = pt^{\mathrm{T}}. \tag{10.20}$$

Regressing X on t in the first equation gives $p^{\mathrm{T}} = (t^{\mathrm{T}}t)^{-1}t^{\mathrm{T}}X^{\mathrm{T}}$, and regression of X on p in the second equation gives $t^{\mathrm{T}} = (p^{\mathrm{T}}p)^{-1}p^{\mathrm{T}}X$. Simplifying this and normalizing so that $\hat{p}$ is of unit length gives the PCA NIPALS algorithm. Lyttkens (1966) showed that this algorithm was convergent to the required values for p and t.

Equations (10.19) and (10.20) represent a system of three variables where each is linearly dependent on the other two. The NIPALS method was soon shown to be applicable to many such systems (e.g. Lyttkens, 1973).

Wold (1975) developed a methodology for designing NIPALS models for multivariate regression. The basis of a model is an 'arrow diagram'. The variables are split into blocks, each of which is considered to be related to a latent variable. Each variable, both latent and observed, is represented as a box (or circle). Arrows are placed between the boxes, these indicating causal relationships. Hence arrows will lie between a latent variable and its block of observed variables. If the observed variable is considered to be causal to the latent variable then the arrow points towards the latent variable. If the variations of the latent variable are seen to explain the observed variations then the arrow points in the opposite direction. Interrelationships between latent variables may also be indicated by arrows. Wold's method allows the arrow diagram to be decomposed into a set of regressions which, when iterated, will provide the latent variable loadings. A number of methods are defined for forming the regression at this level, termed modes A, B and C. Wold (1980) terms mode A as being a regression where the 'outgoing' variables are explained by the latent variable, mode B being a regression where the incoming variables explain the latent variable. Mode C is the case where both A and B are used in a model.

An example should clarify matters. Figure 10.10 shows a path model for PCR. The latent variable (first principal component) is caused by the X variations and it is this which is assumed to cause the Y variations. Hence the X arrows point into the latent variable and the Y arrows point outward. The latent variable is therefore a weighted sum of the X_i. Let $\hat{p}$ and $\hat{q}$ be the X and Y weight vectors, X and Y the X and Y data matrices, and t the latent variable vector. Then we have:

$$t = Xp, \quad \text{normalized to length 1;} \tag{10.21}$$

which, by rearranging, gives

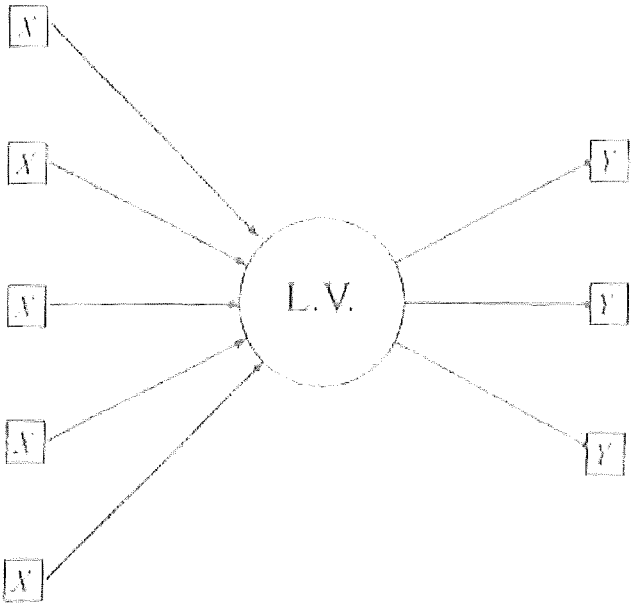

Figure 10.10 Path model diagram for principal components regression. L.V. = latent variable, X and Y are observed variables. The direction of the arrows indicates the causal relationship; for instance, variables X are causal to the latent variable, and variations in the latent variable explain variations in Y

$$\mathbf{X} = \boldsymbol{t}\boldsymbol{p}^{\mathrm{T}}; \tag{10.22}$$

$$\mathbf{Y} = \boldsymbol{t}\boldsymbol{q}; \tag{10.23}$$

and, by MLR,

$$\hat{\boldsymbol{p}}^{\mathrm{T}} = (\boldsymbol{t}^{\mathrm{T}}\boldsymbol{t})^{-1}\boldsymbol{t}^{\mathrm{T}}\mathbf{X} \tag{10.24}$$

and

$$\hat{\boldsymbol{q}} = (\boldsymbol{t}^{\mathrm{T}}\boldsymbol{t})^{-1}\boldsymbol{t}^{\mathrm{T}}\mathbf{Y}. \tag{10.25}$$

So, by initializing $\hat{\boldsymbol{p}}$ to a non-zero value and iterating over

$$\hat{\boldsymbol{t}} = \mathbf{X}\hat{\boldsymbol{p}}, \quad \text{normalized}, \tag{10.26}$$

and equations (10.24) and (10.25) until convergence of $\hat{\boldsymbol{p}}$, we can obtain a PCR model. In fact, $\hat{\boldsymbol{q}}$ is not used inside the loop, so it may be calculated after $\hat{\boldsymbol{p}}$ has converged. With $\hat{\boldsymbol{q}}$ removed from the loop we can see that the iterative part is equivalent to the NIPALS PCA algorithm, with the final $\hat{\boldsymbol{q}}$ regression forming the PCR. [A discussion of the case where there are several interacting variables with associated blocks is given by Wold (1982); Bookstein (1980) gives an insight into the geometric interpretation of NIPALS methodology, now commonly called PLS.]

The method we shall refer to as PLS is one of incorporating the Y data into the latent variable modelling step by bringing an extra projection into the iterative sequence, as follows:

$$w^{\mathrm{T}} = (t^{\mathrm{T}}t)^{-1}t^{\mathrm{T}}\mathrm{X}, \tag{10.27}$$

$$t = \mathrm{X}w, \tag{10.28}$$

$$q^{\mathrm{T}} = (t^{\mathrm{T}}t)^{-1}t^{\mathrm{T}}\mathrm{Y}, \tag{10.29}$$

$$t = \mathrm{Y}q, \tag{10.30}$$

To put it in a visual manner, the location of t gets pulled first towards the X variables then towards the Y variables each time round the iteration. In the case of PCR, the only alteration to t is to pull it towards the X variables, so we end up converging to principal components. The pull of the Y variables tends to bring convergence such that the latent variables chosen are of relevance to both X and Y. Both sets of loadings (w and q) are required for prediction in the case of PLS.

A simplified PLS model exists for cases when only a single Y variable is to be modelled. PLS 1 is non-iterative and can therefore be used to generate a result quickly in these cases (details are given in Martens and Næs, 1989).

PLS has come to the fore as an important modelling method because of the improved latent variables it generates. Since these variables are extracted in decreasing relevance to Y, the problem of variable selection is bypassed to a significant degree. Each PLS factor is extracted and its effect subtracted from the remaining X (and Y) variation before the next is obtained. The PLS factors obtained in this manner are always generated in decreasing order of importance to the overall model and the problem of selection is reduced to that of the number to use. This problem is handled easily by standard validation methods. De Jong (1993a) has shown that for any number of latent variables, PLS will fit at least as well as PCR.

Standard spectroscopic methods have tended to benefit from the use of PLS for gaining useful insights into data (Bhandare *et al.*, 1993; Haaland and Thomas, 1988a, b). In this discipline it is usual to obtain intensities at a large number of wavelengths. The wavelengths are considered as variables and the intensities their values. There is often a high degree of collinearity within such data sets and the phenomenon of interest is often distributed throughout a large number of the measured wavelengths. MLR-based methods are therefore of limited use, as a large number of wavelengths would have to be dropped to escape the collinearity problem, and this would tend also to lose the interesting phenomena.

10.4.9 Artificial neural networks

The X and Y matrices of the statistical models are analogous to the training inputs and outputs of a neural network used to create the model, and the test inputs and outputs used to predict values. Cheng and Titterington

(1994) take a look at neural networks from a statistical point of view. Indeed, Sarle (1994) suggests that neural networks are no more than non-linear regression and discriminant models that can be implemented by means of standard statistical software; Ripley (1994) agrees with this statement.

The human brain is a very complex organ. It contains around 10^{10} neurons (that is, the basic processing, or nerve, cell). Neurons consist of two types (Beale and Jackson, 1990):

- interneuron cells, with input and output locally (over a distance up to 100 μm);
- output cells, connecting to different regions of the brain, to muscles, or connecting from sensory organs (e.g. the eye) into the brain.

If enough active inputs are received to an individual neuron at any one time, that neuron is activated (it 'fires'), otherwise it remains inactive. This is analogous to the McCulloch–Pitts model of the neuron, proposed as long ago as 1943. Their model of neural activity is described in detail by Aleksander (1989).

Thus the brain is massively parallel, any one processing job being shared between many neurons. The consequence of this is that any one single neuron is not generally very important; if one neuron were to malfunction for some reason, it would be relatively unlikely to affect other neurons significantly. This kind of processing, spread over many processing units, is known as 'distributed processing', and is widely considered to be fairly tolerant of errors (to have 'fault tolerance').

The traditional computer has one (or maybe two or a few more) processors. The consequences of this are clear: in addition to the inferior processing power, if the only processor should malfunction then the whole system would be unviable. As the term 'neural network' implies, this field of computing was originally aimed towards modelling networks of real neurons in the brain (Hertz, Krogh and Palmer, 1991).

Neural networks, which naturally lend themselves to parallel processing (although they are often presently run, by necessity of available equipment, on single-processor computers), clearly overcome this problem of possible failure by allowing processing to continue in the event of a failure of one of the constituent neurons.

Traditionally, neural networks are seen as a branch of artificial intelligence (AI), although some AI academics disagree over this. It surely could be said that it is one of the most unquestionably AI areas in computing science. In most other areas of AI, learning takes place in a fully understood and predictable way using explicitly represented knowledge (Luger and Stubblefield, 1989), whereas for neural networks the way in which the net learns is not known, nor is it predictable or repeatable, because of the random nature of the network initialization (discussed below).

Neural networks can learn to recognize patterns within sets of data (which may be an encoded image, a spectrum or any encoded set of related data). Fault tolerance helps in pattern recognition, allowing the net to cope with differences between input examples it receives, placing greatest importance only on the parts of the input data which are important for distinguishing the data it is learning.

So what is meant by 'learning'? A good definition is given by Judd (1990, p. 3), who states that 'Learning is the capacity of a system to absorb information from its environment without requiring some external intelligent agent to program it'. By 'program' it can be understood that Judd is referring to more direct methods of information input, such as that used in an expert system. Judd goes on to say that, unfortunately, all learning algorithms so far reported are unacceptably slow for large networks. Given the size of the human brain it is no great wonder that attempts to simulate this sort of complexity have not succeeded! Nonetheless, the principle of neurons as elements of processing is not lost on reduction to a smaller sized network, such as those modelled by the various packages available on desktop computers. A tutorial review of neural networks, with particular reference to multivariate PyMS applications, is given by Goodacre, Neal and Kell (1996).

With standard back propagation, the most common type of neural network, there are a number of input nodes (equal to the number of inputs), each connected to every node of a hidden layer, which are in turn each connected to the output node(s) (Fig. 10.11). Each node in the input layer brings into the network the value of one independent variable. The hidden layer nodes (called 'hidden' because they are hidden from the outside world) do most of the work (Smith, 1993). Each output node passes a single dependent variable out of the network.

In neural net jargon, the neuron is known as a 'perceptron' (Rosenblatt, 1958). The learning rule for these 'multilayer perceptrons' is called the back-propagation rule. This is usually ascribed to Werbos in his thesis of 1974 (Werbos, 1993), but was popularized by Rumelhart and McClelland (1986) as recently as 1986, since when there has been a revival in interest in neural networks.

The network is initialized with random weights on the connections between the perceptrons, and the input is applied to the input nodes. By using a training set of data and comparing the output from the net with the desired (known) output it is possible to calculate new values for the weights to increase the accuracy by decreasing the error between known and actual values. Each of these iterations is known as an 'epoch'. An error function is defined to represent the difference between the network's output and the desired output. The aim therefore is to reduce the value of this function as far as possible. The back-propagation rule does this by calculating the value of the error function for one particular input and propagating the error back from one layer to the previous one. Thus each connection has its weights

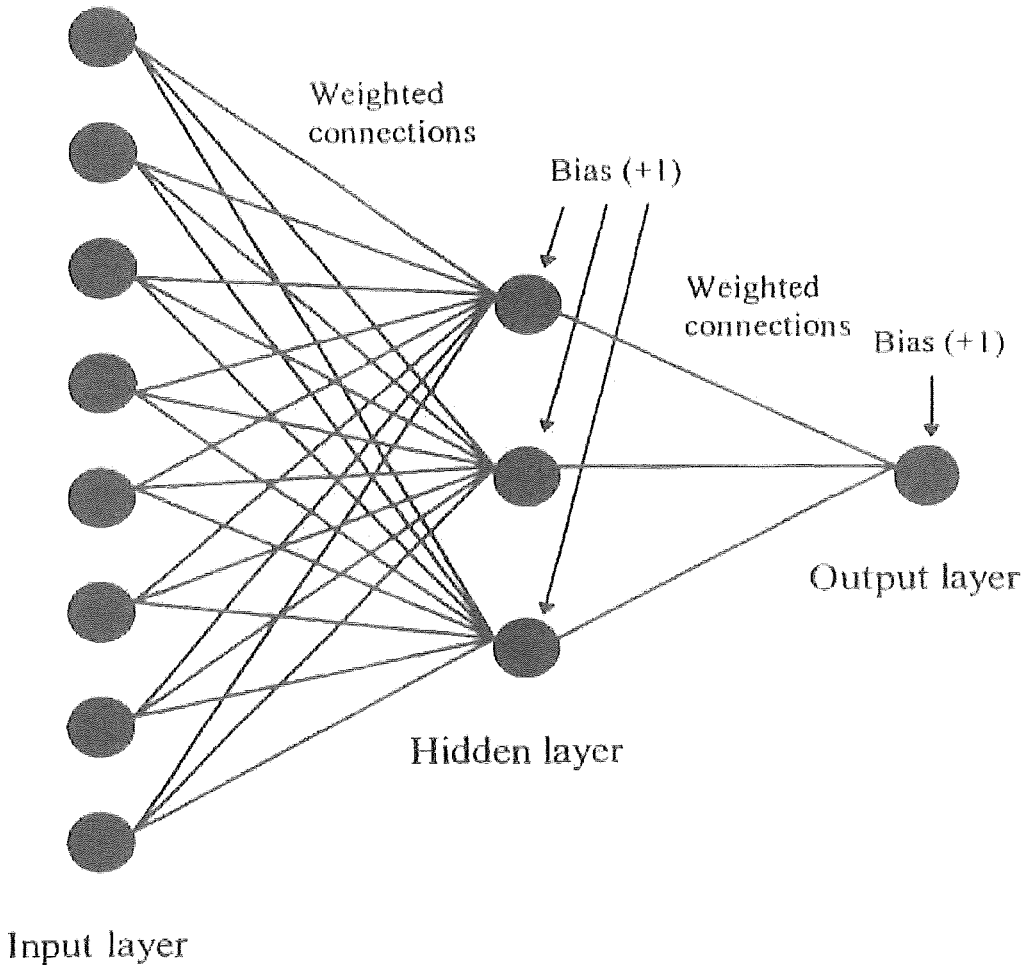

Figure 10.11 Structure of a standard back-propagation 8–3–1 neural network.

adjusted. The mathematics involved in this procedure are often badly presented, but some of the most understandable explanations are by Beale and Jackson (1990) and Bishop (1995).

One problem frequently encountered with neural networks is that of overtraining. This occurs when the net trains so closely to the training set that any other data will not be recognized. In this case, when the net is tested, it will be found that examples which were in the training set will give highly accurate results, and other examples may be wildly inaccurate.

Take, for example, the training and test data shown in Fig. 10.12. Overtraining results in the fit shown with a dashed line, where the points in the test set are poorly fitted to the line. The ideal line is shown by a solid line.

According to Hecht-Nielsen (1989, p. 116), the exact origin of this problem has still not been fully elucidated, but it seems to be related to the manner in which the afflicted networks form their mapping approximations. When testing, it is normal to use the training set as well as a test set which has not been seen by the network during training – this helps to show up any over-

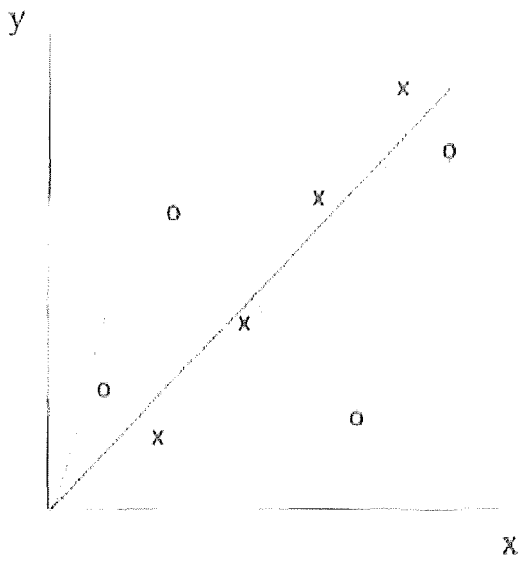

Figure 10.12 Graph showing how a network may be overtrained to the data in the training set. ○ = training set; × = test set; —— = desired train; ----- = overtrained.

training as well as showing how well the network has learnt what it has seen and how good at generalization (dealing with previously unseen data) it is.

It is clear, then, that neural networks, being good at dealing with large numbers of inputs, lend themselves to the analysis of spectra such as those obtained by the PyMS technique and NMR. PyMS, for example, produces a spectrum of 150 elements, therefore the network used will have 150 input nodes. Goodacre, Kell and Bianchi (1992, 1993) have used neural networks for the detection of adulteration in olive oil from PyMS data. Francelin, Gomide and Lanças (1993) compare three different types of neural network for the classification of vegetable oils, including olive oil, after gas chromatography analysis.

10.4.10 Chemometrics

The use of PLS and related methods in a wide range of sciences has led to the emergence of a new discipline, that of chemometrics. This is devoted to the study of pragmatic multivariate methods in sciences, and the literature has grown rapidly over recent years (e.g. Brereton, 1992; Brown *et al.*, 1996; Ramos *et al.*, 1986). With the increasing use of PLS has come the realization that it is not well understood in terms of its statistical properties. The method was designed to avoid making assumptions about structure within the data except those that are built into the path model. No assumption of normality is placed on the variables, for instance.

Frank and Friedman (1993) and Helland (1988) have studied the more commonly used chemometric regression methods (namely PLS and PCR) from a statistical viewpoint. Their studies highlight the general similarity

between these methods and that of ridge regression. This tends to confirm the proposition that there is a continuum of regression methods extending from MLR to PCR, with PLS lying within this continuum. Stone and Brooks (1990) derived the continuum regression (CR) algorithm, for which there are two controlled parameters: the number of factors (termed ω) and a position indicator, α. When $\alpha = 0, \frac{1}{2}, 1$, CR is equivalent to MLR, PLS and PCR, respectively. The optimal values for α and ω are estimated using cross-validation. Malpass *et al.* (1994) have investigated the continuum regression methodology and have simplified the selection of optimal α values (1993). Seasholtz and Kowalski (1993) have highlighted the importance of choosing parsimonious models. That is to say that if two models are formed then the one which requires fewer parameters for its description is more likely to provide good future predictions.

On the algorithmic front, a number of new PLS methods have been proposed, for example those of Lindgren, Geladi and Wold (1993) and de Jong (1993b). Extensions have been proposed to improve the situation when non-linear data are encountered. The methods discussed so far assume that the observed and latent variables are linearly related. When non-linearity is encountered, the methods tend to fail to make a good model. This phenomenon shows up in the requirement of large numbers of factors in model formation, and in curvature in plots of the residuals after modelling. Taavitsainen and Korhonen (1992) have shown how PLS can be extended to incorporate certain types of non-linearity at the latent variable stage, and Oman, Næs and Zube (1993) have used a PCR-based method augmented with products of the components for similar purposes.

10.4.11 Variable selection

Variable selection is a technique whereby those variables which are considered to be most important in the creation of a model are used, whereas others are discarded. There are many arguments for selecting variables and a number of ways of selecting them, some of which are described in this chapter. To date, comparatively few researchers have used variable selection methods, most concentrating on improving prediction by using all the variables.

Why select variables? The rationale behind using latent variable methods is the exclusion of effects that contribute only noise. Latent variable methods assume that there are underlying effects which are expressed to varying degrees in the measured variables. If, however, some of the measured variables do not reflect any relevant underlying effect then there is no reason to include them in the modelling process. Indeed, since latent variable methods rarely assign zero relevance to any given variable, one may expect such irrelevant variables to introduce noise despite the application of such methods.

The principle of parsimony (de Noord, 1994; Flury and Riedwyl, 1988; Seasholtz and Kowalski, 1993) states that if a simple model (that is, one with relatively few parameters or variables) fits the data then it should be preferred to a model that involves redundant parameters. A parsimonious model is likely to be better at prediction of new data and to be more robust against the effects of noise (de Noord, 1994). Despite this, the use of variable selection is still rare in chromatography and spectroscopy (Brereton and Elbergali, 1994). Note that the terms 'variable selection' and 'variable reduction' are used by different researchers to mean essentially the same thing.

Chatfield (1995) warns of the dangers of selecting variables in such a way as to enable some sort of model to be made from pure noise. This is a risk with some methods such as genetic algorithms (Bangalore *et al.*, 1996; Broadhurst *et al.*, 1997; Horchner and Kalivas, 1995; Jouan-Rimbaud *et al.*, 1995; Kubinyi, 1994a, b, 1996); if these methods are used, thorough validation of the results is necessary to avoid this problem.

Some variable selection techniques for classification. It can easily be shown that certain variables in a data set not only contribute little to the model but actually detract from the optimum model. Let us take a simple, imaginary, case to demonstrate this. If we have a set of data describing two different varieties of olive oil, Leccino and Frantoio, and only three variables, we can look at the data and see which of the variables is most valuable for discriminating between the two varieties (Table 10.3).

If we take the standard deviation (StDev) of the Leccino and Frantoio oils for variables 1, 2, and 3 and then calculate the average of these, we have a value which represents the 'inner variance' or 'reproducibility'. The higher

Table 10.3 Finding the most valuable variables for discrimination between two olive oil varieties, Leccino (Le) and Frantoio (Fr). StDev = standard deviation; w is defined in text in equation (10.31) and indicates the characteristicity of a variable.

Variety	Variable		
	1	2	3
Le 1	3.4	5.4	6.4
Le 2	3.2	2.8	4.5
Le 3	3.5	8.6	5.9
Fr 1	3.0	3.9	5.1
Fr 2	3.2	7.5	3.5
Fr 3	3.1	5.5	4.9
StDev Le	0.152	2.905	0.984
StDev Fr	0.1	1.803	0.871
Average StDev	0.126	2.354	0.928
StDev All	0.186	2.162	1.027
W	0.678	1.088	0.903

this value, the higher the inner variance and the lower the reproducibility (Eshuis, Kistemaker and Meuzelaar, 1977). By taking the standard deviation over all the samples, we calculate a value which represents the 'outer variance' or 'specificity'. The higher this value the greater the outer variance and the greater the specificity. The ratio between inner variance and outer variance represents the 'characteristicity'. [The terminology 'characteristicity', 'reproducibility' and 'specificity' has been adopted from Eshuis, Kistemaker and Meuzelaar (1977).] So, by dividing our value for the average of the standard deviations by the standard deviation of the whole we get a value w which is an indication of the characteristicity of the variable (for varieties, var. 1, ... var. n):

$$w = \frac{\text{average [St Dev (var. 1), St Dev (var. 2), ..., St Dev (var. } n)]}{\text{St Dev (all samples)}} \quad (10.31)$$

Now, if w has a value greater than 1 then the inner variance is greater than the outer variance; therefore this variable is a hindrance to correct discrimination and so it should definitely be discarded.

The PCA scores plots of the data shown in Table 10.3 demonstrate the effect of selecting variables. Taking all three variables [Fig. 10.13(a)] it is apparent that discrimination is possible, but not easy. Eliminating the worst variable (with $w > 1$) makes discrimination easier [Fig. 10.13(b)]. If the best variable is removed, discrimination is impossible [Fig.10.13(c)]

When looking at data sets with many varieties, where one variety may contain more samples than another, it could be desirable to weight w in favour of varieties with a greater representation. In this case, we can use:

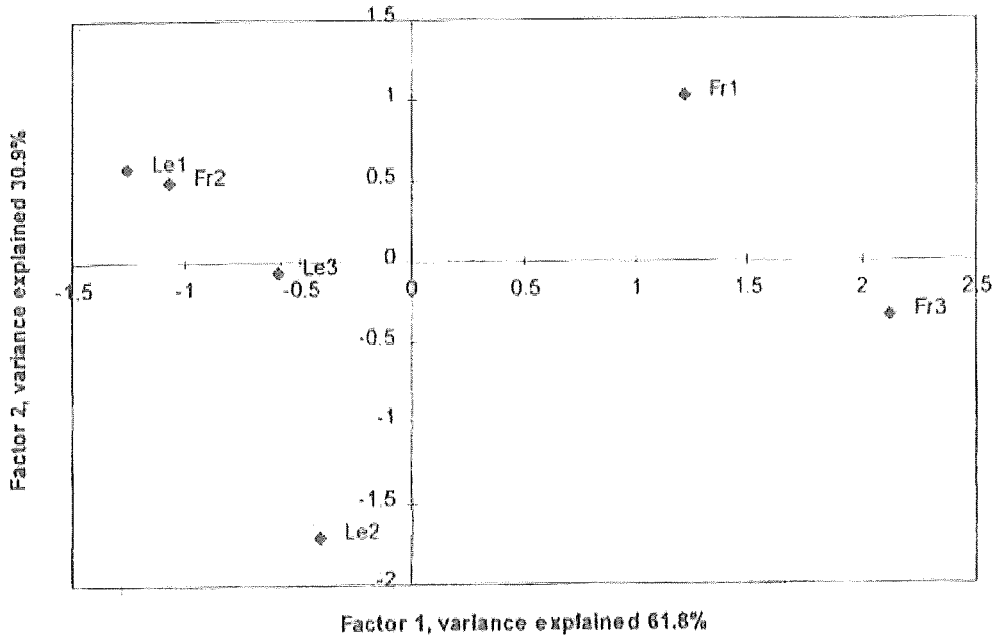

Figure 10.13(a)

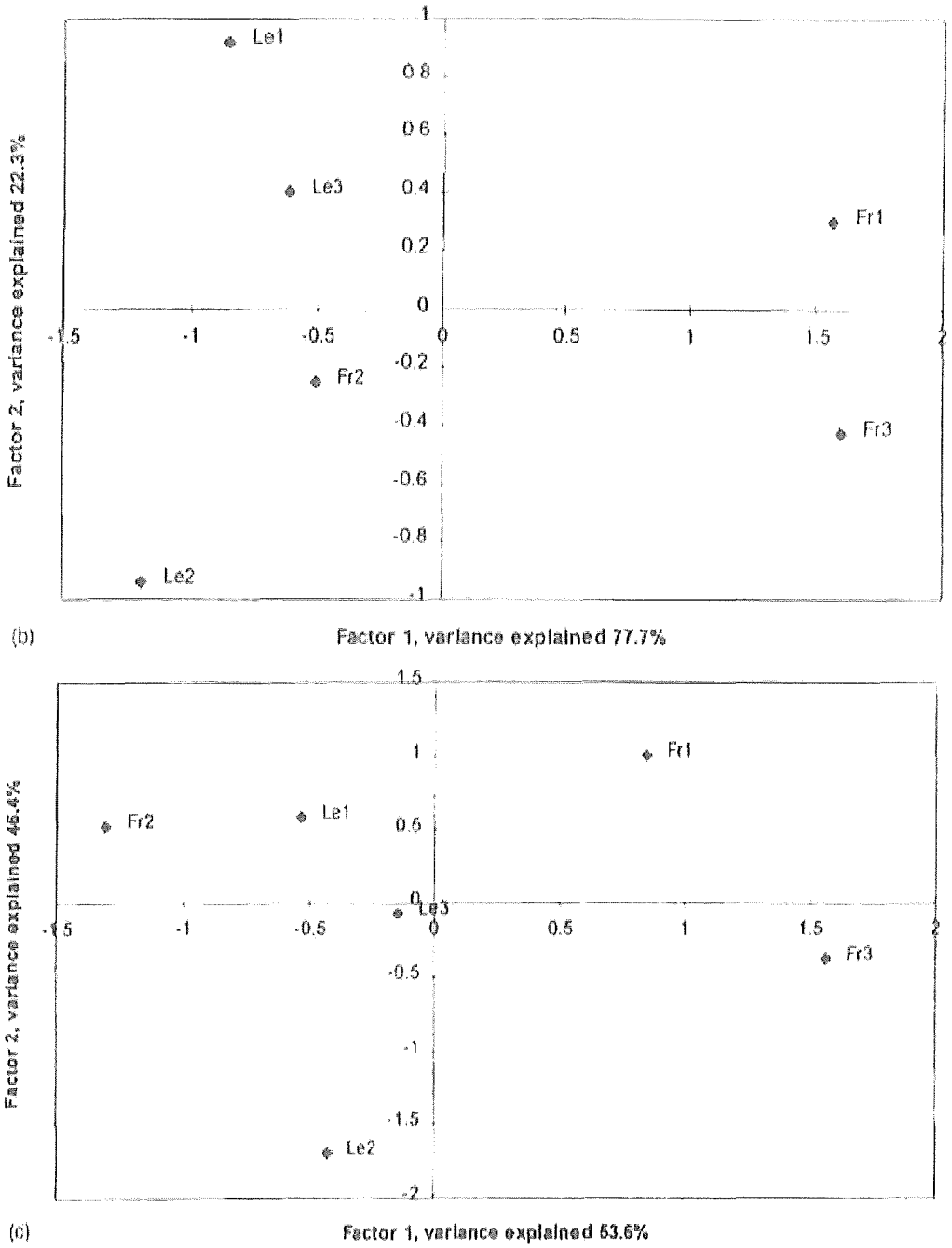

Figure 10.13 Principal components analysis scores plots: (a) using all three example variables (first two principal components; discriminations of the varieties, particularly sample Le2, would be very difficult); (b) using the best two variables, unweighted w selected [equation (10.31) in text] (discrimination of the varieties is now possible using only the first principal component); (c) discarding the best variable, unweighted w selected (linear discrimination of the varieties would not appear to be possible from this chart). Details of the variables are given in Table 10.3.

$$w = \frac{[\text{StDev(var.1)} \times N(\text{var.1})] + [\text{StDev(var.2)} \times N(\text{var 2})] + \ldots + [\text{StDev(var.}n\text{)} \times N(\text{var.}n)]}{N(\text{total}) \times \text{StDev(all samples)}} \tag{10.32}$$

Where N (var. n) is the number of samples for variety n and N (total) is the total number of samples.

Another method of variable selection for use in classification problems involves use of the Fisher ratio, whereby a value is calculated for each variable according to the following formulae.

Calculation of the between-group variation, V_b:

$$V_b = \sum_{i=1}^{g} n_i(\bar{y}_i - \bar{y})^2. \tag{10.33}$$

Calculation of the within-group variation, V_w:

$$V_w = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2. \tag{10.34}$$

Calculation of the Fisher coefficient, F:

$$F = \left(\frac{1}{g-1} V_b\right)\left(\frac{1}{n-g} V_w\right)^{-1}, \tag{10.35}$$

where g is the number of groups;
n_i is the number of elements in group i;
$\bar{y}_i$ is the mean value of group i;
$\bar{y}$ is the total mean;
y_{ij} is the value of object j in group i.

The three selection methods are henceforth referred to as weighted w, unweighted w, and Fisher.

It may often be found that factors other than that searched for (e.g. olive oil variety) may be having some influence on the data (for example, time of harvesting or region of origin). Although variables containing data so affected may have a value for w of less than 1, they may still be having a great influence on the model by causing oils of, say, a similar harvesting date, to cluster together, or at least to be pulled away from their hoped-for varietal clustering, in a PCA scores plot. In order to eliminate this effect and to increase the parsimony of the model, we would wish to discard many variables with a value of w less than 1. We cannot easily know which variables are affected most in this way, and the optimal model complexity is not fixed for a given problem; it is data set dependent (de Noord, 1994). One solution is to start from a minimum number of variables (two or three), selected with a low threshold value of weighted or unweighted w, and work upwards towards $w = 1$ to see at what point the best model is reached.

It could be expected, when trying to identify varieties, that the optimum model will be achieved at or near a value for w which selects variables that

contribute to the discrimination of varieties but does not select any that also contribute a large amount to the discrimination of other factors (such as region or harvesting date). In practice it is found that the ideal threshold ranges from that which selects only the best three or four variables right up to $w = 1$, depending on the data and the desired factor (whether variety, region, etc.). Undoubtedly, this is at least in part arising from the principle of parsimony.

10.4.12 Exploitation of multivariate spectroscopies in the identification of the geographical origin of olive oils

For the next olive oil harvest (1996/97 season) the Italian oil producers will provide the consumers with DOC (*Denominazione di Origine Controllata*) or CBO (Certified Brands of Origin) virgin olive oils. [The so-called DOC (Controlled Denomination of Origin) classification of extra virgin olive oil is being introduced in Italy according to law 169/1992.] Similar provisions can be found in EU regulations 2081/1992 and 2082/1992, governing the DOP [*Denominazione di Origine Proteggètta* (Protected Denomination of Origin)] for agricultural food products. These highly priced oils will be obtained from either a single (or a high percentage content from a single) variety or from several varieties growing in a specified region. Clearly, to enforce the legal situation suitable methods for determining the geographical origins of an extra virgin olive oil will be essential.

The methods used to attain the correct identification of olive oils will also have to take into account the potential large variability arising from variety, location and environmental differences in the compositional characteristics of 'pure' virgin olive oils; thus the availability of reliable methods for authentication of the geographical origin of the oils will be crucial. As discussed above, one possible method is Curie-point PyMS (Section 10.3.2) combined with a powerful multivariate or chemometric analysis technique such as ANNs (Section 10.4.9) (Fig. 10.14).

The use of variable selection has been shown to improve the prediction of the variety and region of origin of olive oils considerably, using both ^{13}C NMR (Shaw *et al.*, 1996, 1997) and PyMS data. The results that are shown here were produced with use of PLS and PCR software written in-house by Jones, in conjunction with Microsoft Excel 5 macros written by Shaw.

The NMR data used consisted of five varieties; four Italian: Coratina (14), Dritta (12), I-77 (16) and Moraiolo (16) and one Israeli (8). The samples were run in duplicate to ensure reproducibility. In such circumstances it is important to ensure that the duplicates are kept together, not split between test and training sets. The Dritta oil, however, was all from one sample and so was unavoidably split between the test and training sets. The regions used were from the same data, being a mixture of varieties from Abruzzo (12),

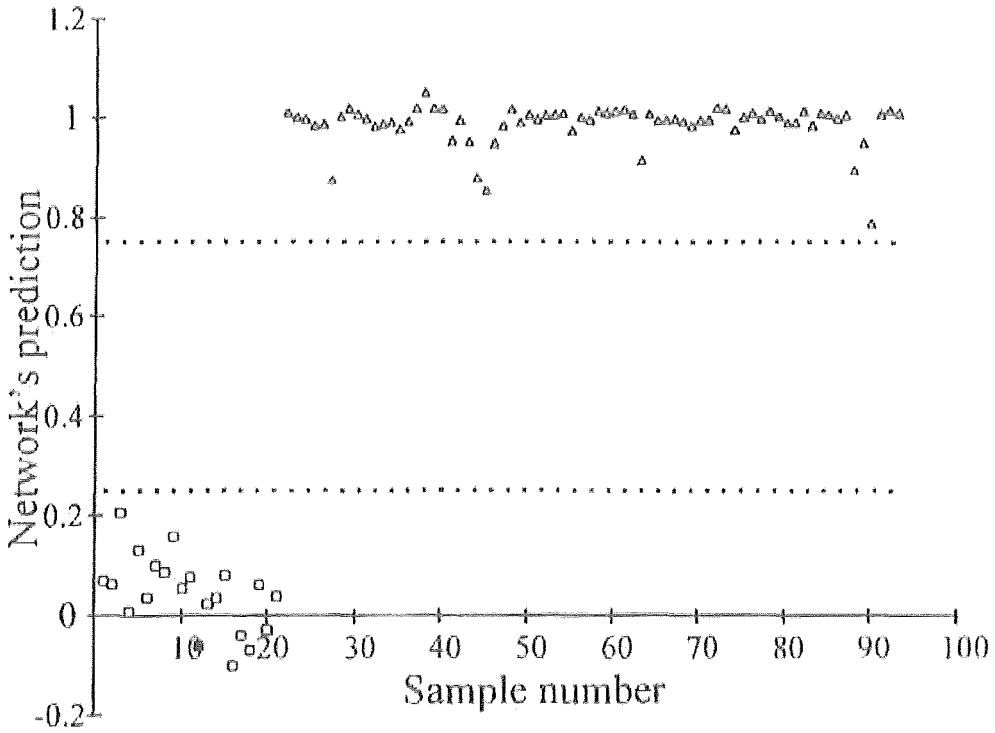

Figure 10.14 Prediction of the region of origin for extra virgin olive oils by means of artificial neural networks (ANNs). Raw data were separated into training and test sets by means of the Duplex partitioning method. A three-layered ANN [150 input nodes (±10% headroom), 8 hidden nodes and a single output node] was trained with 183 objects (39 Abruzzo, coded 0, and 144 Sardinia, coded 1) and interrogated periodically using 93 separate objects [21 Abruzzo (□) and 72 Sardinia (△)] previously unseen by the training net. The output after 110 000 epochs is shown above, with a training error of 0.009 root mean square. The lines represent network estimates of 0.25 and 0.75. Squares below the 0.25 line are deemed as being Abruzzo, and triangles above the 0.75 line as Sardinia. In this strict test all 93 are correctly identified.

Puglia (14), Toscana (12), Israel (8) and the 12 measurements of the Dritta sample, which was also from Abruzzo.

Figure 10.15(a) shows how only the best six or seven variables allow a 100% prediction of variety I-77, whereas if all variables are used only around 85% can be achieved. Figure 10.15(b) shows that similar success can be achieved with regions, again with less than half the variables; region Toscana has proved to be easier to predict than other regions, which is fortuitous as it is these oils that are the most highly prized by connoisseurs. For these two examples, one Y variable was used, a 1 being used to represent the variety being predicted, and a 0 all other varieties.

The remaining examples are predictions of five varieties simultaneously, rather than just one. The output matrix, Y, was therefore encoded by using five variables of 1s and 0s, using a 1 to represent each variety in the appropriate column.

MLR performs significantly worse than PLS or PCR on the data used; as has been pointed out by Martens and Næs (1989), the MLR predictor has seriously deficient performance where there is collinearity in the X data.

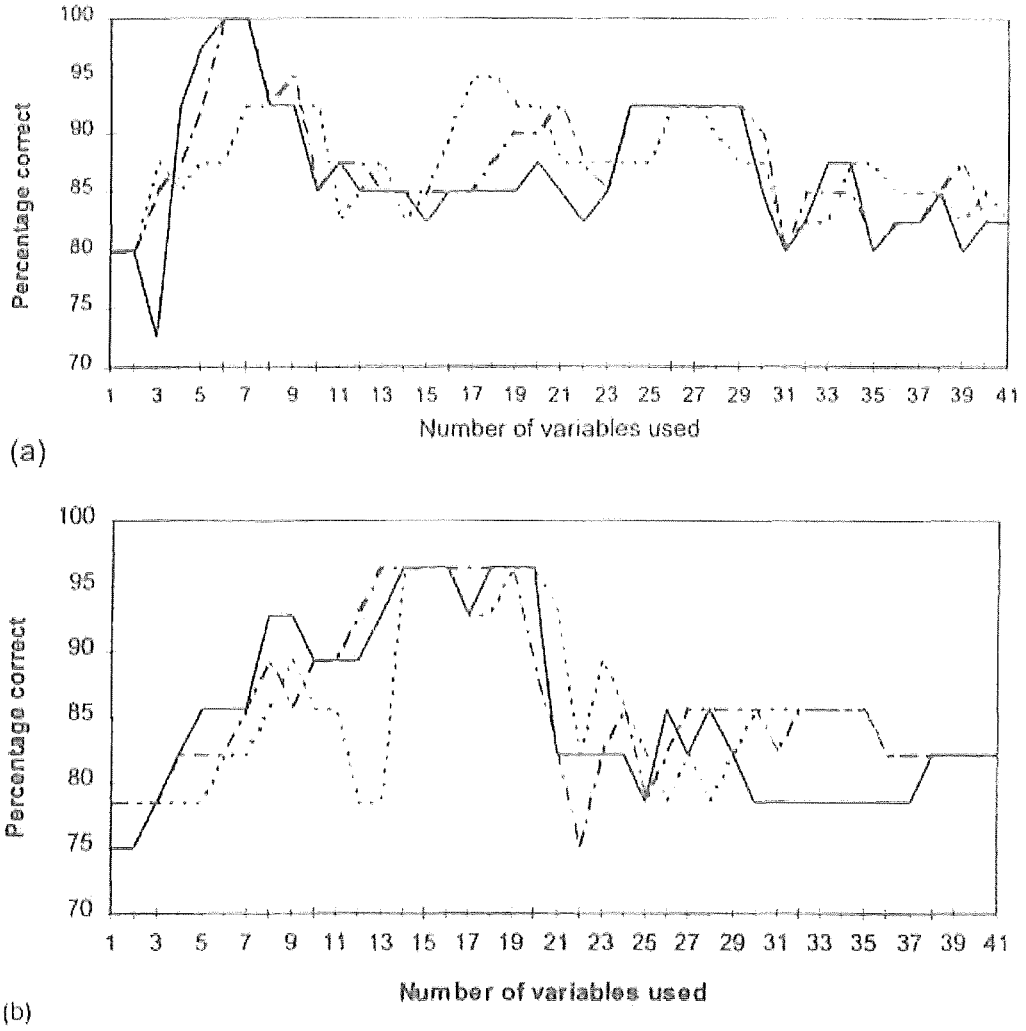

Figure 10.15 PLS 1 (partial least squares) prediction of: (a) olive oil variety I–77 from ^{13}C NMR data containing five varieties (all oils were correctly predicted with Fisher and weighted w selection using the best six or seven variables); (b) olive oil region of origin Toscana (Tuscany) from ^{13}C NMR data containing four regions (at best, all but one are correctly predicted). Selection procedures: —— = Fisher; - - - = weighted w; - - - - = unweighted w. Selection procedures are described in text, Section 10.4.11, equations (10.31)–(10.35).

Variable selection, however, greatly improves the prediction obtained by MLR, as can be seen in Fig. 10.16. Without variable selection, these results would be worse than a random guess!

The same data, using PLS 2 rather than MLR, gives a better prediction (Fig. 10.17), using weighted and unweighted w selection methods. Although the best prediction using weighted w (91.6%, all but three) is achieved with all but one of the variables, the inclusion of that one variable significantly reducing the prediction, to 75% (all but eight).

Since we know to what most of the carbon signals correspond, it is possible to draw some conclusions on the chemical significance of the order of variable selection.

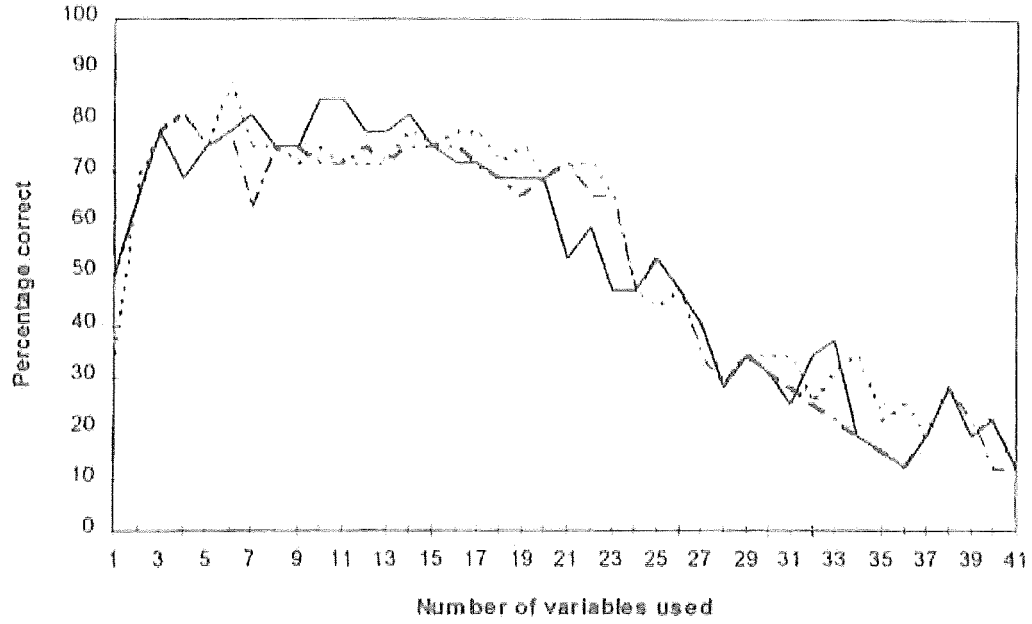

Figure 10.16 Multiple linear regression prediction of olive oil variety from ^{13}C NMR data containing five varieties (at best, all but four predictions are correct with selection by unweighted w). Selection procedures: — = Fisher; - - - - = weighted w; · · · · = unweighted w. Selection procedures are described in text, Section 10.4.11, equations (10.31)–(10.35).

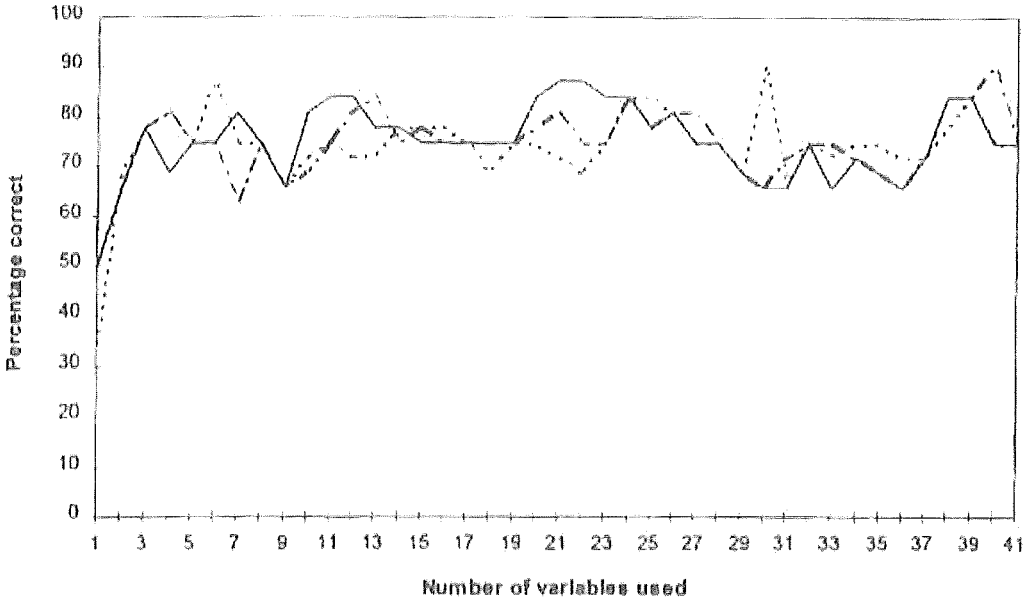

Figure 10.17 PLS 2 (partial least squares) prediction of olive oil variety from ^{13}C NMR data containing five varietes (at best, all but three predictions were correct with use of weighted and unweighted w). Selection procedures: — = Fisher; - - - - = weighted w; · · · · = unweighted w. Selection procedures are described in text, Section 10.4.11, equations (10.31)–(10.35).

For region discrimination the most important variables are found in the aliphatic region of the spectrum, from both oleic and saturated chains [Fig. 10.18(a)]. The fatty acids corresponding to the most significant 10 variables are identified. All the assignments are either for the α or the $\alpha+\beta$ positions of the glycerol backbone. Since there are two α positions for each β position it seems likely that it is the strength of the signal (strong signals probably contain less noise) which causes those variables corresponding to the α position to be selected above those corresponding to the β position.

For variety discrimination, the carbonyl region is much more prominent, but again oleic acid predominates [Fig. 10.18(b)]. The most significant finding here is that position of oleic acid (α or β) is not indicated by any of the 10 most significant signals in the aliphatic region, but is indicated by CS_2 and CS_4 in the carbonyl region, as is the position of linoleic with CS_5. Therefore, it would appear that variety discrimination is aided by the knowledge of the position of monounsaturated fatty acids on the glycerol backbone, whereas this is not so important for region discrimination. It follows, then, that one of the main distinguishing features of olive oil varieties is the position of the monounsaturated fatty acids on the glycerol backbone, whereas for regions the main factor is the relative proportions of fatty acids in the oil. It is unfortunate that the assignment of CS_22 is unknown, as this is important for both variety and region discrimination.

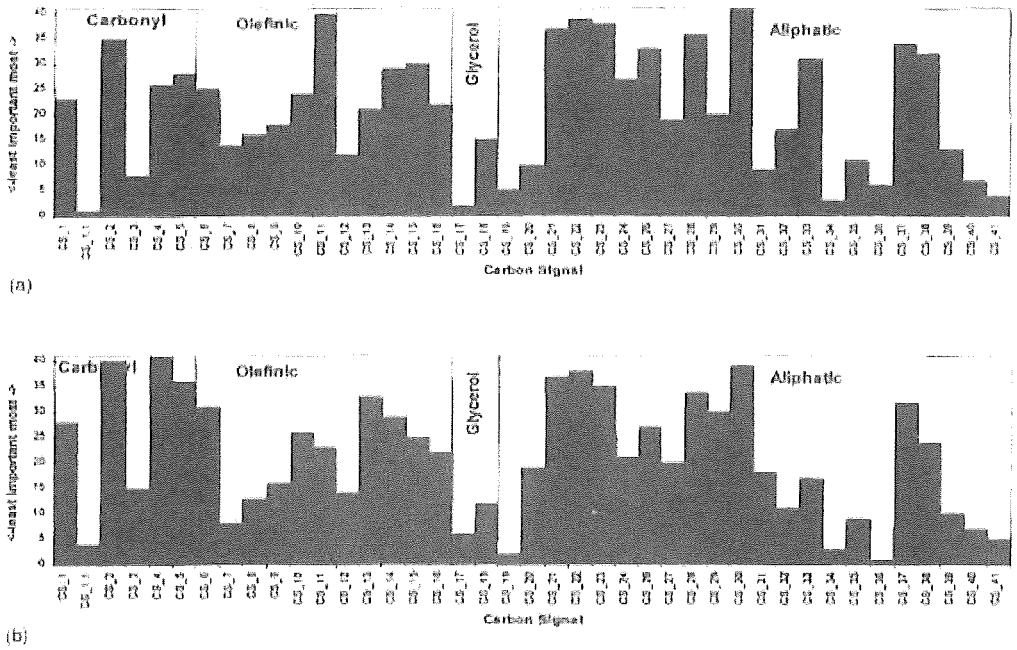

Figure 10.18 Relative significance of carbon signals (CS) for (a) region discrimination; (b) variety discrimination. Assignments are $\alpha+\beta$ except where indicated. O = oleic; S = saturated; L = linoleic.

10.5 Concluding remarks and future prospects

For reasons of space we have been unable to cover all of the major chromatographic and spectroscopic methods which might be applied to olive oils and other lipids or other chemometric methods which might be used to turn the complex, multivariate data so obtained into information. In particular, we recognize the immense potential of vibrational spectroscopies (Kemsley, Appleton and Wilson, 1994; Winson *et al.*, 1997a, b) such as Fourier transform infra-red, near infra-red and Raman (Section 10.2.4). Although the chemometric methods we have described can make excellent predictive models the non-linear ones in particular do not easily lend themselves to explaining how they do it. This is known as the assignment problem. By contrast, rule-based methods such as classification and regression trees (Breiman *et al.*, 1984) and fuzzy multivariate rule induction (Alsberg *et al.*, 1997) can provide models which are much easier to interpret according to the IF . . . THEN rules which they can produce. The production of simpler rules and models is contingent on the extraction of appropriate features in a pre-processing step; to this end we have found a peak parameter representation to be of value (Alsberg, Winson and Kell, 1997), and less familiar statistical and multivariate methods that may prove to be of particular use in the extraction of relevant information from such spectra include wavelet analysis and regression, Fourier regression, Fourier deconvolution and B-spline splitting (Alsberg, Woodward and Kell, 1997). Finally, genetic programming methods (Koza, 1992, 1994a, b), although computationally intensive, have the potential to search the high-dimensional space with high efficiency (Edmonds, Burkhardt and Adjei, 1995) to find the models which best relate the spectroscopic data available to the biological or chemical properties of interest (Gilbert *et al.*, 1997).

Acknowledgements

We thank the UK BBSRC, the Higher Education Funding Council for Wales and the Ministry of Agriculture, Fisheries and Foods for financial support. Thanks are also due to the Italian Ministry of Agriculture.

References

Afifi, A. A. and Clark, V. (1996) *Computer-aided Multivariate Analysis*, 3rd edn, Chapman & Hall, London.

Akitt, J. W. (1983) *NMR and Chemistry: An Introduction to the Fourier Transform–Multinuclear Era*, 2nd edn, Chapham & Hall, London.

Aldridge, W. N. (1992) The Toxic Oil Syndrome (TOS, 1981): from the disease towards a toxological understanding of its chemical aetiology and mechanism. *Toxicology Letters*, **64–65**, 59–70.

Aleksander, I. (1989) *Neural Computing Architectures: the Design of Brain-like Machines*, MIT Press, Cambridge, MA.

Allen, D. M. (1971) Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13**, 469–75.

Alsberg, B. K., Winson, M. K. and Kell, D. B. (1997) Improving interpretation of multivariate and rule induction models by using a spectral peak parameter representation. *Chemometrics and Intelligent Laboratory Systems*, **36**, 95–109.

Alsberg, B. K., Woodward, A. M. and Kell, D. B. (1997) An introduction to wavelet transforms for chemometricians: a time-frequency approach. *Chemometrics and Intelligent Laboratory Systems*, **37**, 215–39.

Alsberg, B. K., Goodacre, R., Rowland, J. J. and Kell, D. B. (1997) Classification of pyrolysis mass spectra by fuzzy multivariate rule induction; comparison with regression, K-nearest neighbour, neural and decision-tree methods. *Analytica Chimica Acta*, in press.

Anon. (1994) Per l'olio di oliva fine delle frodi? *L'Informatore Agrario*, **18**, 41.

Aparicio, R. and Alonso, V. (1994) Characterization of virgin olive oils by SEXIA expert system. *Progress in Lipid Research*, **33**, 29–38.

Aparicio, R. and Morales, M. T. (1995) Sensory wheels – a statistical technique for comparing QDA panels – application to virgin olive oil. *Journal of the Science of Food and Agriculture*, **67**, 247–57.

Aparicio, R., Alonso, V. and Morales, M. T. (1994). Detailed and exhaustive study of the authentication of European virgin olive oils by SEXIA expert-system. *Grasas y Aceites*, **45**, 241–52.

Aparicio, R., Alonso, V. and Morales, M. T. (1996) Developments in olive oil authentication, in *Food Authenticity '96*, Norwich.

Aparicio, R., Ferreiro, L. and Alonso, V. (1994) Effect of climate on the chemical-composition of virgin olive oil. *Analytica Chimica Acta*, **292**, 235–41.

Aparicio, R., Gutierrez, F. and Morales, J. R. (1992) Relationship between flavor descriptors and overall grading of analytical panels for virgin olive oil. *Journal of the Science of Food and Agriculture*, **58**, 555–62.

Aparicio, R., Navarro, M. S. and Ferreiro, M. S. (1991) Definite influence of the extraction methods on the chemical composition of virgin olive oil. *Grasas y Aceites*, **42**, 356–62.

Armanino, C., Leardi, R. and Lanteri, S. (1989) Chemometric analysis of Tuscan olive oils. *Chemometrics and Intelligent Laboratory Systems*, **5**, 343–54.

Aylott, R. I., Clyne, A. H., Fox, A. P. and Walker, D. A. (1994). Analytical strategies to confirm Scotch whisky authenticity. *Analyst*, **119**, 1741–6.

Baeten, V., Meurens, M., Morales, M. T. and Aparicio, R. (1996) Detection of virgin olive oil adulteration by Fourier transform Raman spectroscopy. *Journal of Agricultural and Food Chem* **44**, 2225–30.

Bangalore, A. S., Shaffer, R. E., Small, G. W. and Arnold, M. (1996) Genetic algorithm-based method for selecting wavelengths and model size for use with partial least squares regression. Application to near infrared spectroscopy. *Analytical Chemistry*, **68**, 4200–12.

Baroni, M., Clementi, S., Cruciani, G. *et al.* (1992) Predictive ability of regression-models. 2. Selection of the best predictive PLS model. *Journal of Chemometrics*, **6**, 347–56.

Battiti, R. (1994) Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions Neural Networks*, **5**, 537–50.

Beale, R. and Jackson, T. (1990) *Neural Computing: An Introduction*, IOP, Bristol.

Belton, P. S. (1995) NMR in context. *Annual Reports on NMR Spectroscopy*, **31**, 3–18.

Bhandare, P., Mendelson, Y., Peura, R. A. *et al.* (1993) Multivariate determination of glucose in whole blood using partial least-squares and artificial neural networks based on mid-infrared spectroscopy. *Applied Spectroscopy*, **47**, 1214–21.

Bianchi, G., Giansante, L. and Lazzari, M. (1996) Analisi per la tutela di genuinità, origine geografica e varietale degli oli vegetali. *L'Informatore Agrario*, **20**, 45–8.

Bianchi, G., Angerosa, F., Camera, L. *et al.* (1993) Stable carbon isotope ratios (13C/12C) of olive oil components. *Journal of Agricultural and Food Chemistry*, **41**, 1936–40.

Bianchi, G., Gussoni, M., Limiroli, R. *et al.* (1994a) NMR and chemical studies of the morphologically different parts of the olive fruit (*Olea Europaea* L.). *Acta Horticulturae*, **356**, 260–3.

Bianchi, G., Tava, A., Vlahov, G. and Pozzi, N. (1994b) Chemical-structure of long-chain esters from sansa olive oil. *Journal Of the American Oil Chemists' Society*, **71**, 365–9.

Bishop, C. M. (1995) *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.

Bookstein, F. L. (1980) Data analysis by partial least squares, in *Evaluation of Econometric Models* (eds J. Kmenta and J. B. Ramsey), Academic Press, London, pp. 75–90.

Bosaeus, I., Belfrage, L., Lindgren, C. and Andersson, H. (1992) Olive oil instead of butter increases net cholesterol excretion from the smallbowel. *European Journal of Clinical Nutrition*, **46**, 111–15.

Boschelle, O., Giomo, A., Conte, L. and Lercker, G. (1994) Caratterizzazione della cultivar di olivo del Golfo di Trieste mediante metodi chemiometrici applicati ai dati chimica-fisici. *La Rivista Italiana delle Sostanze Grasse*, **71**, 57–65.

Boskou, D. (1996) *Olive oil, chemistry and technology*, AOCS Press, Champaign, IL.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*, Wadsworth International, Belmont, CA.

Brekke, T., Barth, T., Kvalheim, O. M. and Sletten, E. (1990) Multivariate analysis of carbon-13 nuclear magnetic resonance spectra. Identification and quantification of average structures in petroleum distillates. *Analytical Chemistry*, **62**, 56–61.

Brereton, R. G. (1992) *Multivariate Pattern Recognition in Chemometrics*, Elsevier, Amsterdam.

Brereton, R. G. (1995) Deconvolution of mixtures by factor-analysis. *Analyst*, **120**, 2313–36.

Brereton, R. G. and Elbergali, A. K. (1994) Use of double windowing, variable selection, variable ranking and resolvability indices in window factor analysis. *Journal of Chemometrics*, **8**, 423–37.

Brevard, C. and Grainger, P. (1981) *Handbook of High Resolution NMR*, John Wiley, New York.

Broadhurst, D., Goodacre, R., Jones, A. *et al.* (1997) Genetic algorithms as a method for variable selection in MLR and PLS regression, with applications to pyrolysis mass spectrometry. Submitted for publication in *Analytica Chimica Acta*.

Brown, P. J. (1993) *Measurement, Regression, and Calibration*, Oxford Science Publications, Oxford.

Brown, S. D., Sum, S. T., Despagne, F. and Lavine, B. K. (1996) Chemometrics. *Analytical Chemistry*, **68**, R21–R61.

Campbell, I. D. and Dwek, R. A. (1984) *Biological Spectroscopy*, Benjamin Cummings, London.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society*, **158**, 419–66.

Cheng, B. and Titterington, D. M. (1994) Neural networks: a review from a statistical perspective. *Statistical Science*, **9**, 2–54.

Clint, M. and Jennings, A. (1970) The evaluation of eigenvectors of real symmetric matrices by simultaneous iteration. *The Computer Journal*, **13**, 76–80.

Collins, E. J. T. (1993) Food adulteration and food safety in Britain in the 19th and 20th centuries. *Food Policy*, (April), 95–109.

Cruciani, G. and Watson, K. A. (1994) Comparative molecular-field analysis using grid force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase B. *Journal of Medicinal Chemistry*, **37**, 2589–601.

Davies, A. M. C. (1995) The better way of doing principal component regression. *Spectroscopy Europe*, **7**, 36–8.

Defalguerolles, A. and Jmel, S. (1993) Variable selection criteria – based on specific Gaussian graphical models in principal components analysis. *Canadian Journal of Statistics – Revue Canadienne de Statistique*, **21**, 239–56.

de Jong, S. (1993a) PLS fits closer than PCR. *Journal of Chemometrics*, **7**, 551–7.

de Jong, S. (1993b) SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **18**, 251–63.

de Noord, O. E. (1994) The influence of data preprocessing on the robustness and parsimony of multivariate calibration models. *Chemometrics and Intelligent Laboratory Systems*, **23**, 65–70.

EC (1991) Commission Regulation (EEC) no. 2568/91 of 11 July 1991 on the characteristics of olive oil and olive-residue oil and on the relevant methods of analysis. *Official Journal of the European Communities* **L248**, 1–83.

Edmonds, A. N., Burkhardt, D. and Adjei, O. (1995) Genetic programming of fuzzy logic production rules, in *IEEE International Conference on Evolutionary Computation*, vols 1–2, IEEE, Perth, pp. 765–70.

Eshuis, W., Kistemaker, P. G. and Meuzelaar, H. L. C. (1977) Some numerical aspects of reproducibility and specificity, in *Analytical Pyrolysis* (eds C. E. R. Jones and C. A. Cramers), Elsevier, Amsterdam, pp. 151–6.

Firestone, D., Carson, K. L. and Reina, R. J. (1988) Update on control of olive oil adulteration and misbranding in the United-States. *Journal of the American Oil Chemists Society*, **65**, 788–92.

Firestone, D. and Reina, R. J. (1987) Update on control of olive oil adulteration in the United-States. *Journal of the American Oil Chemists' Society*, **64**, 682–82.

Firestone, D., Summers, J. L., Reina, R. J. and Adams, W. S. (1985) Detection of adulterated and misbranded olive oil products. *Journal of the American Oil Chemists' Society*, **62**, 1558–62.

Flury, B. and Riedwyl, H. (1988) *Multivariate Statistics: A Practical Approach*, Chapman & Hall, London.

Forina, M. and Tiscornia, E. (1982) Pattern recognition methods in the prediction of Italian olive oil origin by their fatty acid content. *Annali di Chimica*, **72**, 143–55.

Francelin, R. A., Gomide, F. A. C. and Lanças, F. M. (1993) Use of artificial neural networks for the classification of vegetable oils after GC analysis. *Chromatographia*, **35**, 160–6.

Frank, I. E. and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–35.

Fraser, G. E. (1994) Diet and coronary heart disease: beyond dietary fats and low-density-lipoprotein cholesterol. *American Journal of Clinical Nutrition*, **59**, S1117–23.

Friebolin, H. (1993) *Basic One- and Two-dimensional NMR Spectroscopy*, 2nd edn. VCH, Weinheim.

Galli, C., Petroni, A. and Visioli, F. (1994) Natural antioxidants, with special reference to those in olive oil, and cell protection. *European Journal of Pharmaceutical Sciences*, **2**, 67–8.

Garcia, M. V. A. and López, R. A. (1993) Characterization of European virgin olive oils using fatty-acids. *Grasas y Aceites*, **44**, 18–24.

Garcia, J. M., Gutiérrez, F., Castellano, J. M. *et al.* (1996) Influence of storage temperature on fruit ripening and olive oil quality. *Journal of Agricultural and Food Chemistry*, **44**, 264–7.

Geisser, S. (1975) The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320–8.

Gigliotti, C., Daghetta, A. and Sidoli, A. (1994) Caratterizzazione geografica e merceologica di oli di oliva mediante valutazione della composizone trigliceridica per HPLC. *La Rivista Italiana delle Sostanze Grasse*, **71**, 51–6.

Gilbert, R. J., Goodacre, R., Woodward, A. M. and Kell, D. B. (1997) Genetic programming, a novel method for the quantitative analysis of pyrolysis mass spectral data. *Analytical Chemistry*, in press.

Goodacre, R. (1994a) Characterization and quantification of microbial systems using pyrolysis mass spectrometry: introducing neural networks to analytical pyrolysis. *Microbiology Europe*, **2**, 16–22.

Goodacre, R. (1994b) Characterization and quantification of microbial systems using pyrolysis mass spectrometry: introducing neural networks to analytical pyrolysis. *Microbiology Europe*, **2**, 16–22.

Goodacre, R. and Kell, D. B. (1996) Pyrolysis mass spectrometry and its applications in biotechnology. *Current Opinion in Biotechnology*, **7**, 20–8.

Goodacre, R., Hammond, D. and Kell, D. B. (1997) Quantitative analysis of the adulteration of orange juice with sucrose using pyrolysis mass spectrometry and chemometrics. *J. Anal. Appl. Pyrol.*, **40/41**, 135–58.

Goodacre, R., Howell, S. A., Noble, W. C. and Neal, M. J. (1996) Sub-species discrimination, using pyrolysis mass spectrometry and self-organising neural networks, of *Propionibacterium acnes* isolates from normal human skin. *Zentralblatt für Bakteriologie*, **284**, 501–15.

Goodacre, R., Kell, D. B. and Bianchi, G. (1992) Neural networks and olive oil. *Nature*, **359**, 594–594.

Goodacre, R., Kell, D. B. and Bianchi, G. (1993) Rapid assessment of the adulteration of virgin olive oils by other seed oils using pyrolysis mass-spectrometry and artificial neural networks. *Journal of the Science of Food and Agriculture*, **63**, 297–307.

Goodacre, R., Neal, M. J. and Kell, D. B. (1994) Rapid and quantitative analysis of the pyrolysis mass spectra of complex binary and tertiary mixtures using multivariate calibration and artificial neural networks. *Analytical Chemistry*, **66**, 1070–85.

Goodacre, R., Neal, M. J. and Kell, D. B. (1996) Quantitative analysis of multivariate data using artificial neural networks: a tutorial review and applications to the deconvolution of pyrolysis mass spectra. *Zentralblatt für Bakteriologie*, **284**, 516–39.

Goodacre, R., Neal, M. J., Kell, D. B. *et al.* (1994) Rapid identification using pyrolysis mass spectrometry and artificial neural networks of *Propionibacterium acnes* isolated from dogs. *Journal of Applied Bacteriology*, **76**, 124–34.

Goodacre, R., Trew, S., Wrigley-Jones, C. *et al.* (1995) Rapid and quantitative analysis of metabolites in fermentor broths using pyrolysis mass spectrometry with supervised learning: application to the screening of *Penicillium chryosgenum* fermentations for the overproduction of penicillins. *Analytica Chimica Acta*, **313**, 25–43.

Gourlay, A. R. and Watson, G. A. (1973) *Computational Methods for Matrix Eigenproblems*. John Wiley, Chichester, Sussex.

Grob, K., Biedermann, M., Bronz, M. and Schmid, J. P. (1994a) Recognition of mild deodorization of edible oils by the loss of volatile components. *Zeitschrift für Lebensmittel-Untersuchung und -Forschung*, **199**, 191–4.

Grob, K., Giuffré, A. M., Leuzzi, U. and Mincione, B. (1994b) Recognition of adulterated oils by direct analysis of the minor components. *Fat Science Technology*, **96**, 286–90.

Guinda, A., Lanzón, A. and Albi, T. (1996) Differences in hydrocarbons of virgin olive oils obtained from several olive varieties. *Journal of Agricultural and Food Chemistry*, **44**, 1723–6.

Gussoni, M., Greco, F., Consonni, R. *et al.* (1993) Application of NMR microscopy to the histochemistry study of olives (*Olea Europaea* L.). *Magnetic Resonance Imaging*, **11**, 259–68.

Gussow, J. D. (1995) Mediterranean diets: are they environmentally responsible? *American Journal of Clinical Nutrition*, **61** (supplement), 1383S–9S.

Haaland, D. M. and Thomas, E. V. (1988a). Partial least squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry*, **60**, 1193–202.

Haaland, D. M. and Thomas, E. V. (1988b) Partial least squares methods for spectral analyses. 2. Application to simulated and glass spectral data. *Analytical Chemistry*, **60**, 1202–8.

Harris, R. K. (1986) *Nuclear Magnetic Resonance Spectroscopy*. Longman Scientific and Technical, Harlow, Essex.

Haumann, B. F. (1996) Olive oil: Mediterranean product. *Inform*, **7**, 890–903.

Hazen, K. H., Arnold, M. A. and Small, G. W. (1994) Temperature-insensitive near-infrared spectroscopic measurement of glucose in aqueous solutions. *Applied Spectroscopy*, **48**, 477–83.

Hecht-Nielsen, R. (1989) *Neurocomputing*, Addison-Wesley, Reading, MA.

Heikka, R., Minkkinen, P. and Taavitsainen, V. M. (1994) Comparison of variable selection and regression methods in multivariate calibration of a process analyzer. *Process Control and Quality*, **6**, 47–54.

Helland, I. S. (1988) On the structure of partial least squares regression. *Communications on Statistical Simulations*, **17**, 581–607.

Hertz, J., Krogh, A. and Palmer, R. G. (1991) *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA.

Hoerl, A. E. and Kennard, R. W. (1970a) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoerl, A. E. and Kennard, R. W. (1970b) Ridge regression: application to nonorthogonal problems. *Technometrics*, **12**, 69–82.

Horchner, U. and Kalivas, J. H. (1995) Further investigation on a comparative-study of simulated annealing and genetic algorithm for wavelength selection. *Analytica Chimica Acta*, **311**, 1–13.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417–41, 498–520.

IFR (1994) *Annual Report 1994*, Institute of Food Research.

Irwin, W. J. (1982) *Analytical Pyrolysis: A Comprehensive Guide*, Marcel Dekker, New York.

Ismail, A. A., van de Voort, F. R., Emo, G. and Sedman, J. (1993) Rapid quantitative determination of free fatty acids in fats and oils by Fourier transform infrared spectroscopy. *Journal of the American Oil Chemists' Society*, **70**, 335–41.

Jolliffe, I. T. (1982) A note on the use of principal components in regression. *Applied Statistics*, **31**, 300–3.

Jolliffe, I. T. (1986) *Principal Component Analysis*, Springer, Berlin.

Jouan-Rimbaud, D., Massart, D. L., Leardi, R. and de Noord, O. E. (1995) Genetic algorithms as a tool for wavelength selection in multivariate calibration. *Analytical Chemistry*, **67**, 4295–301.

Judd, J. S. (1990) *Neural Network Design and the Complexity of Learning*, MIT Press, Cambridge, MA.

Kafatos, A. and Comas, G. (1991) Biological effects of olive oil on human health, in *Olive Oil* (ed. A. K. Kiritsakis), American Oil Chemists' Society, Champaign, IL, pp. 157–81.

Kajioka, R. and Tang, P. W. (1984) Curie-point mass spectrometry of *Legionella* species. *Journal of Applied and Analytical Pyrolysis*, **6**, 59–68.

Kemp, W. (1986) *NMR in Chemistry: A Multinuclear Introduction*, Macmillan Education, London.

Kemsley, E. K., Appleton, G. P. and Wilson, R. H. (1994) Quantitative-analysis of emulsions using attenuated total reflectance (ATR). *Spectrochimica Acta Part a – Molecular Spectroscopy*, **50**, 1235–42.

Kiritsakis, A. K. (1984) Effect of selected storage conditions and packaging materials on olive oil quality. *Journal of the American Oil Chemists' Society*, **61**, 1868–70.

Kiritsakis, A. K. (1991) *Olive oil*, AOCS, Champaign, IL.

Kiritsakis, A. and Dugan, L. R. (1985) Studies in photoxidation of olive oil. *Journal of the American Oil Chemists' Society*, **62**, 892–6.

Kiritsakis, A. and Markakis, P. (1991) Olive oil analysis, in *Essential Oils and Waxes* (eds. H. E. Linskens and J. F. Jackson), Springer, Berlin, pp. 1–20.

Koza, J. R. (1992) *Genetic Programming: On The Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA.

Koza, J. R. (1994a) Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, **4**, 87–112.

Koza, J. R. (1994b) *Genetic Programming II: Automatic Discovery of Reusable Programs*. MIT Press, Cambridge, MA.

Kubinyi, H. (1994a) Variable selection in QSAR studies. I. An evolutionary algorithm. *Quantitative Structure–Activity Relationships*, **13**, 285–94.

Kubinyi, H. (1994b) Variable selection in QSAR studies. 2. A highly efficient combination of systematic search and evolution. *Quantitative Structure–Activity Relationships*, **13**, 393–401.

Kubinyi, H. (1996) Evolutionary variable selection in regression and PLS analyses. *Journal of Chemometrics*, **10**, 119–33.

Kvalheim, O. M., Aksnes, D. W., Brekke, T. *et al.* (1985) Crude oil characterization and correlation by principal component analysis of ^{13}C nuclear magnetic resonance spectra. *Analytical Chemistry*, **57**, 2858–64.

Lai, Y. W., Kemsley, E. K. and Wilson, R. H. (1994) Potential of Fourier transform-infrared spectroscopy for the authentication of vegetable-oils. *Journal of Agricultural and Food Chemistry*, **42**, 1154–9.

Lai, Y. W., Kemsley, E. K. and Wilson, R. H. (1995) Quantitative-analysis of potential adulterants of extra virgin olive oil using infrared-spectroscopy. *Food Chemistry*, **53**, 95–8.

Li-Chan, E. (1994) Developments in the detection of adulteration of olive oil. *Trends in Food Science and Technology*, **5**, 3–11.

Lindgren, F., Geladi, P., Berglund, A. *et al.* (1995) Interactive variable selection (IVS) for PLS2. Chemical applications. *Journal of Chemometrics*, **9**, 331–42.

Lindgren, F., Geladi, P. and Wold, S. (1993) The Kernel algorithm for PLS. *Journal of Chemometrics*, **7**, 45–59.

Linos, A., Kaklamanis, E., Kontomerkos, A. *et al.* (1991) The effect of olive oil and fish consumption on rheumatoid-arthritis – a case control study. *Scandinavian Journal of Rheumatology*, **20**, 419–26.

Luger, G. F. and Stubblefield, W. A. (1989) *Artificial Intelligence and the Design of Expert Systems*, Benjamin Cummings, Redwood City, CA.

Lyon, D. H. and Watson, M. P. (1994) Sensory profiling – a method for describing the sensory characteristics of virgin olive oil. *Grasas y Aceites*, **45**, 20–5.

Lyttkens, E. (1966) On the fix-point property of Wold's iterative estimation method for principal components, in *Multivariate Analysis* (ed. K. R. Krishnaiah), Academic Press, New York, pp. 335–50.

Lyttkens, E. (1973) The fix-point method for estimating interdependent systems with the underlying model specification. *Journal of the Royal Statistical Society, Series A*, **135**, 353–94.

MAFF (1995) *Manual of Nutrition: Reference Book 342*, 10th edn, Ministry of Agriculture, Fisheries and Food, The Stationery Office, London.

Malpass, J. A., Salt, D. W., Ford, M. G. *et al.*, (1994) Continuum regression: a new algorithm for the prediction of biological activity, in *Advanced Computer-Assisted Techniques in Drug Discovery* (ed. H. van der Waterbeemd), VCH, Weinheim, pp. 163–89.

Malpass, J. A., Salt, D. W., Wynn, E. W. *et al.* (1995) Prediction of biological activity using continuum regression, in *Trends in QSAR and Molecular Modelling 92* (ed. C. G. Wermuth), ESCOM, pp. 314–16.

Mark, H. (1991) *Principles and Practice of Spectroscopic Calibration*, John Wiley, New York.

Martens, H. and Næs, T. (1989) *Multivariate Calibration*, John Wiley, Chichester, Sussex.

Martín-Moreno, J. M., Willett, W. C., Gorgojo, L. *et al.* (1994) Dietry fat, olive oil intake and breast cancer risk. *International Journal of Cancer*, **58**, 774–80.

Massy, W. F. (1965) Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, **60**, 234–56.

Meuzelaar, H. L. C., Haverkamp, J. and Hileman, F. D. (1982) *Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials*, Elsevier, Amsterdam.

Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (1994) Machine learning: neural and statistical classification, in *Ellis Horwood Series in Artificial Intelligence* (ed. J. Campbell), Ellis Horwood, Chichester, Sussex.

Morales, M. T., Alonso, M. V., Rios, J. J. and Aparicio, R. (1995) Virgin olive oil aroma – relationship between volatile compounds and sensory attributes by chemometrics. *Journal of Agricultural and Food Chemistry*, **43**, 2925–31.

Morris, A. O. (1982) *Linear Algebra – An Introduction*, Van Nostrand Reinhold, London.

Mottram, R. F. (1979) *Human Nutrition*, 3rd edn, Edward Arnold, London.

Murphy, D. J. (1995) New oils for old. *Chemistry in Britain*, **31**, 300–2.

Norinder, U. (1996) Single and domain mode-variable selection in 3D QSAR applications. *Journal of Chemometrics*, **10**, 95–105.

Oman, S. D., Næs, T. and Zube, A. (1993) Detecting and adjusting for nonlinearities in calibration of near-infrared data using principal components. *Journal of Chemometrics*, **7**, 195–212.

Ozaki, Y., Cho, R., Ikegaya, K. *et al.*, (1992) Potential of near-infrared Fourier transform Raman spectroscopy in food analysis. *Applied Spectroscopy*, **46**, 1503–7.

Peri, C. and Rastelli, C. (1994) Implications for the future and recommendations for modifications to current regulations concerning virgin olive oil. *Grasas y Aceites*, **45**, 60–1.

Perrin, J.-L. (1992) Les composés mineurs et les antioxygènes naturels de l'olive et de son huile. *Revue Française des Corps Gras*, **39**, 25–32.

Rade, D., Strucelj, D., Mokrovcak, Z. and Hrboka, Z. (1995) Influence of olive storage and processing on some characteristics of olive oil. *Prehrambeno-Tehnoloska I Biotehnoloska Revija*, **33**, 119–22.

Ramos, L. S., Beebe, K. R., Carey, W. P. *et al.* (1986) Chemometrics. *Analytical Chemistry*, **58**, 294R–315R.

Ranalli, A. and Martinelli, N. (1994) Extraction of the oil from the olive pastes by biological and not conventional industrial technics. *Industrie Alimentari*, **33**, 1073–83.

Ripley, B. D. (1994) Neural networks and related methods for classification. *Journal of the Royal Statistical Society, Series B – Methodological*, **356**, 409–37.

Rosenblatt, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408.

Rumelhart, D. E. and McClelland, J. L. (1986) *Parallel Distributed Processing, Experiments in the Microstructure of Cognition*, MIT, Cambridge, MA.

Sadeghi-Jorabchi, H., Hendra, P. J., Wilson, R. H. and Belton, P. S. (1990) Determination of the total unsaturation in oils and margarines by Fourier-transform Raman spectroscopy. *Journal of the American Oil Chemists' Society*, **67**, 483–6.

Sadeghi-Jorabchi, H., Wilson, R. H., Belton, P. S. *et al.* (1991) Quantitative analysis of oils and fats by Fourier-transform Raman spectroscopy. *Spectrochimica Acta A*, **47**, 1449–58.

Salter, G. J., Lazzari, M., Giansante, L. *et al.* (1997) Determination of the geographical origin of Italian extra virgin olive oil using pyrolysis mass spectrometry and artificial neural networks. *Journal of Analytical Applied Pyrolysis*, **40/41**, 159–70.

Salunkhe, D. K., Chavan, J. K., Adsule, R. N. and Kadam, S. S. (1991) *World Oilseeds: Chemistry, Technology and Utilization*, Van Nostrand Reinhold, New York.

Sarle, W. S. (1994) Neural networks and statistical models, in *Nineteenth Annual SAS Users Group International Conference*,

Sato, T. (1994) Application of principal-component analysis on near-infrared spectroscopic data of vegetable-oils for their classification. *Journal of the American Oil Chemists' Society*, **71**, 293–8.

Schwaiger, I. and Vojir, F. (1994) Anwendung Multivariater Statistischer Verfahren zur Überprüfung der Authentizität von Speiseölen. *Deutsche Lebensmittel-Rundschau*, **90**, 143–6.

Seasholtz, M. B. and Kowalski, B. (1993) The parsimony principle applied to multivariate calibration. *Analytica Chimica Acta*, **277**, 165–77.

Segre, A. L., Mannina, L., Barone, P. and Sacchi, R. (1996) Quality and geographical origin of virgin olive oil as determined by high-field ^{1}H-NMR. *Bruker Report*, **143**, 27–8.

Shahidi, F. (1990) *Canola and Rapeseed: Production, Chemistry, Nutrition and Processing Technology*, Van Nostrand Reinhold, New York.

Shaw, A. D., di Camillo, A., Vlahov, G. *et al.* (1996) Discrimination of different olive oils using ^{13}C NMR and variable reduction, in *Food Authenticity '96*, Norwich.

Shaw, A. D., di Camillo, A., Vlahov, G. *et al.* (1997) Discrimination of the variety and region of origin of extra virgin olive oils using ^{13}C NMR and multivariate calibration with variable reduction. *Analytica Chimica Acta*, in press

Shepherd, J. and Packard, C. J. (1992) Atherosclerosis in perspective: the pathophysiology of human cholesterol metabolism, in *Human Nutrition: A Continuing Debate* (eds M. Eastwood, C. Edwards and D. Parry), Chapman & Hall, London, pp. 33–50.

Simpkins, W. and Harrison, M. (1995a) The state of the art in authenticity testing. *Trends in Food Science and Technology*, **6**, 321–8.

Simpkins, W. and Harrison, M. (1995b) The state of the art in authenticity testing. *Trends in Food Science and Technology*, **6**, 321–8.

Smith, M. (1993) *Neural Networks for Statistical Modeling*, Van Nostrand Reinhold, New York.

Snee, R. D. (1977) Validation of regression models: methods and examples. *Technometrics*, **19**, 415–28.

Sreerama, N. and Woody, R. W. (1994) Protein secondary structure from circular dichroism spectroscopy: combining variable selection principle and cluster analysis with neural network, ridge regression and self-consistent methods. *Journal of Molecular Biology*, **242**, 497–507.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, **36**, 111–33.

Stone, M. and Brooks, R. J. (1990) Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society, Series B*, **52**, 237–69.

Stryer, L. (1981) *Biochemistry*, 2nd edn, Freeman, San Francisco, CA.

Taavitsainen, V.-M. and Korhonen, P. (1992) Nonlinear data analysis with latent variables. *Chemometrics and Intelligent Laboratory Systems*, **14**, 185–94.

Trichopoulou, A. (1995) Olive oil and breast-cancer. *Cancer Causes and Control*, **6**, 475–6.

Trichopoulou, A., Gnardellis, C., Katsouyanni, K. *et al.* (1995a) Consumption of olive oil and specific food groups in relation to breast-cancer risk in Greece – response. *Journal of the National Cancer Institute*, **87**, 1022.

Trichopoulou, A., Katsouyanni, K., Stuver, S. *et al.* (1995b) Consumption of olive oil and specific food groups in relation to breast-cancer risk in Greece. *Journal of the National Cancer Institute*, **87**, 110–16.

Trichopoulou, A., Kouris-Blazos, A., Vassilakou, T. *et al.* (1995c) Diet and survival of elderly Greeks: a link to the past. *American Journal of Clinical Nutrition*, **61**(supplement), 1346s–50s.

Tsimidou, M. (1995) The use of HPLC in the quality control of virgin olive oil. *Chromatography and Analysis*, (Aug/Sept), 5–7.

Tsimidou, M. and Karakostas, K. X. (1993) Geographical classification of Greek virgin olive oil by nonparametric multivariate evaluation of fatty-acid composition. *Journal of the Science of Food and Agriculture*, **62**, 253–7.

van de Voort, F. R. (1994) FTIR spectroscopy in edible oil analysis. *INFORM*, **5**, 1038–42.

van de Voort, F. R., Ismail, A. A. and Sedman, J. (1995) A rapid, automated method for the determination of *cis* and *trans* content of fats and oils by Fourier-transform infrared-spectroscopy. *Journal of the American Oil Chemists' Society*, **72**, 873–80.

van de Voort, F. R., Ismail, A. A., Sedman, J. *et al.* (1994a) The determination of peroxide value by Fourier transform infrared spectroscopy. *Journal of the American Oil Chemists' Society*, **71**, 921–6.

van de Voort, F. R., Ismail, A. A., Sedman, J. and Emo, G. (1994b) Monitoring the oxidation of edible oils by Fourier transform infrared spectroscopy. *Journal of the American Oil Chemists' Society*, **71**, 243–53.

Visioli, F. and Galli, C. (1994) Oleuropein protects low density lipoprotein from oxidation. *Life Sciences*, **55**, 1965–71.

Visioli, F. and Galli, C. (1995) Natural antioxidants and prevention of coronary heart-disease – the potential role of olive oil and its minor constituents. *Nutrition Metabolism and Cardiovascular Diseases*, **5**, 306–14.

Visioli, F., Vinceri, F. F. and Galli, C. (1995) Waste-waters from olive oil production are rich in natural antioxidants. *Experientia*, **51**, 32–4.

Vlahov, G. (1996) Improved quantitative C-13 nuclear magnetic resonance criteria for determination of grades of virgin olive oils. The normal ranges for diglycerides in olive oil. *Journal of the American Oil Chemists' Society*, **73**, 1201–3.

Vlahov, G. and Angelo, C. S. (1996) The structure of triglycerides of monovarietal olive oils: a ^{13}C-NMR comparative study. *Fett/Lipid*, **98**, 203–5.

Wehrli, F. W., Marchand, A. P. and Wehrli, S. (1988) *Interpretation of Carbon-13 NMR Spectra*, 2nd edn, John Wiley, Chichester, Sussex.

Weiss, S. H. and Kulikowski, C. A. (1991) *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems*, Morgan Kaufmann, San Mateo, CA.

Werbos, P. J. (1993) *The Roots of Back-propagation: From Ordered Derivatives to Neural Networks and Political Forecasting*, John Wiley, Chichester, Sussex.

Williams, D. A. R. (1986) *Nuclear Magnetic Resonance Spectroscopy*, John Wiley, Chichester, Sussex.

Williams, K. P. J., Pitt, G. D., Batchelder, D. N. and Kip, B. J. (1994) Confocal Raman microspectroscopy using a stigmatic spectrograph and CCD detector. *Applied Spectroscopy*, **48**, 232–5.

Williams, K. P. J., Pitt, G. D., Smith, B. J. E. and Whitley, A. (1994) Use of a rapid scanning stigmatic raman imaging spectrograph in the industrial environment. *Raman Spectroscopy*, **25**, 131–8.

Winson, M. K., Goodacre, R., Timmins, É. *et al.* (1997a) Diffuse reflectance absorbance spectroscopy taking in chemometrics (DRASTIC). A hyperspectral FT-IR-based approach to rapid screening for metabolite overproduction. *Analytica Chimica Acta*, in press.

Winson, M. K., Todd, M., Rudd, B. A. M. *et al.* (1997b) A DRASTIC (diffuse reflectance absorbance spectroscopy taking in chemometrics) approach for the rapid analysis of microbial fermentation products: quantification of aristeromycin and neplanocin A in *Streptomyces citricolor* broths., in press.

Wold, H. (1966) Estimation of principal components and related models by iterative least squares, in *Multivariate Analysis* (ed. K. R. Krishnaiah), Academic Press, New York, pp. 391–420.

Wold, H. (1975) Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach, in *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett* (ed. J. Gani), Academic Press, London, pp. 117–42.

Wold, H. (1980) Model construction and evaluation when theoretical knowledge is scarce (theory and application of partial least squares), in *Evaluation of econometric models* (eds J. Kmenta and J. B. Ramsey), Academic Press, London, pp. 47–74.

Wold, H. (1982) Soft modeling: the basic design and some extensions, in *Systems under Indirect Observation: Causality, Structure, Prediction. Part II* (eds K. G. Jöreskog and H. Wold), North Holland, Amsterdam, pp. 1–53.

Wold, S. (1978) Cross validatory estimation of the number of components in factor and principal components models. *Technometrics*, **20**, 397–405.

Wonnacott, T. H. and Wonnacott, R. J. (1981) *Regression: A Second Course in Statistics*, John Wiley, Chichester, Sussex.

Yoder, C. H. and Schaeffer, C. D. J. (1987) *Introduction to Multinuclear NMR*, Benjamin/Cummings, Menlo Park, CA.

Zamora, R., Navarro, J. L. and Hidalgo, F. J. (1994) Identification and classification of olive oils by high-resolution C-13 nuclear magnetic resonance. *Journal of the American Oil Chemists' Society*, **71**, 361–4.

Zupan, J. and Gasteiger, J. (1993) *Neural Networks for Chemists: An Introduction*, VCH, Weinheim.