EVALUATION OF THAI FISH SAUCE QUALITIES BY NEAR-INFRARED SPECTROSCOPY

INTRODUCTION

Fish sauce is a traditional seasoning that has been widely used in Southeast Asian countries such as Thailand, Vietnam, Laos, Philippines and Southern part of China. It is a water-soluble proteinaceous fraction occurring after several months of storage of heavily salted fish at tropical temperature. Fish sauce has been used mainly as a condiment for rice dishes, but in addition to giving pleasant flavor, it is also important supplement of animal proteins to many people. Fish sauce fermentation is a means of the preservation and producing valuable products for fish species such as anchovy and sardine, which are not normally used as foods (Gildberg, 2001; Park *et al*., 2001).

In Thailand, there are many factories producing fish sauces for domestic consumption and export. Therefore, the government agencies control the factories to assure the quality standard of fish sauces. The qualities of commercial Thai fish sauces are identified by chemical and physical parameters, i.e. total nitrogen content, glutamic acid, pH, density etc. The analyses involved in the determination of these parameters are expensive, time consuming and require specialized personnel. According to disadvantages, nondestructive analysis methods have been desired. Near-infrared (NIR) spectroscopy is one of the efficient nondestructive techniques, which has many applications in both quantitative and qualitative analyses of foods and food products such as grains, seeds, fruits and vegetables (Osborne *et al*., 1993; Siesler *et al*., 2002). Few analyses have been done on fermented soy sauce to analyze the constituents such as protein, sodium chloride, glutamic acid but has not found in fermented fish sauce (Kawano, 1995; Iizuka and Aishima 1999). Fish sauce is also one of the fermented products like soy sauce. Therefore, it is possible to use NIR spectroscopy to analyze the constituents of fish sauce as well.

A key step in the implementation of a successful NIR analysis is the use of chemometric methods to extract analyte information from the spectral background arising from the sample matrix. Wold and Sjöström (1998) explained that the areas where chemometrics has been most successful according to all measures are i) multivariate calibration, ii) structure-(re)activity modeling, iii) pattern recognition, classification, and discriminant analysis, and iv) multivariate process modeling and monitoring.

In the NIR spectral analysis, the chemometric algorithm based on direct application of regression by partial least squares (PLS) is the most widely used methods for multivariate calibration. It is a popular full-spectral calibration method (Chalmers and Griffiths, 2002). However, the selection of a wavelength region is still important because using whole spectra region does not always yield optimal calibration results. Recently, new wavelength interval selection methods, namely moving window partial least squares regression (MWPLSR) and searching combination moving window partial least squares (SCMWPLS) have been proposed (Jiang *et al*., 2002; Kasemsumran *et al*., 2003; Kasemsumran *et al*., 2004; Du *et al*., 2004).

MWPLSR is a method to search for informative spectral regions for the multi-component spectral analysis (Jiang *et al*., 2002). Informative regions mean that they contain useful information for a PLS model building and are helpful to improve the performance of the model. MWPLSR builds a series of PLS models in a window that moves over the whole spectral region and then locates useful spectral intervals such as informative regions in terms of the least complexity of PLS models reaching a desired error level. SCMWPLS is a method used for searching the optimized combinations of informative regions selected by MWPLSR. In order to eliminate uninformative regions, SCMWPLS optimizes the informative regions and combines them altogether (Kasemsumran *et al*., 2003; Du *et al*., 2004; Kasemsumran *et al*., 2004).

The multivariate pattern recognition techniques such as Artificial Neural Networks (ANNs), Linear Discriminant Analysis (LDA), Soft Independent Modeling of Class Analog (SIMCA) and K Nearest Neighbors (KNN) together with NIR spectra have been used to classify foods and agricultural products such as soy sauce (Iizuka and Aishima, 1999), olive oil (Christy *et al*., 2004), mayonnaise (Indahl *et al*., 1999) and sugar beet (Roggo *et al*., 2003a). There are many types of pattern recognition algorithms which essentially differ in the way they define classification rules. Supervised pattern recognitions such as LDA, SIMCA, KNN and ANNs are pattern recognition techniques which require prior knowledge about the category membership of samples to develop the classified model. LDA is designed to find explicit boundaries between classes while KNN method does this implicitly. The SIMCA method puts the emphasis more on similarity within a class than on discrimination between the classes (Vandeginste *et al*., 1998).

However, classification of NIR data is usually an ill-posed problem i.e. the number of variables is larger than the number of samples to be estimated. Therefore, many studies try to optimize models by the choice of a suitable chemometric method (Friedman, 1989; Roggo *et al*., 2003b). Multivariate data analysis methods such as Principal Component Analysis (PCA) and Factor Analysis (FA) can be used to reduce the number of original variables by creating new variables, and this is one of possibilities to avoid the use of ill-conditioned data (Lee *et al*., 2005).

The aim of the present study was to develop nondestructive quantitative and qualitative methods for Thai fish sauces by use of NIR spectroscopy and multivariate methods. Quantitative models were developed to determine the chemical, physical and sensory properties of commercial Thai fish sauces and qualitative models were built by different supervised pattern recognition techniques to classify commercial Thai fish sauces into three groups based on the total nitrogen content. The outcomes of this research will be useful for Thai fish sauce industries, especially for their export. Moreover, it will be induced the engineers to develop the NIR sensor for on-line detecting qualities of fish sauces. Therefore, the objectives of this study were:

1. To determine the chemical and physical properties as well as the NIR spectra of commercial Thai fish sauces.

2. To develop the predictive models between the chemical properties and NIR spectra and between the physical properties and NIR spectra of commercial Thai fish sauces.

3. To develop the qualitative models for classifying commercial Thai fish sauces into three groups based on their total nitrogen content by different supervised pattern recognition techniques.

4. To establish a descriptive sensory profile and evaluate the sensory properties of commercial Thai fish sauces through trained panel descriptive analysis.

5. To categorize Thai fish sauces based on their sensory characteristics.

6. To describe the correlations between sensory, chemical and physical properties of Thai fish sauces studied by NIR spectroscopy combined with chemometrics.

7. To develop the predictive models between the perceived sensory properties and the NIR spectra of commercial Thai fish sauces.

LITERATURE REVIEW

The qualities of commercial Thai fish sauces are identified by chemical, physical and sensory parameters. However, the analyses involved in the determination of these parameters are destructive, expensive, time consuming, require specialized personnel and unsuitable for on-line application. Therefore, the development of fast, nondestructive, accurate, and on-line/at-line techniques is desired. NIR spectroscopy could form the basis for such techniques due to the ability to measure numerous samples within short time, easy to handle, requires little training of operators and can be operated on line and at line (Liu *et al*., 2003; Nilsen and Esaiassen, 2005). Therefore, attention will be focused on fish sauces, near-infrared spectroscopy and its application in foods, multivariate techniques for developing the quantitative and qualitative models for Thai fish sauces.

1. Fish Sauces

Fish sauce is a clear brown liquid hydrolysate from salted fish and is commonly used as a flavor enhancer or salt replacement in various food preparations. It is also an important supplement of animal protein to many people (Lee, 1990; Lopetcharat *et al*., 2001).

1.1 Fish Sauce Production

Fish sauce fermentation is a means of the preservation and producing valuable products for fish species such as anchovy and sardine, which are not normally used for food (Lee, 1990; Ravipim, 1991). Generally, traditional fish sauce fermentation begins by mixing fish with salt in a weight ratio of 1:1-3:1 and by natural fermentation at ambient temperature (30 - 40 °C) for 6 - 12 months or longer (Lopetcharat *et al*., 2001; Tsai *et al*., 2006). However, the ratio of fish to salt can be different in different regions. During the fermentation, fish proteins are hydrolyzed both by endogenous proteases and exogenous ones of microbial origin. The resulting product has a distinctive odor and flavor, which develops progressively as the

fermentation progresses (Fukami *et al*., 2004). At the end of fermentation, the primary extract is a high quality fish sauce. However, after the liquid is run-off, the residue is leached several times with brine solution to yield poorer quality sauces, which may be improved in their color, flavor and storage properties by using additives such as caramel or roasted rice (Sanceda *et al*., 1990). The process diagram of fish sauce production is shown in Figure 1.

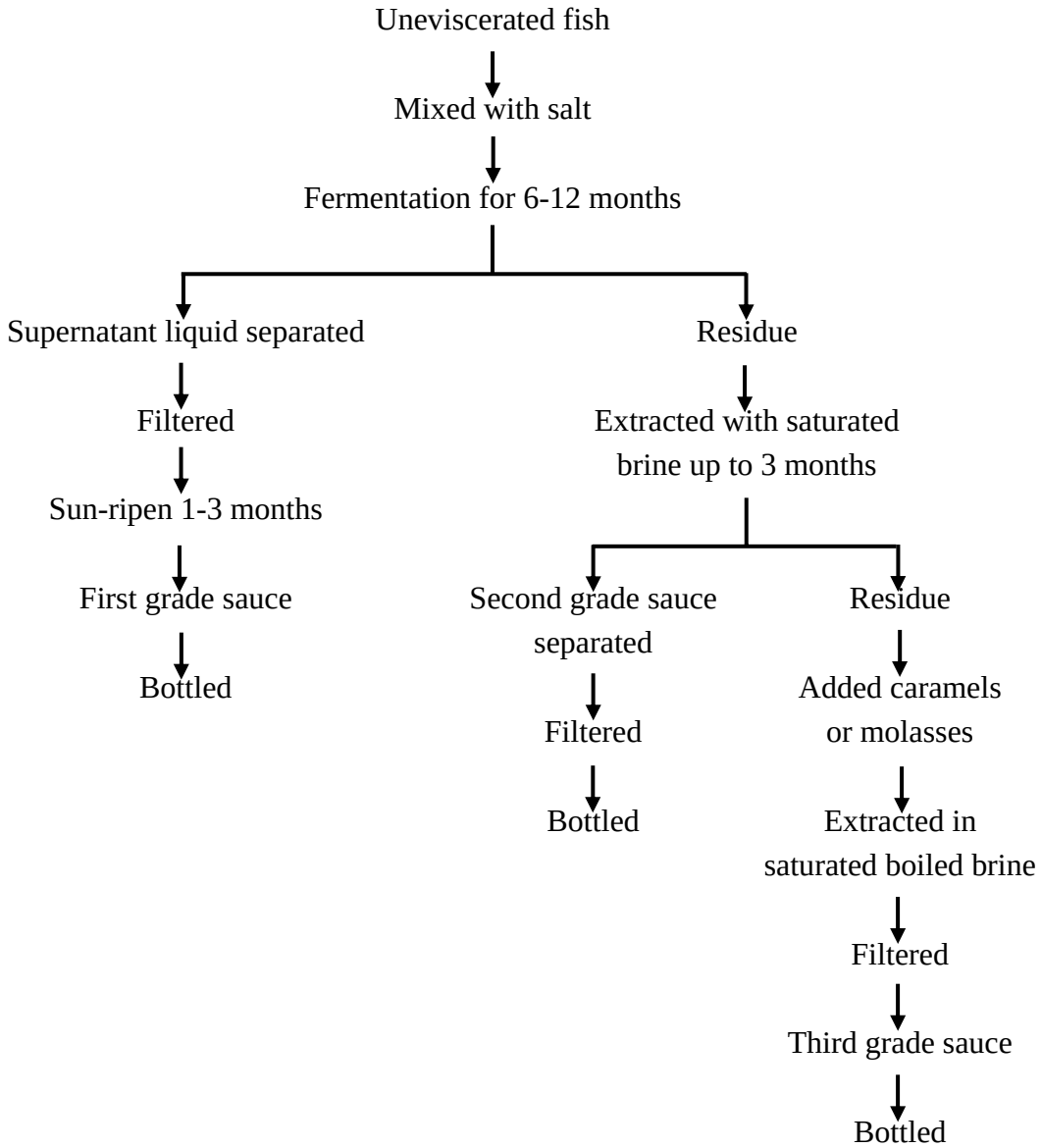


Figure 1 Fish sauce production.

Source: Ravipim (1991)

1.2 Development of Color, Flavor, and Aroma

Fish sauce fermentation process normally takes a long time to ensure the solubilisation as well as the color and flavor development. These characteristics are significant factors in consumer acceptability of fish sauce products.

Color of fish sauce becomes darker during the fermentation. This is due to the non-enzymatic browning reaction. Jones (1962) reported that two types of non-enzymatic browning reaction in fish sauce were the Maillard reaction and the Lipid-amino reaction. Lipids and their degradation products, carbonyl compounds and peroxide, appear to be important sources of carbonyl groups for the non-enzymatic browning reaction in fish sauce (Orejana, 1979). Primary and secondary lipid oxidation products are the biological amino acids, proteins, peptides, free amino acids and phospholipids. These react to produce interaction compounds and this makes the color of the product brown. It is also causes a change in flavor and loss in aromatic nutrient elements (Aubourg, 1998). Klomklao *et al.* (2006) reported that the Maillard reaction contribution to the increase in the a*(redness) was coincidental with a decrease in the lightness (L*-value), particularly when the fermentation time increased. In addition, the increase in browning was found to depend on salt concentration. The higher concentration of salt used, the lower was the increase in browning. This due to the addition of a higher amount of salt slowed down the breakdown of the fish meat by autolysis or microbial activities (Gildberg, 2001). Even though reducing sugar content in fish sauce is low, carbohydrate derivative such as glucse-6-phosphate and other substances present in the metabolic pathways, can also act as reactant to initiate the Maillard reaction (Kawashima and Yamanaka, 1996).

The aroma of fish sauce is a gauge for measuring the quality of the product because the very salty taste tends to overpower other flavoring constituents. Volatile acids were reported to be the most abundant flavor compounds in the fish sauces (Sanceda *et al*., 2003). The volatile compounds in fish sauces may vary due to species and origin of fish used, as well as the manner of production employed (Peralta *et al*., 1996). It is assumed that the volatiles contributing to flavor and aroma of fish sauce

are produced by nonenzymatic reactions by endogenous enzymes of fish origin and those of microorganisms serving during fermentation (Fukami *et al*., 2004). Many species belonging to *Bacillus, Micrococcus, Staphylococcus, Streptococcus, Pediococcus,* and other halophilic bacteria that produce lactic acid are found in fish sauce. However, it is unclear how these bacteria act on the production of the characteristic of taste and flavor of fish sauce during fermentation (Saisithi *et al*., 1966; Ijong and Ohta, 1996).

Previous studies have indicated that the volatile compounds in fish sauces are composed of three distinctive notes; i) ammonical, ii) cheesy, and iii) meaty. The ammonical note is produced by ammonia, amines, and other basic nitrogen-containing compounds (Saisithi *et al*., 1966; Dougan and Howard, 1975). The cheesy note is associated with low molecular weight volatile fatty acids (Dougan and Howard, 1975; Beddows *et al*., 1976). The source of the meaty note has not been fully clarified but it was believed that it could be produced by atmospheric oxidation of precursors present in the mature fish sauces (Peralta *et al.*, 1996; Shimoda *et al*., 1996).

Stone *et al.* (1974) subdivided three major contributing factors (ammonical, cheesy, and meaty), which describe the fish sauce odor into eight attributes as the descriptors for quantitative descriptive analysis (QDA). They were burnt, fishy, sweaty, feacal, rancid, cheesy, meaty, and ammonical notes. Fukami *et al.* (2002) reported that 2-methylpropanal, 2-methylbutanal, 2-pentanone, 2-ethylpyridine, dimethyl trisulfide, 3-(methylthio)-propanal, and 3-methylbutanoic acid were principal contributors to the distinctive odor of fish sauce. These volatile compounds contributing odor in fish sauce are shown in Table 1.

Table 1 Volatile compounds in a fish sauce characterized by AEDA and GC-MS.

Odor characters	Compounds
Fishy	2-ethylpyridine, dimethyl trisulfide
Sweaty	2-methylpropanal, 2-methylbutanal,2-ethylpyridine, dimethyl trisulfide
Fecal	2-ethylpyridine, dimethyl trisulfide
Rancid	2-methylpropanal, 2-methylbutanal,2-ethylpyridine, dimethyl trisulfide
Cheesy	2-ethylpyridine, 2-pentanone, volatile acids
Meaty	2-ethylpyridine, 2-methylpropanal, 2-methylbutanal
Burnt	2-ethylpyridine, dimethyl trisulfide, 2-methylpropanal, 2-methylbutanal

Source: Fukami *et al*. (2002)

1.3 Chemical and Nutritional Compositions of Fish Sauce

Fish sauce has been used in various prepared foods and sauce with the merit of its characteristic, flavorable taste and nutritive value. The nutritional value is limited due to the high salt content, but daily consumption of fish sauce renders it one of the main protein sources in some regions where carbohydrates are the fundamental part of diet (Amano, 1962). Furthermore, it is also a major source of mineral elements such as calcium, phosphorus, ferric and also the B-group of vitamins (Campbell-Platt, 1987; Ravipim, 1991). The chemical compositions of fish sauce are shown in Table 2. The high salt concentration of fish sauce limits its consumption. However, the importance of salt used in fermented products was determined as a bacteriostatic agent for many bacteria including pathogenic and spoilage bacteria (Ijong and Ohta, 1996).

Table 2 Chemical characteristics of fish sauce.

Components	Compositions
Moisture	63-79 g/100mL
Total nitrogen content	0.35-2.59 g/100mL
Nitrogen recovery	42-70 g/100 mL
pH	4.90 - 6.23
Sodium chloride	15-22 g/100mL
Calcium (Ca)	30-130 mg/100g
Phosphorus (P)	90-130 mg/100g
Ferric (Fe)	3-30 mg/100g
Potassium (K)	800 mg/100g
Thiamine	0.03 mg/100g
Riboflavin	0.3 mg/100g
Niacin	6-12 mg/100g
Retinol	30 μg/100g
Biotin	30 μg/100g

Source: Campbell-Platt (1987); Park *et al.* (2001)

During the fermentation, the rise of pH is probably due to the increase of alkaline volatile basic nitrogen. The total nitrogen content is mainly derived from breakdown of the fish protein by the enzymes of the fish. High quality Thai fish sauces must have a total nitrogen content of 20 gN/L based on the Kjeldahl method (Thai Industrial Standard, 1983). Most of the nitrogenous compounds in fish sauce are free amino acids and small peptides, which contribute to the brown color development, and the specific aroma and flavor (Finne, 1992; Lopetcharat *et al*., 2001). Both the bound and free amino acid content contributes to the nutritional quality of fish sauce. Therefore, it is considered as an important source of dietary proteins and amino acids (Sanceda *et al*., 1990). The free amino acid content of commercial fish sauces are shown in Table 3.

Table 3 Amino acid compositions of various commercial fish sauces sold in some Asian countries.

Amino acids (mg/mL)	Sources of fish sauce				
	Thailand	Vietnam	Myanmar	Laos	China
Taurine	1.19	1.71	1.42	0.37	1.17
Aspartate	5.83	10.02	2.87	0.54	5.11
Threonine*	3.84	5.84	1.31	0.28	2.85
Serine	2.33	4.83	0.11	0.24	0.95
Glutamate	14.89	15.84	5.60	0.31	11.64
Proline	1.35	3.22	0.67	0.18	1.27
Glycine	2.67	4.61	2.37	0.43	2.65
Alanine	5.74	9.85	4.69	1.79	5.97
Cysteine	0.17	0.45	0.13	ND	1.25
Valine*	4.78	7.09	2.89	0.32	4.93
Methionine*	2.22	2.30	0.75	0.32	2.06
Isoleucine*	3.34	3.74	1.84	0.72	3.48
Leucine*	4.39	4.27	2.71	1.43	4.71
Tyrosine	0.91	1.28	0.64	0.05	0.88
Phenylalanine*	3.23	4.15	1.10	0.36	3.28
Tryptophan	ND	0.50	0.01	ND	ND
Lysine*	7.67	12.69	4.05	1.23	6.53
Histidine*	2.75	3.70	0.12	0.09	1.91
Arginine*	0.03	2.17	0.08	0.07	ND

* Essential amino acid, ND: Not detected.

Source: Park *et al.* (2001)

Park *et al.* (2001) investigated the values of total nitrogen, total amino acids, total nucleosides and bases, and creatinine or the total of creatine and creatinine of fish sauce from seven Southeast Asian countries. They found that these compositions can be used to categorize fish sauces into three distinct groups. The three groups were the “high-content” group (Thai, Vietnam, and Japan), the “intermediate” group (China and Korea), and the “low-content” group (Myanmar and Laos). They suggested that creatinine which originates from creatine during fish sauce fermentation could be used as a practical chemical marker for the quality control in fish sauce

factories. Its concentration is useful to estimate the fish content in sauces. It is much more convenient than total nitrogen or other compounds because creatine determination is much easier and time saving.

1.4 Types of Thai Fish Sauce and Quality Standards

Thailand is the largest producer of fish sauces (Saisithi, 1994). According to the Fisheries Economics Division (2002), the quantity of fish sauces exported by country in 1999 showed that the total value of exporting was more than 600 millions baht and the quantity exported was more than 38 thousand tons. Thai fish sauces can be categorized into three types based on the production process and raw material, according to the Notification of the Ministry of Public Health (Ministry of Public Health, 2000). These three types of Thai fish sauce and their definitions are listed in Table 4.

Table 4 Definition of three types of Thai fish sauce.

Types of fish sauce	Definitions
1) Pure fish sauce	Fish sauce in which fermentation is derived from fish and fish residues.
2) Fish sauce made from other animals	Fish sauce where fermentation is derived from marinating other types of animals rather than anchovy fish.
3) Mixed fish sauce	Fish sauce with added with non-hazardous additive or flavoring agents

Source: Ministry of Public Health (2000)

The required chemical quality standards of Thai fish sauces by the Notification of the Ministry of Public Health (Ministry of Public Health, 2000) are listed in Table 5.

Table 5 Chemical standard for three types of Thai fish sauce.

Types of fish sauce	Sodium chloride (g/L)	Total nitrogen (g/L)	Glutamic acid per total nitrogen
1) Pure fish sauce	≥ 200	≥ 9	0.4 - 0.6
2) Fish sauce made from other animals	≥ 200	≥ 9	0.4 - 0.6
3) Mixed fish sauce	≥ 200	≥ 4	0.4 – 1.3

Source: Ministry of Public Health (2000)

Furthermore, the Thai Industrial Standard Institute (1983) categorized pure fish sauces into two grades of qualities based on the total nitrogen content; i) first grade (Total nitrogen content not less than 20 grams per liter) and, ii) second grade (Total nitrogen content not less than 15 grams per liter). The specification of physical and chemical standards for the first and second grade fish sauces are listed in Table 6.

Table 6 Specification standard for first and second grade fish sauces.

Specification	Fish sauces	
	First grade	Second grade
Relative density at 27°C	≥ 1.20	≥ 1.20
Acidity (pH)	5.0 – 6.0	5.0 – 6.0
Sodium chloride (g/L)	≥ 230	≥ 230
Total nitrogen (g/L)	≥ 20	≥ 15
Glutamic acid per total nitrogen	0.4 – 0.6	0.4 – 0.6
Nitrogen from amino acid content	≥ 10	≥ 7.5

Source: Thai Industrial Standard Institute (1983)

2. Near-Infrared Spectroscopy

Near-infrared (NIR) spectroscopy is an efficient nondestructive technique, which has many applications in both quantitative and qualitative analyses of foods and food products such as grains, seeds, fruits and vegetables (Osborne *et al*., 1993; Siesler *et al*., 2002). In this section, the principles of near-infrared spectroscopy, NIR calibration basics and applications of NIR spectroscopy are described.

2.1 Principle of Near-Infrared Spectroscopy

The American Society of Testing and Materials (ASTM) defines the NIR region of the electromagnetic spectrum as the wavelength range of 780-2526 nm corresponding to the wave number range 12820-3959 cm^{-1}. It covers the wavelength range adjacent to the mid infrared and extends up to the visible region (Figure 2). The NIR method is based on the phenomenon that the functional groups such as O-H, C-H, S-H and N-H absorb near infrared light. Absorption in the NIR region always occurs due to the overtone and combination vibrations of molecules in chemical compounds. Therefore, it is possible to identify compounds and individual chemical groups by their absorption and by the region of large wavelengths (Osborne *et al*., 1993; Reich, 2005).

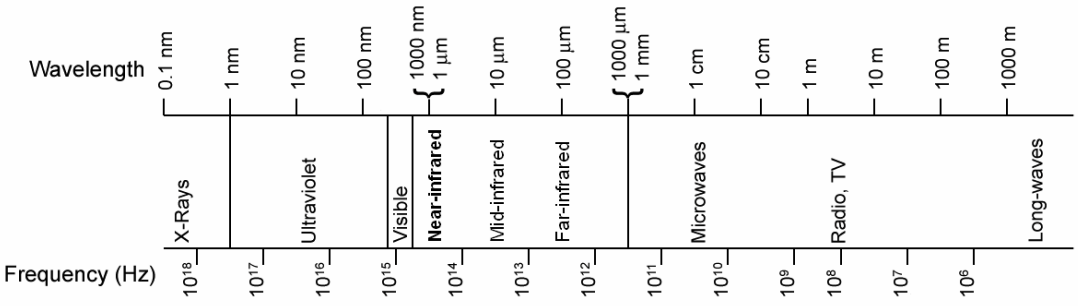


Figure 2 An overview of the electromagnetic spectrum.
Source: Modified from CRISP (2006)

2.2 Principle of NIR Measurement

The aim of NIR measurement is to find out the quantification or qualification of an unknown sample. A NIR spectrometer is generally composed of a light source, a monochromator, a sample holder or sample presentation interface, and a detector, allowing for transmittance or reflectance measurements (Reich, 2005). Approaching to NIR measurement, light comes from a source and illuminates the sample. The light interacts with the sample and modified in specific way. The modified reflected or transmitted light is directed to a detector that is sensitive to NIR light. The signal is converted into electrical information that can be read by a computer. This is generating spectral data. The spectral data are in the computer applied to the calibration models for quantitative and qualitative analyses. Figure 3 shows the principle of NIR measurement.

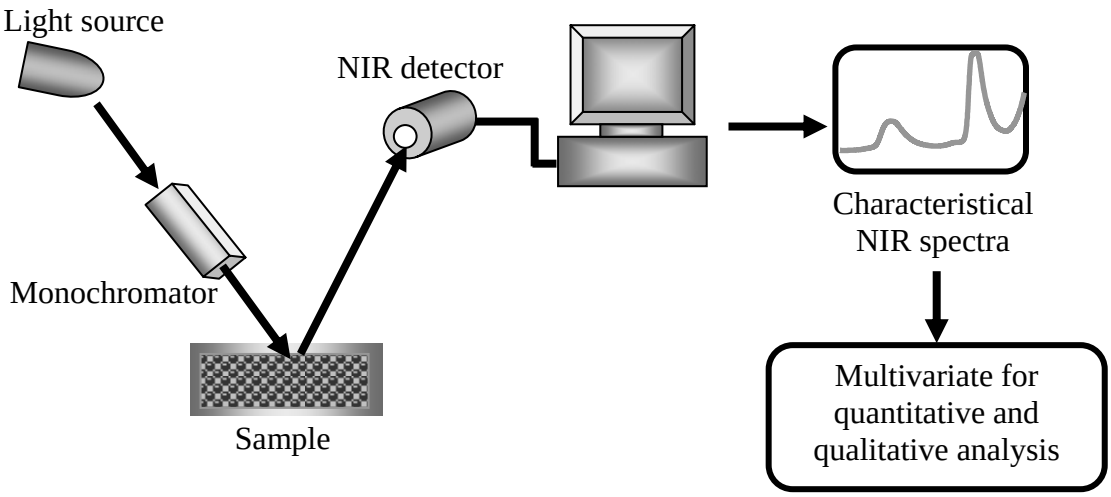


Figure 3 Principle of NIR measurement.
Source: Modified from STATOIL (2006)

The appropriate NIR measuring mode is dictated by the optical properties of the samples. Figure 4 illustrates some of measuring modes known as “transmittance”, “reflectance” and “transflectance”. In the case of transmittance, incident light illuminates one side of sample and the transmitted light may be detected

from the other side (Figure 4A and 4B).This mode is widely used for transparent materials. In the case of reflectance, incident light illuminates the surface of the sample and the diffusely reflected light from the surface may be detected (Figure 4C). In this mode, the sample should be opaque such as a powdered sample. Transflectance was developed by combining the transmittance and reflectance modes. In this case, incident light is transmitted through the sample then scattered back from a reflector (Figure 4D and 4E). Turbid liquids or semi-solids and solids may be measured in diffuse transmittance (Figure 4B), diffuse reflectance (Figure 4C) or transflectance (Figure 4D and 4E), depending on their absorption and scattering characteristics. In any case, absorbance (*A*) values relative to a standard reference material and measured, with *A* corresponding to log 1/R and log 1/T for reflectance and transmittance spectra, respectively (Kawano, 2001; Reich, 2005).

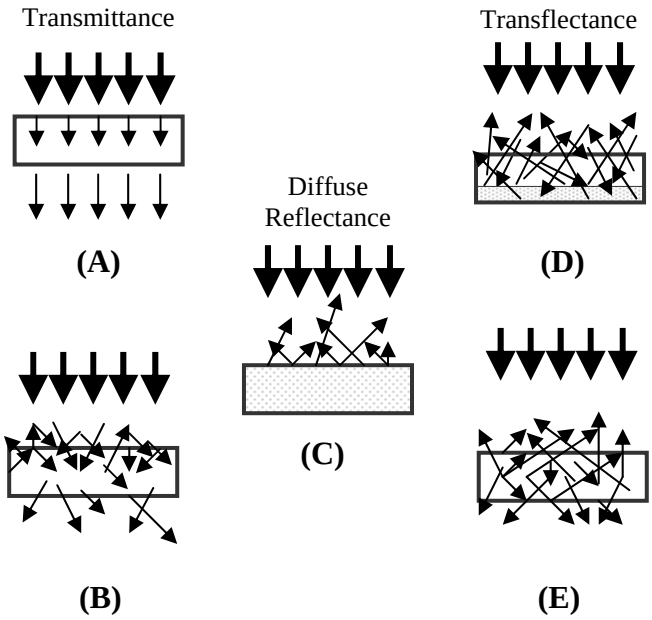


Figure 4 NIR measuring modes.
Source: Modified from Reich (2005)

Sample selection, sampling, sample preparation, and sample presentation to the instrument are fundamental to accurate and precise testing by both NIR and reference methods. Most analysis is carried out on what is purported to be a sample that is representative of the whole population and provides information on what the

operator needs to know. This can be achieved only by continuous sampling or repeated random sampling of the materials. Sampling of liquid samples depends on the viscosity and clarity of the liquid. Sample preparation includes documentation, blending, sub sampling, removal of unwanted material, grinding and storage. Sample presentation is also the important factor affecting NIR measurements. The definition of sample presentation is meant how to present or set a sample to a NIR instrument. Therefore, several types of sample cells, such as quartz cuvettes with defined optical path length for liquids, specifically designed sample cells with quartz windows for semi-solids and powders, and adjusted sample holders for tablets and capsules have been developed by the manufactures and the users of NIR instruments (Kawano, 2001; Williams, 2001; Reich, 2005).

2.3 NIR Calibration Basics

For quantitative and qualitative analyses, NIR spectroscopy needs a calibration equation. The calibration procedure involves collecting a number of samples, obtaining both reference and NIR data on each sample and deriving a calibration equation from these data by using chemometrics (Kawano, 2001). Chemometrics is the discipline using computers and mathematics to derive meaningful chemical information from samples of varying complexity (Workman, 2001). A main part of chemometrics is multivariate data analysis, which is essential for quantitative and qualitative assays based on NIR spectroscopy (Heise and Winzen, 2001).

Workman (2001) and Reich (2005) described that the calibration process basically involves the following steps: i) selection of a representative calibration sample set; ii) determination of standard concentration and spectra acquisition; iii) development of the mathematical model for quantitative or qualitative analysis; and iv) validation of the model. Figure 5 shows the calibration and validation process.

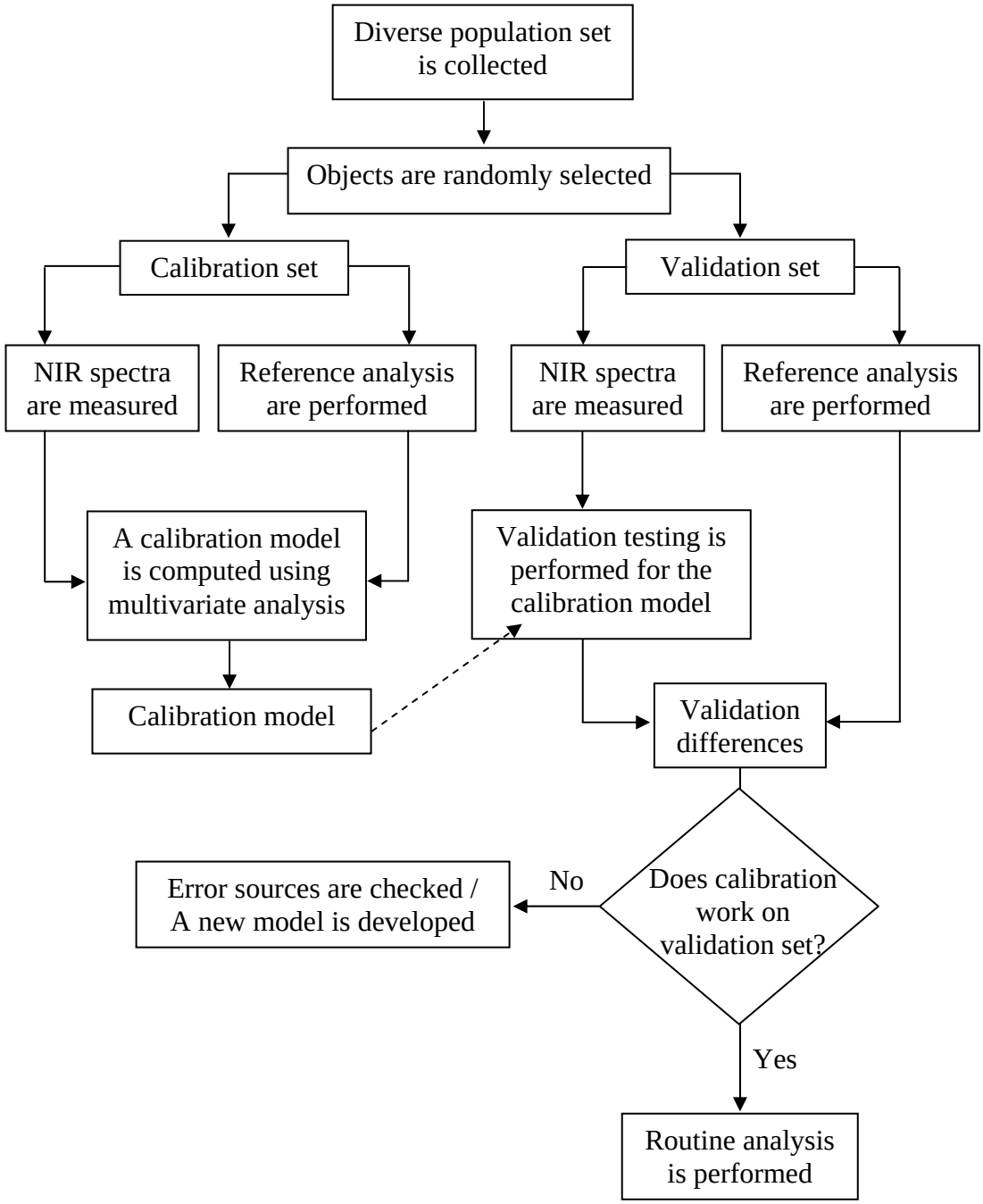


Figure 5 Flow diagram of NIR calibration and validation process.

Source: Modified from Workman (2001)

2.3.1 Multivariate Calibration for Quantitative Analysis

Quantitative analysis of NIR spectroscopic data is based on Beer's law, which states that a linear relationship exists between the molar concentration of a substance and the absorbance of this substance at a given wavelength (Heise and Winzen, 2001). The multivariate regression methods most frequently used in quantitative NIR analysis are principal component regression (PCR) and partial least squares (PLS) regression (Naes *et al*., 2002). PCR and PLS are full-spectrum calibration methods which achieve the same end by reducing the amount of spectral data in another way, constructing a small number of factors without discarding any useful information. The main difference between the two methods is that PCR uses the principal components provided by principal component analysis (PCA) to perform regression on the sample property to be a predicted, whereas PLS finds the directions of greatest variability by comparing both spectral and target property information with the new axes, called PLS factors. In some case, the spectral data and the target property may not be linearly related as a result of physical sample properties or instrumental effects. These can be solved by non-linear calibration methods, such as PLS2, locally weighted regression (LWR), and artificial neural networks (ANNs) (Osborne *et al*., 1993; Reich, 2005).Wavelength selection is composed of the decision of a subset of spectral channels with which the established calibration model gives the minimum errors in prediction. The selection of spectral intervals has been addressed in several works (Heise and Winzen, 2001). The details of PLS regression and wavelength selection methods have been used in this study will be described in section 3.3 and 3.4, respectively.

2.3.2 Multivariate Classification for Qualitative Analysis

Unlike concentrations in quantitative analysis, qualitative sample properties that have to be related to spectral variations have discrete values that represent a product identity or a product quality, for example "good" or "bad" (Heise and Winzen, 2001). Multivariate classification methods, also known as pattern recognition methods are used for grouping samples with similar characteristics. These

methods are subdivided into i) supervised and ii) non-supervised learning algorithms, depending on whether or not the class to which the sample belong is known. Supervised pattern recognition refers to techniques where prior knowledge about the category membership of samples is used for classification. Algorithm of this type such as Linear Discriminant Analysis (LDA), Artificial Neural Networks (ANNs), Soft Independent Modeling of Class Analog (SIMCA) and K Nearest Neighbors (KNN) are typically used for constructing spectral libraries. Non-supervised methods do not require prior knowledge about the group labels in the data, but instead produces the grouping itself. These methods are useful at an early stage of an investigation to explore subpopulations in a data set. Principal component analysis (PCA) is often performed with visual techniques. In this study, supervised pattern recognition have been used to construct classification models for qualitative analysis of Thai fish sauces. The details of these methods will be described in section 3.5.

2.4 Applications of Near-Infrared Spectroscopy in Foods

Near-infrared (NIR) spectroscopy applications have been developed in agriculture and food industries. The main advantages of NIR spectroscopy for food analysis are its speed, the absence of (or reduction in) sample pretreatment, and the avoidance of chemical use (Osborne *et al*., 1993; Alvarez *et al*., 2002). NIR spectroscopy is a technique that needs reference measurements to build the calibration model. Developing this calibration model which is done by chemometrical tools is time consuming but for routine analysis can be made in a few minutes (Luypaert *et al*., 2003).

In food industry, NIR spectroscopy is being widely used for quality control of raw materials, intermediate products and final products (Siesler *et al*., 2002). Most applications have dealt with the quantitative determination of main chemical components such as protein content, moisture and fat in cereals, animal feeds, fats, meat and milk, as well as carbohydrate in fruit juices and alcohol in beverages (Burns and Ciurczak, 2001). Applications of NIR for the prediction of functional properties and quality variables in food have also emerged. Luypaert *et al*. (2003) mentions the

use of NIR for the prediction of epigallocatechin gallate and epicatechin which are considered to be responsible for the antioxidant activity of green tea. McCaig (2002) applied NIR spectrometer to measure L*a*b* color values for food and agricultural products such as spices, seasonings and ground coffee. NIR spectroscopy can be applied to evaluate texture quality of food products. Geesink *et al.* (2003) investigated the potential of NIR-based models to predict shear force of pork samples.

Sensory analysis of food involves the measurement, interpretation and understanding of human responses to the properties of food perceived by the senses such as sight, smell, taste, touch and hearing (Marens and Martens, 2001). Predictions of sensory quality based on NIR spectroscopy have been reported. Several researches have been applied NIR spectroscopy to predict sensory attributes such as hardness, juiciness, tenderness, flavor and acceptability in both beef and pork meat (Brøndum *et al*., 2000; Ellekjaer *et al*., 1994; Liu *et al.,* 2003). Correlation between sensory attributes on cooked fish and NIR spectroscopy has been shown in the work of Warm *et al.* (2001). Quantitative descriptive analysis (QDA) is a sensory evaluation technique which is a much used tool in the combination of NIR and sensory analyses. The QDA technique is based on the principle of a panelist's ability to verbalize perceptions of a product in a reliable manner. The method composes of a formal screening and training procedure, development and use of a sensory language, and the scoring of products on repeated trials to obtain a complete, quantitative description (Hootman, 1992).

In addition, NIR is widely applied for qualitative analysis. There are four main types of qualitative application of NIR spectroscopy. These are i) determination of the general composition of a substance; ii) detection of impurities; iii) detection of areas of maximum differences in concentration; and iv) classification of unknown substances (Williams and Norris, 1990). A large number of NIR applications for qualitative analysis are related to quality and authenticity studies of food products (Indahl *et al*., 1999). Food authentication is a wide-ranging issue that has come to prominence in recent years. Previous papers have described the application of NIR to the analysis of food authentication. Downey *et al*. (2002) gave an application with

sunflower oil, Wang *et al*. (2006) focused on quantification and discrimination of soybean oil adulteration in camellia oils. Cozzolino *et al*. (2005) used NIR to identify and authenticate fishmeal batches made with different fish species. Furthermore, NIR combined with pattern recognition techniques have also demonstrated a capability for discrimination between sets of similar biological materials such as, finishing oils (Blanco and Pages, 2002), coffee varieties (Downey *et al*., 1997), wines (Cuadrado *et al*., 2004), green tea (Luypaert *et al*., 2003) and apple juice samples (Reid *et al*., 2005). Those researches were based on the classification of chemically different samples and geographical origin.

3. Multivariate Data Analysis

Multivariate analysis is often used in spectroscopy to extract information from complex spectra containing overlapping absorption peaks, interference effects, and instrumental artifacts from the data collected (Paradkar *et al*., 2002). In the construction of a calibration model for quantitative analysis, partial least squares (PLS) regression is the most popular multivariate method (Du *et al*., 2004). Discriminant analysis is another multivariate procedure commonly used for the classification of similar objects into groups or clusters based on a statistical measure. Several pattern recognition mainly supervised pattern recognition has been useful in classifying foods and agricultural products. Multivariate methods such as principal component analysis (PCA) and factor analysis (FA) are useful techniques for reducing the number of original variables by create new variables. PCA and FA will find groups and sets of variables with similar properties, and the supervised pattern recognition techniques can discriminate them into unique population (Lee *et al*., 2005).

In this section, five multivariate techniques which widely combined with NIR spectroscopy applications are described. These are i) Principal component analysis; ii) Factor analysis; iii) Partial least square regression; iv) Wavelength interval selection methods; and v) Supervised pattern recognition techniques.

3.1 Principal Component Analysis

Principal component analysis (PCA) is one of variable-directed techniques for forming new variables which are linear composites of the original variables. The objective of the PCA is to reduce the dimensionality of the data set. It involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables that are called *principal components* (PCs). The values of new variables are called *principal components scores* (Sharma, 1996; Johnson, 1998).

Camo (1996) concluded that the principle of PCA is the following: find the directions in space along which the distance between data points is the largest. This can be translated as finding the linear combinations of the initial variables that contribute most to make the samples different from each other (Figure 6). The first PC accounts for the maximum variance in the data. The second PC accounts for the maximum variance that has not been accounted for by the first variable, and so on.

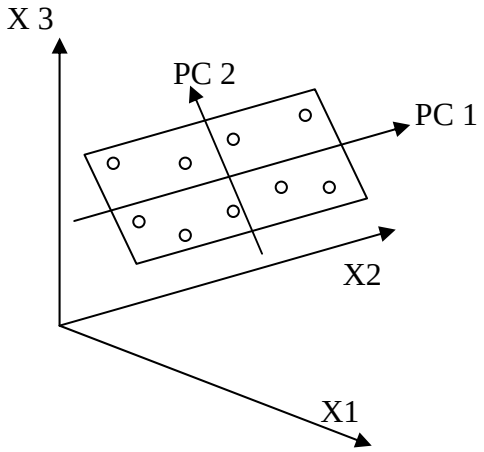


Figure 6 Principal component decomposition. The new axes called principal components (PCi) are linear combinations of the original variables (Xi), calculated so that the first PCs point in the direction of greatest dispersion of the samples.

Source: Camo (1996)

PCA is based on a decomposition of the data matrix **X** into two matrices **V** and **U** as shown in Figure 7. The two matrices **V** and **U** are orthogonal. The matrix **V** is usually called the *loadings matrix*, and the matrix **U** is called the *scores matrix*. The loadings can be understood as the weights for each original variable when calculating the principal component. The matrix **U** contains the original data in a rotated coordinate system (Jolliffe, 1986).

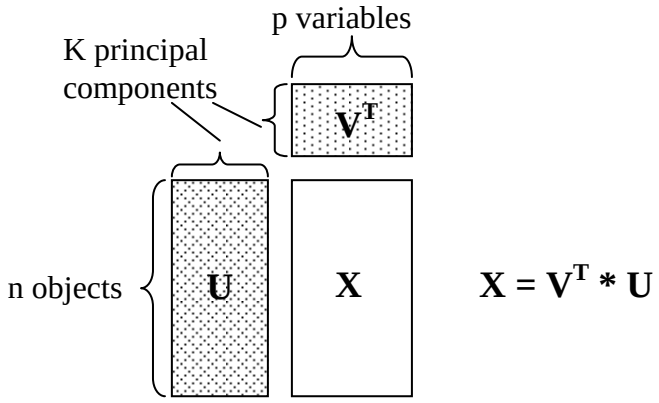


Figure 7 Scheme for explanation of Principal Component Analysis (PCA). PCA is based on a decomposition of the data matrix **X** into two matrices **V** and **U**. The matrix **V** is usually called the loadings matrix, and the matrix **U** is called the scores matrix.

Source: Lohninger (2006)

Camo (1996) explained that in order to interpret the results of a PCA, it is useful to understand the combination of loadings and scores called the structure part of the data. Loadings and scores can be interpreted as follows:

(1) Loadings describe the data structure in terms of *variable correlation*. Each variable has a loading on each PC. In geometrical terms, a loading is the cosine of the angle between the variable and the current PC: the smaller the angle, the larger the loading. If two variables have high loading along the same PC, it means that their angle is small, which in turn means that the two variables are highly correlated. In

additional, if both loadings have the same sign, the correlation is positive. Else, it is negative.

(2) Scores describe the data structure in terms of *sample patterns*, and more generally show *sample differences or similarities*. Each sample has a score on each PC. It reflects the sample location along that PC; it is the coordinate of the sample on the PC. Sample with close scores along the same PC are similar (they have close values for the corresponding variables). Conversely, samples whose score differ much are quite different from each other with respect to those variables.

The most reason for performing a PCA is to use it as a tool for screening multivariate data. An examination of the results graphical display used PC scores as input often reveal abnormalities and outliers in the data set. In additional, the PC scores can be analyzed individually to see whether distributional assumptions such as normality of the variables and independence of the experimental units hold. Principal component analysis is also helpful to researchers who want to partition experimental units into subgroups so that similar experimental units belong to the same subgroup (Johnson, 1998).

3.2 Factor Analysis

Factor analysis (FA) is another of variable-directed technique which is often used to create new variables that summarize all of the information that might be available in the original variables. Factor analysis creates a new set of uncorrelated variables from a set of correlated variables. Therefore if the original variables are already uncorrelated, then there is little reason to consider carrying out a FA (Johnson, 1998).

Like principal component analysis, the essential purpose of factor analysis is to describe the variation among many variables in term of a few underlying but unobservable random variables called *factors*. However, one important difference is PCA is not based on any particular statistical model, whereas factor analysis is based

on a model. Furthermore, factor analysis is concerned with explaining the covariance and/or correlation structure among the measured variables, whereas PCA is primarily concerned with explaining the variance of the variables. Factor analysis can also be viewed as a statistical procedure for grouping variables into subsets such that the variables within each set are mutually highly correlated, whereas at the same time variables in different subsets are relatively uncorrelated (Jobson, 1992; Manly, 2005).

Malinowski (1991) concluded that the principle behind factor analysis is quite simple, although the concrete realization depends on the requirements of the specific situation, and may be quite demanding: in principle, a data matrix **X** is split up into a product of two data matrices:

$$\mathbf{X} = \mathbf{U}\,\mathbf{V}$$

The matrices **U** and **V** are called *scores* and *loading matrices*, respectively. This can be visualized by the following figure:

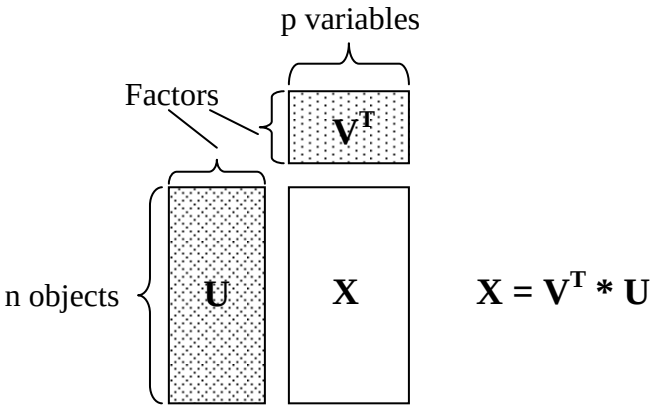


Figure 8 Scheme for explanation of Factor Analysis (FA). FA is based on a decomposition of the data matrix **X** into two matrices **V** and **U**. The matrices **U** and **V** are called scores and loading matrices, respectively.

Source: Lohninger (2006)

Factor score is a composite of all of the original variables that were important in making the new factor. Factor scores are standardized to have a mean of 0 and a standard deviation of 1. The factor scores can be used in subsequent analysis such as discriminant analysis. Factor loading describe the correlation between the original variables and the factors, and the key to understanding the nature of a particular factor. Squared factor loading indicate what percentage of the variance in an original variables is explained by a factor (Hair *et al*., 1998)

The main applications of factor analysis are i) to determine whether a smaller set of uncorrelated variables exists that will explain the relationships that exist between the original variables, ii) to determine the number of underlying variables, iii) to interpret these new variables, iv) to evaluate individuals or experimental units in the data set on these new variables, and v) to use these new variables in other statistical analyses of the data (Johnson, 1998; Hair *et al.,* 1998).

3.3 Partial Least Squares Regression

Partial least squares (PLS) regression is the most popular multivariate method and has been widely used in NIR analysis (Chalmers and Griffiths, 2002). PLS is a full-spectral calibration method. PLS models are based on principal components of both the independent data **X** and the dependent data **Y**. The central idea is to calculate the principal component scores of the **X** and the **Y** data matrix and to set up a regression model between the scores (and not the original data).

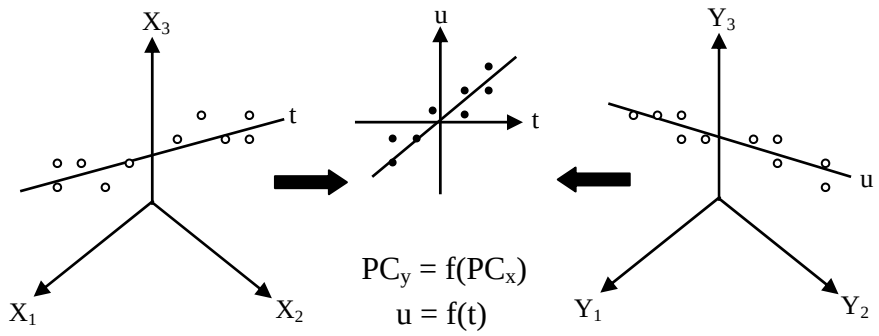


Figure 9 PLS component decomposition. PLS models are based on principal components of both the independent data **X** and the dependent data **Y**. The matrices **t** and **u** are the principal component scores of the **X** and the **Y** data matrix. A regression set up between both scores called PLSR model.

Source: Camo (1996)

Both PLS and Principal component regression (PCR) are projection methods as same as PCA. The difference between PLS and PCR lies in the algorithm. PLS uses both the independent and dependent variables to find the regression model. PCR decomposes the X-matrix by PCA then fits a multiple linear regression (MLR) model using the PCs instead of the original data as predictors (Camo, 1996). The scheme for explanation of PLS is shown in Figure 10.

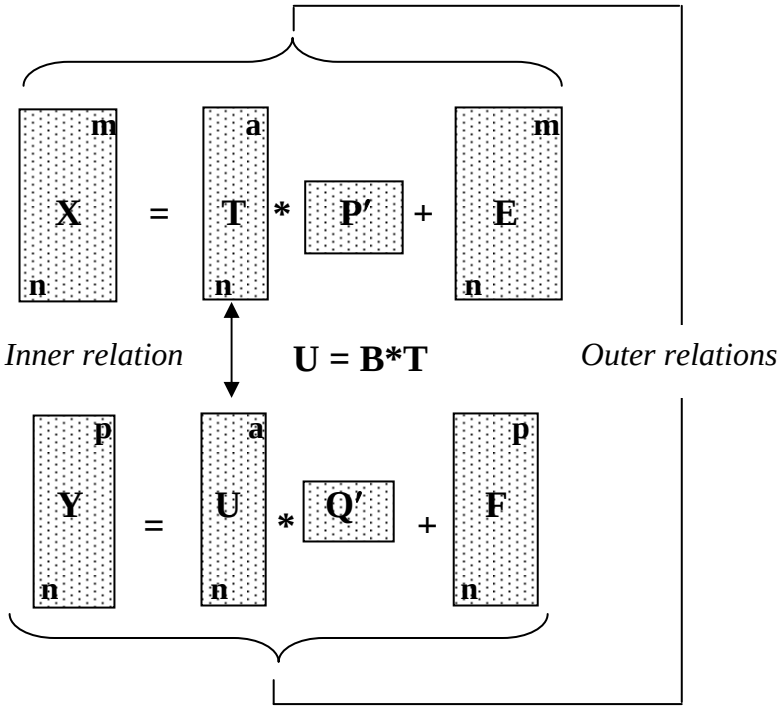


Figure 10 Scheme for explanation of Partial least squares (PLS) regression. The matrix **X** is decomposed into a matrix **T** (the score matrix) and a matrix **P'** (the loadings matrix) plus an error matrix **E**. The matrix **Y** is decomposed into **U** and **Q** and the error term **F**. These two equations are called outer relations. The goal of the PLS algorithm is to minimize the norm of F while keeping the correlation between **X** and **Y** by the inner relation **U = BT.**

Source: Modified from Lohninger (2006)

The important point when setting up a PLS model is to make a decision for the optimum number of principal components involved in the PLS model. While this can be done from variation criteria for other models, for PLS the optimum number of components has to be determined empirically by cross validation of the PLS model using an increasing number of components. The model with the smallest predictive error sum of squares value can be regarded as the "best" model (Lohninger, 2006).

PLS calibration of a multicomponent system can be performed into two different ways; PLS1 and PLS2 regression. PLS1 regression do a separate regression for each analyze such as univariate (in y), whereas PLS2 model the various analyses collectively in one and the same multivariate regression model. Considering the fact that PLS1 calibration usually performs equally well or better in terms of predictive accuracy, Vandeginste *et al*. (1998) advised that a separate PLS1 regression for each analyte is selected when the ultimate requirement of the calibration study is to enable the best possible prediction.

3.4 Wavelength Interval Selection Methods

PLS regression is a popular full-spectral calibration method. It has been widely used for the NIR spectral analysis (Chalmers and Griffiths, 2002). However, the selection of a wavelength region is still important because using whole spectra region does not always yield optimal calibration results (Kasemsumran *et al*., 2003; Du *et al*., 2004; Kasemsumran *et al*., 2004). Recently, new wavelength interval selection methods, namely moving window partial least squares regression (MWPLSR) and searching combination moving window partial least squares (SCMWPLS) have been proposed (Jiang *et al*., 2002; Kasemsumran *et al*., 2003; Du *et al*., 2004; Kasemsumran *et al*., 2004).

3.4.1 Moving Window Partial Least Squares Regression

Moving window partial least squares regression (MWPLSR) was proposed by Jiang *et al.* (2002) to search for informative spectral regions in order to further improve the prediction of PLS model. Informative regions mean that they contain useful information for a PLS model building and are helpful to improve the performance of the model. This method builds a series of PLS models in a window that moves over the spectral direction and then locates useful spectral intervals in terms of the model complexity and the sum of residuals (SSR). Log (SSR) is plotted as a function of the position of the window. A figure containing such residual lines is plotted, which provides the information about informative regions, where residue lines

show low values of SSR. When multiple spectral intervals are selected, two strategies are suggested for coupling the MWPLSR procedure with PLS for multi-component spectral analysis; one is the inclusion of all selected intervals to build a PLS calibration model, and the other is the combination of the PLS models built separately in each interval (Kasemsumran *et al*., 2003; Du *et al*., 2004; Kasemsumran *et al*., 2004). Figure 11 shows the calculations process of MWPLSR.

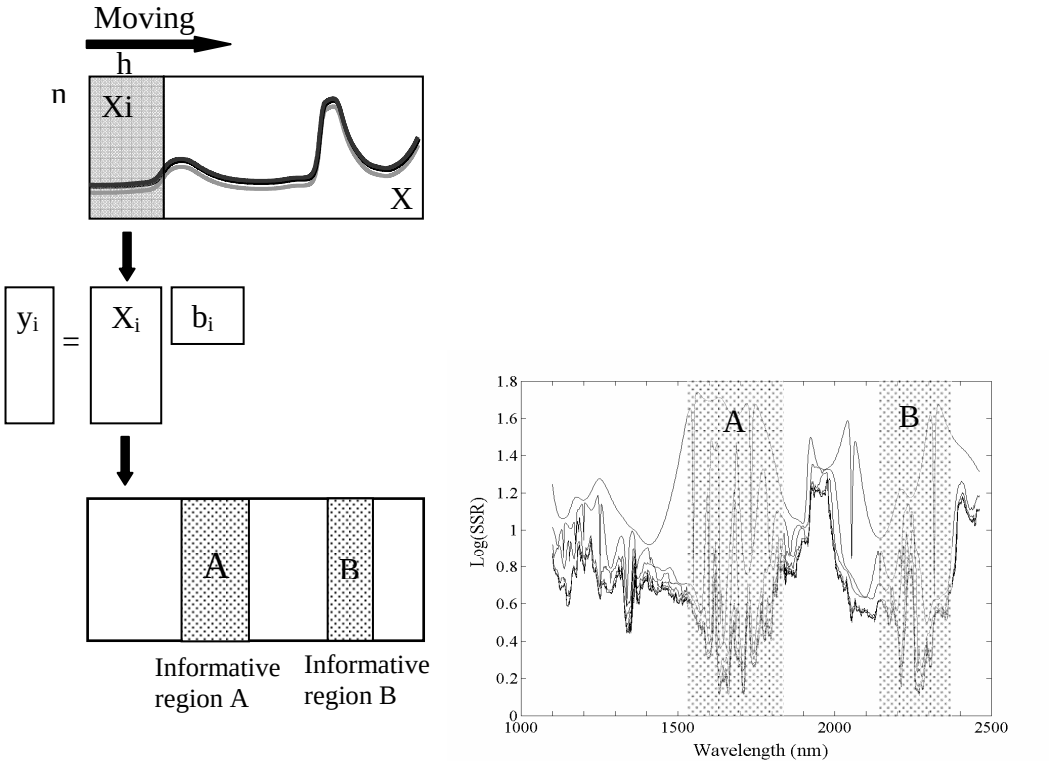


Figure 11 Scheme for explanation of MWPLSR. The sums of squared residuals (SSR) are calculated with the PLS models; $SSR_i = (y_i - X_i \hat{b}_i)^t (y_i - X_i \hat{b}_i)$. Log(SSR) is plotted as a function of the position of the window.

Source: Modified from Kasemsumran *et al.* (2004)

3.4.2 Searching Combination Moving Window Partial Least Squares

Searching combination moving window partial least squares (SCMWPLS) was proposed by Du *et al*. (2004). It is a method used for searching the optimized combinations of informative regions selected by MWPLSR. In order to eliminate uninformative regions, SCMWPLS optimizes the informative regions and combines them altogether. Figure 12 shows the optimized combination of informative regions process performed by SCMWPLS. In this method, MWPLSR is used first to identify informative regions from the whole spectra of a compound, and then directly search for the optimized combination of regions by SCMWPLS. The optimized combination of these informative regions for building the PLS models may be able to improve the ability of calibration model. The advantage of SCMWPLS is due to the fact that this method only focuses on informative regions, not the whole spectra. Therefore, it will not take too much time (Du *et al*., 2004; Kasemsumran *et al*., 2004).

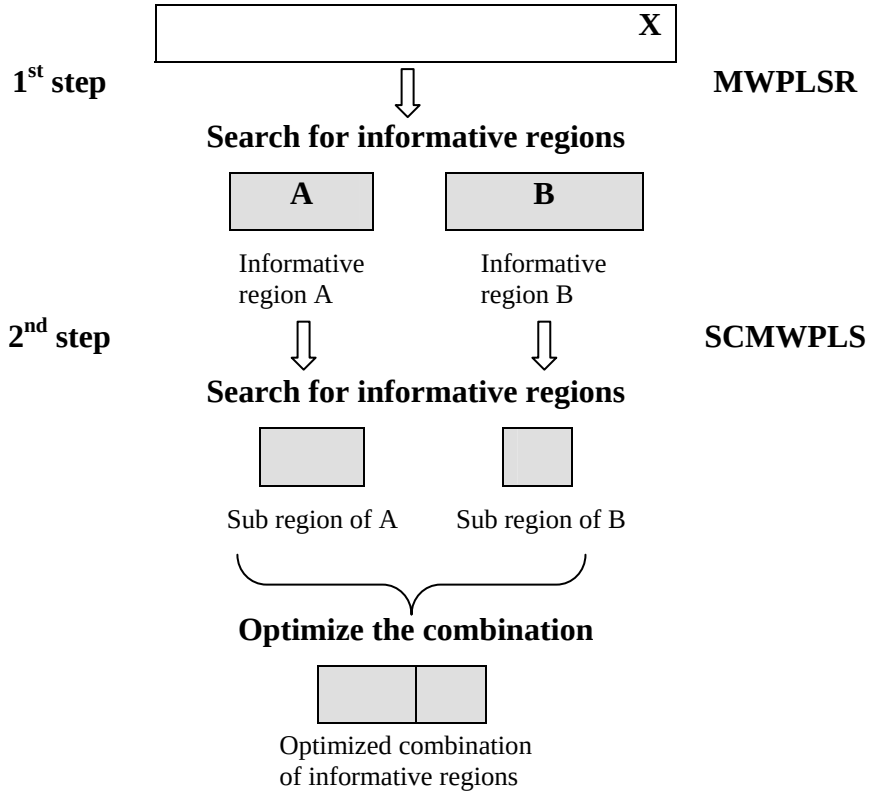


Figure 12 Scheme for explanation of SCMWPLS.
Source: Modified from Du *et al.* (2004)

3.5 Supervised Pattern Recognition

Supervised pattern recognition is one of pattern recognition techniques which require a prior knowledge about the category membership of samples to develop the classification model (Vandeginste *et al*., 1998). Various supervised pattern recognition techniques such as Artificial Neural Networks (ANNs), Linear Discriminant Analysis (LDA), Soft Independent Modeling of Class Analog (SIMCA) and K Nearest Neighbors (KNN) together with NIR spectra have been used to classify foods and agricultural products such as soy sauce (Iizuka and Aishima, 1999), olive oil (Christy *et al*., 2004), mayonnaise (Indahl *et al*., 1999) and sugar beet (Roggo *et al*., 2003a).

Vandeginste *et al*. (1998) explained that supervised pattern recognition techniques essentially consist of the following steps:

(1) Selection of training or learning set. This consists of objects of known classification for which a certain number of variables are measured.

(2) Feature selection. This step means the selection of variables that are meaningful for the classification and elimination of those that have no discriminating (or, for certain techniques, no modeling power).

(3) Derivation of a classification rule by using the training set.

(4) Validation of the classification rule by using an independent test set.

There are many types of supervised pattern recognition which essentially differ in the way they define classification rules. In this section, four supervised pattern recognition techniques such as LDA, SIMCA, KNN and ANNs were described followings:

3.5.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a linear and parametric method which discriminates between two or more groups of samples. LDA focuses on finding optimal boundaries between classes. The boundaries are obtained by linear discriminant analysis. Linear discriminant function is a latent variable which is created as a linear combination of discriminating (independent) variables:

$$D = b_1x_1 + b_2x_2 + ... + b_nx_n + c$$

where the b's are discriminant coefficients, the x's are discriminating variables, and c is a constant. Two classes A and B to be discriminated by linear discriminant function are shown in Figure 13.

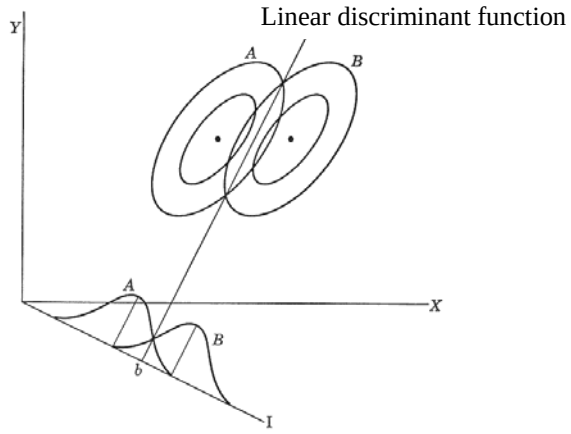


Figure 13 The boundaries are obtained by linear discriminant function.
Source: ASTQ (2006)

LDA is one of feature reduction methods as PCA. Unlike PCA, LDA select direction which achieves maximum separation among the different classes, whereas PCA selects a direction which retains maximum structure in a lower dimension among the data (Roggo *et al.*, 2003b). The only condition to apply LDA is that the number of variables should be lower than the sample number. Therefore, the feature reduction techniques, such as PCA and FA, are needed when NIR data are used.

When the discriminating function is parameterized, it has to be tested either by using an independent set of test data, or by performing cross-validation. In both cases, the results of the test set should be comparable to the training data (Lohninger, 2006).

3.5.2 Soft Independent Modeling of Class Analogy

Soft Independent Modeling of Class Analogy (SIMCA) is a well known pattern recognition method which focuses on modeling the classes rather than on finding an optimal classifier. It is a parametric method. This method is based on making a PCA model for each class in the training set. The optimal number of PCs should be chosen for each model separately. Unknown samples are then compared to the class models, and assigned to classes according to their analogy to the training samples (Camo, 1996; Vandeginste *et al*., 1998; Maesschalck *et al*., 1999). Therefore, the performance of the method depends not only on the difference between classes, but also strongly on the training set for each class (Candolfi *et al.*, 1999). An example of the SIMCA classification approach is shown in Figure 14.

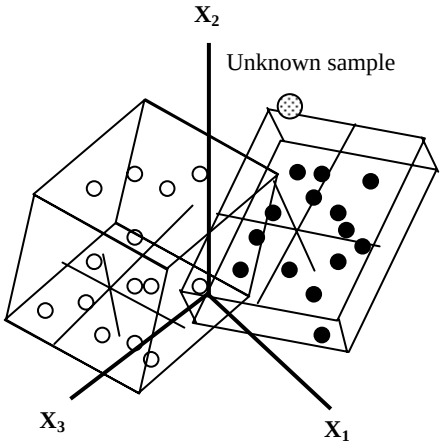


Figure 14 SIMCA classification of two classes as shown in the soft modeling representation of the data. SIMCA uses PCA as a starting point. An unknown sample is identified by its position within a class.

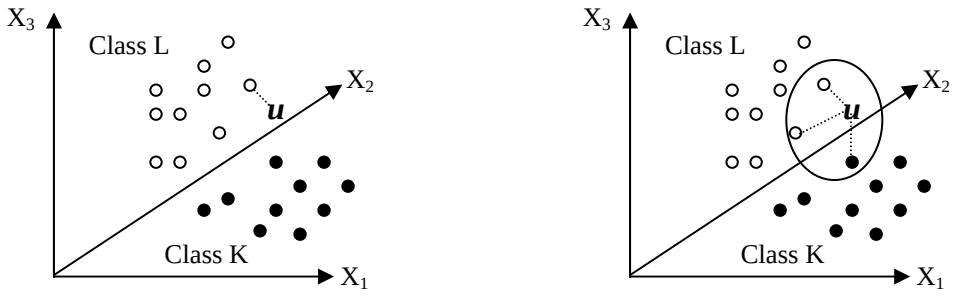
Source: Modified from Anonymous (2001)

A useful tool in the interpretation of SIMCA is the so-called Cooman's plot. The Cooman's plot shows distance between samples and the center of each group (Iizuka and Aishima, 1999). SIMCA has been applied very often and with much success in chemometrics. Examples are food authentication and pharmaceutical identifications. Furthermore, SIMCA can be applied as an outlier test (Vandeginste *et al*., 1998; Maesschalck *et al*., 1999).

3.5.3 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a non-parametric method in pattern recognition. In KNN, the nearest neighbor of an object requires the definition of similarity measure, such as the correlation coefficient and Euclidean distance (Wu and Massart, 1997). Usually the Euclidean distance is the most commonly used but for strongly correlated variables, the correlation coefficient should be preferred (Vandeginste *et al*., 1998). The basic idea of KNN discriminant analysis is as follows: First selects the **K** objects in the training set that is close to an unknown object ***u*** and computes the distance between ***u*** and each of the selected objects. Then ***u*** is classified in the group which the object's nearest neighbor belong (Johnson, 1998). The simplest case is 1-NN classification as shown in Figure 15a. Figure 15b shows an example of a 3-NN classification.

The choice of K is determined by optimization such as determines the prediction ability with different values of K. Usually it is found that small values of K (3 to 5) are to be preferred (Vandeginste *et al*., 1998). The nearest neighbors method is often applied because of its simplicity and surprisingly good results. Furthermore, it is free from statistical assumptions, such as normality of the distribution of the variables (Vandeginste *et al*., 1998).



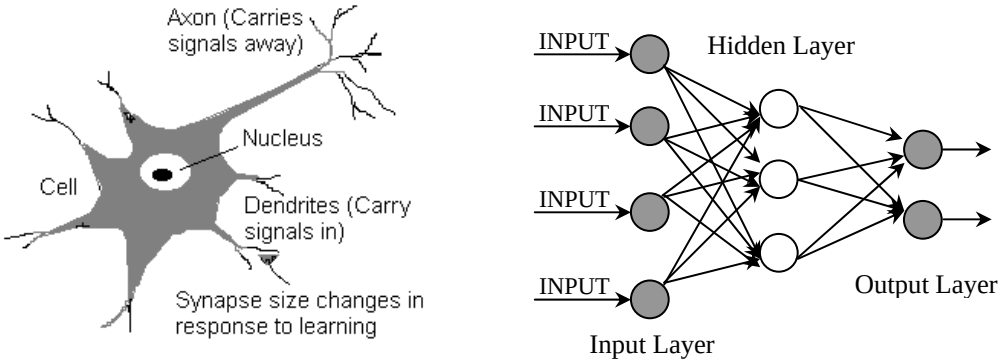
(a) 1-NN classification (b) 3-NN classification

Figure 15 KNN classification (a) 1-NN classification, (b) 3-NN classification. An unknown object ***u*** is classified in the group which the object's nearest neighbor belong.

Source: Modified from Vandeginste *et al*. (1998)

3.5.4 Artificial Neural Networks

Artificial Neural Networks (ANNs) are non-linear and non-parametric classification methods. ANNs are composed of several layers of neurons: input, hidden and output layers (Figure 16b). A neuron is a processing unit of which inputs are transformed by an activation function into the outputs (Roggo *et al.*, 2003b). ANNs have become a popular classification tool. This method can be applied to any classification problem (linear, on-linear, fuzzy, etc.) with the only requirement that the training set is representative and contains enough objects (Wu and Massart, 1996). ANNs are based on the parallel architecture of biological nervous systems, such as the brain and process information (Figure 16a). Like people, ANNs are configured for a specific application, such as pattern recognition, through a learning process.



(a) Biological Neural (b) Artificial Neural Network

Figure 16 Architectures of (a) biological neurons and (b) artificial neural networks. Source: Modified from NIBS (2006)

There are numerous algorithms available to be used for training the ANNs, such that it remains necessary to select appropriate learning rule in order to achieve an acceptable solution (Kim *et al.*, 2000). The feed forward nets trained by the back propagation learning algorithm are the most popular ones (Wu and Massart, 1996). Application of the PC scores instead of the original variables as the net input leads to efficient reduction of the net architecture (i.e. lowers the number of nodes in the input layer and the number of weights) and usually enhances the speed of the training phase (Wu and Massart, 1996; Blanco *et al*., 2000).