Development of Novel Calibrations for FT-NIR Analysis of Protein, Oil, Carbohydrates and Isoflavones in Foods

1. Introduction

The determination of food composition is fundamental to theoretical and applied investigations in food science and technology, and is often the basis of establishing the nutritional value and overall acceptance from the consumer standpoint. Most of the methods described in chapter 1 are useful for the conventional analysis of foods, that is, the determination of the major components (proteins, lipids, moisture, carbohydrates, and minerals). These components are included in standard tables of food composition. Advances in food analysis in the last three decades have resulted from the development of many instrumental methods such as NIR and from the improvements in separation methods (mainly chromatography).

The analyst often assumes that the sample to be analyzed is homogeneous. It is advisable that before starting a determination, the whole sample be mixed to eliminate heterogeneity – mainly in particle size and moisture distribution (Pomeranz and Meloan 1994). In some foods like concentrated sugar solutions, the sample must be heated carefully to dissolve sugar crystals.

1.1. Rationale: Why is it necessary to analyze the composition of soy and other health foods

Soy and other health foods are thought to be potentially important for lowering cholesterol and the prevention, or treatment of atherosclerosis and coronary heart disease. Soy food composition is also important for weight loss/weight control (Liu et al., 1995). Therefore, quality control and routine monitoring of soy and other health food composition is important to the consumers. Monitoring the levels of isoflavones in health foods such as soymilk appears also to be important in populations that are at risk for certain types of cancers. Rapid, accurate, and cost-effective composition analyses of soyfoods and other health foods are essential for improving the efficiency and quality of health food production. This is the first attempt at developing Fourier Transform Near Infrared Reflectance Spectroscopy (FT-NIRS) calibrations for soy-based and other health foods.

Soy tofu is a traditional soyfood originated from China (Liu et al. 1995). During the course of soybean cultivation, the Chinese had gradually transformed soybeans into various forms of soyfoods, including tofu, soymilk, soy paste, soy sauce and soy sprouts. Along with soybean cultivation, methods of soyfood preparation were gradually spread to Far East and West countries. The art of preparing soyfoods has now spread to the rest of the world, due to agricultural innovation and cultural exchanges. For the past several decades, advances in soybean chemistry and innovation in processing and packaging technology have dramatically modernized traditional ways of preparing soyfoods. As new medical research unveils the health benefits of soyfoods, such as the benefits of isoflavones for women's health, there is no doubt that soyfoods will soon become a part of global culture.

It is well known that protein is the dominant component in tofu. In an early report (Koga et al., 1992), the spectral curve of tofu lees in NIR (1100 to 2500 nm region) was correlated with moisture, crude protein, and fiber contents determined by standard chemical methods, with correlation coefficients of 0.976, 0.830, and 0.865, respectively. Some other researchers studied contribution of the total soybean proteins, the storage proteins [glycinin (11S) and b-conglycinin (7S) fractions] to tofu yield and texture. They analyzed protein contents by using SDS-PAGE (SDS-PAGE) coupled with densitometry and reversed phase-high performance liquid chromatography (RP-HPLC) (Mujoo et al., 2003). In order to measure the soy protein in gel form directly, rapidly and accurately, a novel tofu calibration was developed with a Spectrum One NTS FT-NIRS instrument.

Soymilk is another popular liquid soyfood, in which protein, carbohydrates and water are the three main components (Liu et al. 1995). Protein content in soymilk is usually determined by conventional methods such as chemical analysis and UV-Vis Spectroscopy method (Nielsen 1994). In a previous research on capillary electrophoresis, quantitation of bovine whey proteins in commercial powdered soybean milk was performed by adding bovine whey to its formulation using the calibration method of the external standard (Garcia-Ruiz et al., 1999). These techniques are either time-consuming, or not accurate enough for practical applications. A novel calibration was thus developed here with the Spectrum One NTS FT-NIR instrument to accurately measure protein, fat and carbohydrate contents in soymilk. For such a purpose, a transflectance working mode was employed for spectral data acquisition of soymilk. This mode is usually used for thin layer samples in order to reduce the noise level and baseline shift of spectra. If the NIR spectra of liquid samples such as milk are obtained with the regular transmittance or reflectance mode,

accurate quantitation is almost impossible because of the low S/N ratio caused by light scattering and large baseline shift (Ozaki et al., 2001).

The high dietary intake of soya has been associated with a reduced risk of some cancers such as breast cancer for women and heart disease. Isoflavones (mainly including daidzein, genistein and genistin) may be responsible for the protective role of soya (Liu 1997; Song et al. 1998). Monitoring the levels of isoflavones in health foods such as soymilk appears also to be important in populations that are at risk for certain types of cancers (Liu et al. 1995). Rapid, accurate, and cost-effective composition analyses of soy isoflavones are essential for breeding and genetic selection studies aimed at optimizing soybean seed compositions for human health food applications (Choi et al. 2000; Lee et al. 2003), and improving the efficiency and quality of soy health food production. The determination of isoflavones content is commonly done by HPLC analysis (Carrao et al. 2002; Choi et al. 2000; de-Rijke et al. 2001; Lee et al. 2003; Song et al. 1998; Tekel et al. 1999), or other improved methods with regular liquid chromatography (Kao et al. 2002). The HPLC method for isoflavone measurement is expensive, time-consuming, and impractical for measurements of large number of soybean samples that are required by breeding and selection studies.

Few NIRS studies were, however, reported on the analysis of one or two seeds of wheat—wheat grains-- and no work has been published on measurement of low-level components such as isoflavones in soybeans by NIRS, mainly because of the limited spectral resolution and stability of conventional NIR instruments. In the past five years, however, significant improvements in NIR instrumentation have been achieved through applications of novel technologies such as Diode Array and Fourier Transform (Guo et al., 2002); which thereby provided the potential for single seed analysis of both major components and low-level components of soybeans. In this chapter, rapid and accurate analytical methods for protein, oil, moisture, and isoflavone determinations were developed with state-of-the-art FT-NIR instruments. This is the first attempt at developing Fourier Transform Near Infrared Reflectance Spectroscopy (FT-NIRS) calibrations for isoflavones in soybeans.

2. Calibration and Validation methods

In this section *partial least-squares regression models* are employed to develop FT-NIR calibrations for soybean-based and other health foods, soy tofu and milk, as well as soy isoflavones.

The general procedures for calibration development for the Perkin-Elmer model Spectrum One NTS can be described as follows.

- **Step 1.** Data acquisition with standard calibration samples.
- **Step 2**. Select a wavelength range of the NIR spectrum which is suitable for sample composition determination, based on the major NIR absorption bands of chemical/biochemical components of the samples studied.
- **Step 3**. The use of a "Interactive Baseline Correction" spline-function to correct fisrt the baselines of the NIR spectra, and then normalize such corrected spectra.
- **Step 4.** Matrix calculations with the PLS-1 algorithm in order to optimize the calibration parameters after corrections for light scattering effects by multiple-scattering correction (MSC) method.
- **Step 5.** Generate a calibration file with the optimized calibration parameters and make it available in the instrument control panel for sample measurements.

In order to improve both the accuracy and robustness for calibration development, spectral data sets were selected based on equally distributed analyte concentrations, and also the widest possible concentration ranges of all components were taken to statistically maximize the information content of the NIR spectra (Haaland and Thomas, 1988). Despite the fact that the calibration algorithm was initially designed only for PLS-1 simulations with a three component mixture (N=3) system, it is applicable to real samples with multiple components with N>3. Thus, wide concentration ranges for all components are necessary for the standard samples in order to be able to develop high quality NIRS calibrations. Even though the concentration ranges of all components for real systems may not be comparable, it is necessary to make the concentration range of each component as wide as possible.

2.1. Calibration algorithms

2.1.1. Determining the Number of Factors for the Model.

In fact, PLS-1 is a only a partial subset of the full PLS-2. The algorithms have been combined here, with appropriate notes on the aspects in which they differ. Note also that a PLS-2 model of a training set with only one constituent is identical to a PLS-1 model for the same data. One of the most difficult tasks in using PCR and PLS is determining the correct number of loading

vectors (factors) to use to model the data. As more and more vectors are calculated, they are ordered by the degree of importance to the model (either by variance in PCA or concentration weighted variance in PLS). Eventually the loading vectors will begin to model the system noise (which usually provides the smallest contribution to the data). The earlier vectors in the model are most likely to be the ones related to the constituents of interest, while later vectors generally have less information that is useful for predicting concentration. In fact, if these vectors are included in the model, the predictions can actually be worse than if they were ignored altogether. Thus, decomposing spectra with these techniques and selecting the correct number of loading vectors is a very effective way of filtering out noise. However, if too few vectors are used to construct the model, the prediction accuracy for unknown samples will suffer since not enough terms are being used to model all the spectral variations that compose the constituents of interest. Therefore, it is very important to define a model that contains enough vectors to properly model the components of interest without adding too much contribution from the noise.

Models that include noise vectors or more vectors than are actually necessary to predict the constituents'concentrations are called overfit. Models that do not have enough factors in them are known as underfit. Unfortunately, there is usually no clear indicator of how many factors are required to move from "constituent" vectors into "noise" vectors and prevent both underfitting and overfitting. However, there are a variety of methods that can be used to aid in determining this value. One of the most effective is to calculate the PRESS (Prediction Residual Error Sum of Squares) for every possible factor. This is calculated by building a calibration model with a number of factors, then predicting some samples of known concentration (usually the training set data itself) against the model. The sum of the squared difference between the predicted and known concentrations give the PRESS value for that model.

$$PRESS = \sum_{i=1}^{n} \sum_{j=1}^{m} (CP_{i,j} - C_{i,j})^2 \qquad \text{Eq. (2.1)}$$

In the above equation, n is the number of samples in the training set, and m is the number of constituents. Cp is the matrix of predicted sample concentration from the model, and C is the matrix of known concentrations of the samples. The smaller the PRESS value, the better the model is able to predict the concentration of the calibrated constituents. By calculating the

PRESS value for a model using possible factors and plotting the results, a very clear trend should emerge.

2.1.2. Cross validation.

The cross-validation concept is quite simple, but it is also the most calculationally intensive method of optimizing a model; in effect, cross-validation aims to emulate the prediction of "unknown" samples by using the training set data itself. The procedure is as follows:

- Select a sample (or a small group of samples, if the training set is large enough) and remove the spectrum (spectra) and corresponding concentration data from the data matrix. Set the factor counter to I =1.
- Use the remaining spectra and concentration data of the samples to perform the decomposition and calibration calculations for factor I (loading factor).
- Predict the concentrations of the removed samples(s) using the calibration equation from step 2, and calculate PRESS(I).
- Increase the factor counter (I=I+1) and repeat from step 2 until all desired factors (I=f) have been calculated and predicted.
- Place the previously left our sample data back into the training set and select a different sample (or group). Return to step 1 and repeat the calculations. As each sample is left out, add the calculated squared residual error to all the previous PRESS values. Repeat until all samples have been left out and predicted at least once.

There are two main advantages of cross-validation over all other methods. The first is in how it estimates the performance of the model. Since the predicted samples are not the same as the samples used to build the model, the calculated PRESS value is a very good indication of the error in the accuracy of the model when used to predict "unknown" samples in the future. The larger the training set and the smaller the groups of samples left out in each pass (optimally only one sample at a time, but this can be very time consuming), the better this estimate will be. In effect, the model is validated with a large number of "unknown" samples (since each training sample is left out at least once) without having to measure an entirely new set of data. The second benefit of cross validation is better outlier detection. While this will be discussed in more depth in a later section, it can be mentioned that cross validation is the only validation method that can give complete outlier detection for the training set data. Since each sample is left out of

the models during the cross validation process, it is possible to calculate how well the spectrum matches the model by calculating the spectral reconstruction and comparing it to the original training spectrum (via the spectral residual). If the predicted concentrations for a single sample are way off and the spectrum does not match the model very well but the rest of the data works very well, the sample is possibly an outlier. Identifying and removing outlier samples from the training set should always improve the predictive ability of the model. Only if a complete cross validation is performed, the outlier detection on the training set data can be well performed. Unfortunately, cross validation is a very time consuming process. It requires recalculating the models for every sample left out. However, there are a few somewhat acceptable shortcuts. If the number of samples in the training set is large enough, the number of samples rotated out in each pass can be more than one. This obviously does not give the best statistics for each sample, but it does speed the calculations and can be acceptable for determining the number of factors for the model.

2.1.3. Selecting the Factors Based on SECV.

To avoid building a model that is either overfit or underfit, the number of factors where the PRESS plot reaches a minimum would be the obvious choice of the best model (except in the case of Self-Prediction). While the minimum of the PRESS may be the best choice for predicting the particular set of samples, it is not always optimum for prediction of all unknown samples in the future.

The concept of SECV (Standard Error of Cross Validation) or SEP (Standard Error of Prediction) can be better used to indicate the optimized number of factors, instead of PRESS. The definition of SECV is:

$$SECV = \sqrt{\frac{\sum_{i=1}^{n} (Y_{i(k)} - Y_{i(p)})^2}{n}}$$

$Y_{i(k)}$ is the known concentration, $Y_{i(p)}$ is the predicted concentration. n is the number of samples calculated. Eq. (2.2)

It is rather obvious that SECV is comparable in use to PRESS because SECV is the averaged root mean square of PRESS, and thus it follows the same tendency of variation as PRESS does

(in ThermoNicolet's TQ Analyst program, SECV is also called RMSECV: RM stands for root mean). When PRESS reaches its minimum, SECV reaches its minimum, too. However, SECV represents the prediction error for building the calibration model better than PRESS does because of the actual manner in which the prediction error is computed through averaging. Therefore, one may use SECV plots and values to indicate the optimized number of factors for the choice for the best model. However, for a calibration that is required to be both robust and accurate, it is customary to choose the number of factors corresponding to the minimum in the plot of Log (PRESS) against the number of factors. In **Figure 2.1**, which is the SECV vs. factor plot for soy tofu calibration development, one notices that in the range of number of factors from 0 to 15 factors the SECV decreases as each new factor is added to the model. This indicates that the model is underfitted, and there are not enough factors to completely account for the sample constituents of interest. At some point, the SECV plot should reach a minimum (6) and start to ascend again. At this point, the model is beginning to add factors that contain uncorrelated noise which are not related to the constituents of interest, and therefore one has an overfitted situation. When these extra "noise" vectors are included in the model, it is overfitted, and its predictive ability is rapidly diminishing. The number of factors at the minimum SECV value, e.g. n= 6, thus can be the best choice of prediction in this particular example. The correlation for calculated (predicted) protein percent vs. actual proteiin percent with 6 factors is plotted in **Figure 2.2**, and a correlation coefficient very close to 1.0 (0.999) is reached.

2.1.4. Outlier sample detection

Outlier detection is equally important as choosing the optimum number of factors for the model. If one or more of the training samples are in error, it will cause errors in the calibration model and ultimately poor prediction results for unknowns. Outlier samples usually arise from some incorrect measurement, whether it is in the concentration data (i.e. errors in the primary calibration techniques, transcription errors), or in the spectral data (i.e. spectrometer error, sample handing procedures, environmental control such as temperature, humidity, etc.). Including outlier samples in the training set will introduce a bias to the final model. In effect, outlier samples will tend to "pull" the model in their direction, causing the predicted concentrations of valid samples to be less accurate (or even erroneous) than if the sample was completely eliminated from the training set.

Samples that have significantly larger concentration residuals (difference between the actual and predicted concentrations) than the rest of the training set are known as *concentration outliers*. This type of outlier generally arises when the experimenter either makes a mistake in creating the calibration mixtures or there was an error in the analysis of the samples from the primary calibration techniques used to generate the calibration concentration values. Another possibility which frequently occurs is a transcription error: the analyst simply types in the wrong concentration value when building the computerized training set. Some obvious outliers can be simply picked up by visual inspection. While the human eye is excellent at discerning patterns in data, visual inspection is not always a valid basis for a decision of this type. What is really needed is a mathematical way to accurately determine the likelihood that a sample is really an outlier. For clusters of data points, it is possible to use a measure of the Mahalanobis distance (Mahalanobis, 1936). This is calculated as the distance of the potential outlier sample point as measured from the mean of all the remaining points in the cluster. The distance is scaled for the range of variation in the cluster in all dimensions, and then assigns a probability weight to the sample in terms of standard deviation. Any sample which lies outside of 3 standard deviations from the mean can be considered suspicious, e.g. 3% deviation for soy and health food composition. The Mahalanobis distance is also useful in qualitative analysis of spectral data for which the constituent concentrations are not known.

2.2. Spectra Pre-processing

One of the major problems in applying chemometric models to spectra is the fact that the acquired spectrum of a sample is dependent on many different, sometimes uncontrollable factors. For example, samples of powdered solids are usually measured by diffuse reflectance. **Figure 2.2.1** shows a plot of the *SECV vs. Factor Number* for soybean protein in the calibration development for soy tofu that appears distinct from those of other samples because of the particle size distribution and its alignment with the incident beam of light. While the quantitative information related to the constituents is still contained within the spectral data, it may not be immediately apparent. Another example is that the pathlength of the samples sometimes can not be controlled, such as measuring spectra of thin films.

Figure 2.2.1 Calculated (or predicted) protein% vs. Actual (or reference) protein% plot, with 6 factors, in the calibration development for soy tofu.

Chemometric models can sometimes correct for these effects by adding extra loading vectors, but generally the models will perform better if they can be removed or at least minimized before running the data through the calculations. Since they are applied to the data before it is used in the model, they are often called Preprocessing Algorithms. There are a variety of methods that can be used to remove the non-constituent related aberrations in the data. Most algorithms are targeted at removing a specific interference (MSC, for example, specifically attempts to remove the effects of light scattering). Properly applying preprocessing requires understanding the interference in the data and selecting the appropriate algorithms to correct the effects.

2.2.1 Multiplicative Scatter Correction (MSC)

The NIR detector receives light coming from the sample in form of: diffuse reflectance after absorption, specular reflection and scattered light. Only the diffuse reflectance contains chemical composition information, whereas the latter two do not. Therefore, in order to determine accurately chemical composition from NIR measurements, the light scattering and specular components must be corrected for (Williams and Norris, 1987).

The degree of scattering is dependent on the wavelength of the light that is used, and not uniform throughout the spectrum. Typically, this appears as a baseline shift, tilt and sometimes curvature. It is not simply a matter of measurement errors that light scattering effect may cause. In an early research about scatter-correction for NIR reflectance spectra of meat (Geladi et al., 1985), reflectance for fat shows completely different tendencies (up and down) before and after MSC correction (see Figure 9 on page 498 of the research paper). Therefore, without MSC correction, the raw reflectance or absorbance values will make a totally incorrect calibration, and lead to wrong prediction for unknown samples. The MSC method assumes that the wavelength dependency of the light scattering is different from that of the constituent absorption. Theoretically, by using data from many wavelengths in the spectrum, it should be possible to separate the two.

This method attempts to remove the effects of scattering by linearizing each spectrum to some "ideal" spectrum of the sample (Galactic 1996). MSC calculates the average spectrum from all the data in the training set and uses it as the "ideal" spectrum. Thereafter, the spectral responses in each spectrum are used to calculate a linear regression against the corresponding points in the ideal spectrum. The slope and offset values from this regression are subtracted and ratioed respectively in the original training spectrum to give the MSC corrected spectrum.

$$\overline{A_j} = \sum_{i=1}^{n} A_{i,j}$$

which is the Mean Spectrum: Eq. (2.3)

Linear Regression: Eq. (2.4)

$$A_i = m_i \overline{A} + b_i$$

MSC Correction: Eq. (2.5)

$$A_{i(MSC)} = (A_i - b_i) / m_i$$

In these equations, A is the n by p matrix of training set spectral responses for all the wavelengths, A bar is a l by p matrix of the average responses of all the training set spectra at each wavelength, A_i is a l by p matrix of the responses for a single spectrum in the training set, n is the number of training spectra, and p is the number of wavelengths in the spectra. The m_i and b_i values are the slope and offset coefficients of the linear regression of the mean spectrum vector A bar versus the A_j spectrum vector. By adjusting the slope and offset of the sample spectra to the "ideal" average spectrum, the chemical information is preserved while the differences between the spectra are minimized. Thus, the major source of random variance between them can be removed as much as possible.

2.2.2. Correcting Baseline Effects.

None of the available spectrometers collect always data with an ideal, flat baseline. In order to accurately calculate concentrations, it is necessary to remove the baseline shift effect introduced by the spectrometer, especially by specular reflectance in the reflectance mode for PerkinElmer's NIRS spectrometer model Spectrum One NTS. There are a number of methods used by spectroscopists to remove baseline effects from the spectra they collect. The problem with most methods is that they require the spectroscopist to decide that the baseline is corrected by visual inspection. However; there are some methods which are reasonably automated enough to be used as part of a calibration model, such as Linear Regression Baseline Fitting, Two Point Linear Baseline approach, and Derivatives. In Perkin Elmer's Spectrum program, a special function "Interactive Baseline Correction" is designed for users to correct baseline shift for raw spectra,

and another function "Normalization" is used to normalize spectra so that the absorbance values can be used correctly to fit Beer's Law for matrix calculations.

2.2.3. Computer iteration steps for calibration development with PLS-1.

The calibration involves regression with a Partial Least Squares Type 1 (PLS-1), multi-variate algorithm (Galactic Industries Corporation, 1996). The collection of known data, or chemical composition, for each standard samples, together with the measured data by the instrument are called a calibration set (or training set). Such calibration algorithms as PLS-1 base their predictions of each constituent concentration on changes in the spectral data rather than absolute absorbance values. A simpler algorithm called "NIPALS" is useful to illustrate the iteration procedures followed in PLS-1 as well. The NIPALS algorithm involves two stages: an iterative stage that utilizes just the NIR spectral data and a regression stage that utilizes the laboratory composition data along with the results from the previous stage**.** The first iteration stage begins by computing the difference between each raw spectrum and the mean spectrum, $A_i - \overline{A}$, for the entire calibration set. A set of factors F, or eigenvectors F_i are then iterated by setting such factors at the beginning to be equal to the raw spectra, A_i. Both A and F are represented as tables (or matrices) of the NIR absorbance values at specific wavelengths across the NIR spectrum of soybeans. From these matrices, one calculates tables (or matrices) of scores, S_i, defined as a product of two matrices:

$$S_i = A_i F_i' \qquad \text{Eq. (2.6)}$$

where F_i' is the transposed matrix of the eigenvector F_i . In a second iteration step, the eigenvectors Fi are normalized by dividing through the corresponding eigenvalues, $\lambda_{i,i}$, defined as:

$$\lambda_{i,i} = (\Sigma S_i^2)^{1/2} \qquad \text{Eq. (2.7)}$$

Thus $F_i = A_i / \lambda_{i,i}$ are the normalized eigenvectors at this second iteration step. A new set of scores is then calculated with equation (2.6) from the normalized eigenvectors. The new set of scores is subtracted from the corresponding ones obtained at the first iteration step. The iteration is complete when this difference is zero or negligible. If the difference is significant, one re-computes the eigenvectors F_i through matrix multiplication:

$$F_i = (Ai - \overline{A})' \times S_i \,, \qquad \text{Eq. (2.8)}$$

until the difference between two values of S_i for consecutive iterations becomes zero or negligible. Such optimized scores are effectively the absorbance values of individual constituents at selected wavelengths across the NIR spectrum of the soybeans.

The tables of those score values obtained at the first stage are then employed in a second stage to relate the absorbance values of individual constituents to the known chemical composition stored as a chemical composition table, or matrix, C. The model equation at this stage is therefore:

$$C = B \times S + E_c \qquad \text{Eq. (2.9)}$$

where B is regression coefficient matrix and E_c is a matrix table of regression error terms for chemical composition of the constituents. Once the regression coefficients in martix B are determined, the calibration is complete and can be utilized to predict composition values for the constituents of unknown samples.

In the PLS-1 algorithm, an added sophistication is introduced by utilizing from the first pass of the iteration a linear combination of calibration spectra weighted by the corresponding concentrations of one constituent at a time. In this procedure, the loading vectors (sometimes called "spectral weighing vectors"), are defined as:

$$W_j = C_j^{'} A \qquad \text{Eq. (2.10)}$$

where C_j is the composition vector for constituent j. At the next iteration pass, these spectral weighing vectors are normalized as follows:

$$W_j \text{ (pass 2)} = W_j \text{ (pass 1)} / [\, W_j \text{ (pass1)}\, W_j^{'} \text{ (pass 1)}] \qquad \text{Eq. (2.11)}$$

Therefore, by using loading vectors as eigenfactors, concentration information is included in the calculations during the first spectral decomposition stage rather than in a separate second stage. This is the main difference between PLS and the NIPALS (also the PCR method).

Loading factors are actually mimics of the pure component spectra. The first loading factor in the PLS-1 analysis is a first-order approximation to the pure-component spectrum of the corresponding component. Figure 2.3 gives one graph of the first loading factors for the pure components in SPI and H_2O mixture. The pure component spectra of SPI and H_2O generated by the computer program look exactly the same as their real spectra.

The number of calibration loading factors for each constituent can be obtained for the minimum value of the SECV. However, for a calibration that is required to be both robust and accurate, it

is customary to choose the number of factors corresponding to the minimum in the plot of Log (PRESS) against the number of factors.

2.2.4. Standard Error of Prediction (SEP)

Standard Error of Prediction (SEP) has the same definition as SECV, but the samples for SEP are not involved in the cross validation process for calibration development. The samples for SEP are only used to compare predicted values from the developed calibration with known values for calibration validation purposes.

3. Experimental results and data analysis

3.1. NIR analysis of soy and other health foods

3.1.1. Sampling and experiments

FT-NIRS measurements were carried out in quadruplicate for 16 types of food samples, such as: soy crisps, dry roasted soy nuts, soy burgers, soy tofu, island black beans, soymilk powder, rye cakes, rye bread, rye toast, rye cocktail bread, dry tomato, popcorn minicakes, biscuits and lean ham. Their composition values were calculated according to the nutrition tables on those products and used for calibration data, which are listed in **Table 3.1.** The other standard samples were prepared by either dehydrating or rehydrating some of the original samples. The total number of samples used for this calibration development was 28. FT-NIR spectra were collected over a spectral range from 4000 to 12000 cm^{-1} (833 to 2500 nm) at a resolution of 8 cm^{-1} with a PerkinElmer Co.'s FT-NIR spectrometer, model Spectrum One NTS NTS. This spectrometer is optimized for high-sensitivity analysis of solid samples, being equipped with an NIRA, integrating sphere accessory and an extended range InGaAs detector. The beam size was set to be 8.94 mm. The number of scans was 64 for each spectrum.

Figure 3.1. A graph of loading factors for the pure components in SPI and H_2O mixture.

Table 3.1 Composition values of 16 soy and other health foods calculated according to the nutrition tables on those products.

	Protein %	Fat %	Moisture %	Total Carbohydrates %	Fiber %
Soy crisps	25.0	7.1	0.5	50.0	7.1
Dry roasted soy nuts	43.3	26.7	< 1.0	20.0	13.3
Soy burgers	20.0	4.4	59.7	8.9	5.6
Frida's firm tofu	7.1	3.5	82.2	2.4	1.2
Fried tofu	9.1	8.6	76.1	2.0	1.0
Island black beans	18.8	1.6	5.0	53.1	18.8
Soymilk powder	10.0	5.9	1.7	69.1	0.1
Popcorn minicakes	12.5	6.3	< 1.0	75.0	6.3
Rye cakes	13.0	< 0.1	1.0	60.0	26.0
Rye bread	9.7	4.8	26.0	45.2	6.5
Light rye bread	7.3	3.7	22.0	48.8	2.4
Rye toast	10.0	< 0.1	< 1.0	85.0	3.0
Biscuits	10.0	< 0.1	< 1.0	85.0	2.0
Dry tomato	<1.0	< 0.1	4.0	86.0	9.0
Rye cocktail bread	9.7	4.8	26.0	45.2	6.5
Bohllen lean ham	14.5	14.2	68.8	1.8	< 0.1

3.1.2. Calibration results.

The TQ Analyst software developed by Nicolet Instruments was employed to process NIR spectra and develop calibration files. A total of 112 FT-NIR spectra were preprocessed by applying a suitable Multiplicative Scattering Correction (MSC). Partial Least Squares Type 1 (PLS-1) multivariate regression analyses were employed for high-quality calibration model developments. **Figure 3.2** shows an overlay of group spectra for soy and other health foods obtained with Spectrum One NTS after baseline correction and normalization.

Figure 3.2. Overlay of FT-NIR Reflectance spectra for soy and other health foods obtained with Spectrum One NTS.

Standard composition values of major food components, such as: protein, fat, moisture, fiber, total carbohydrates were obtained from nutrition tables on those products. Composition changes of soy and other health foods caused by microwave heating or moisture rehydration were also monitored. The composition ranges for calibration development are: protein 0.5% to 43.3%, fat 0.1% to 26.7%, moisture 0.5 to 82.2%, fiber 0.1% to 26%, total carbohydrates 0.5% to 95%. These are quite wide concentration ranges and cover almost all of the soy and other health foods contents. The optimized parameters for the calibration result are listed in **Table 3.2.**

Table 3.2. Optimized SECV, R^2 values and number of factors for the calibrations developed on Spectrum One NTS, Wavelength range 4080 to 11200 cm^{-1}.

	Protein %	Fat %	Moisture %	Total Carbohydrates %	Fiber %
SECV	1.2	0.7	1.4	1.7	1.0
R^2	0.992	0.994	0.995	0.995	0.985
# Factors	12	12	13	14	13
SEP	1.4	1.0	1.6	1.7	1.3

This calibration for soy and other health foods is characterized by low standard errors (~1%) and high degrees of correlation between NIR calculated values and laboratory reference values (~99%). It will satisfy commercial determination of nutritional contents in soy and other health foods. The purpose of developing this calibration is to introduce a new experimental method for rapidly and accurately measuring different types of soy and other health foods. The results were reported as (see Appendix) "Determination of Soy and Other Health Foods Composition by Fourier Transform Near Infrared Reflectance Spectroscopy", Jun Guo and Ion C. Baianu, Proceedings for the 9th Biennial Conference of the Cellular and Molecular Biology of the Soybean, August 11-14, 2002, P506.

3.2. NIR analysis of soy tofu

3.2.1. Sampling and experiments.

FT-NIRS measurements were carried out in quadruplicate for 19 tofu samples with different protein and water contents. The original tofu sample was a commercial product Fridas' Firm

Tofu, with 7.1% protein, 82.2% water, and ~10% other total solid components such as fat, salts and carbohydrates. The other samples were prepared by short time microwave heating with an interval of 20 seconds, so that the water in tofu could be lost gradually and the protein content increased accordingly. The total number of samples used for this calibration development was 24. The composition values were calculated according to the amount of water loss. The composition ranges for calibration development are: protein 7.1% to 39.8%, moisture 27.1% to 82.2%, and other total solids 10.7% to 33.1%. These are quite wide concentration ranges and cover almost all of the soft and firm tofu contents. FT-NIR spectra were collected over a spectral range from 4000 to 12000 cm^{-1} (833 to 2500 nm) at a resolution of 8 cm^{-1} with Spectrum One NTS. The beam size was set at 8.94 mm. The number of scans was 64 for each spectrum.

3.2.2. Calibration results.

The TQ Analyst software was employed to process NIR spectra and develop calibration files. Totally 96 FT-NIR spectra (shown in

Figure 3.5) were preprocessed by applying a suitable Multiplicative Scattering Correction (MSC). Partial Least Squares Type 1 (PLS-1) multivariate regression analyses were employed for high-quality calibration model developments. The optimized parameters for the calibration result are listed below in **Table 3.3.**

Figure 3.5. Overlay of FT-NIR Reflectance spectra for soy tofu obtained with Spectrum One.

Table 3.3. Optimized SECV, R^2 values and number of factors for the calibrations developed on Spectrum One NTS, wavelength range 4080 to 11000 cm^{-1}.

	Protein %	Moisture %
SECV	0.75	1.19
R^2	0.999	0.998
# Factors	10	8
SEP	0.83	1.35

The calibration for soy tofu is characterized by low standard errors (~1%) and high degrees of correlation between NIR calculated values and laboratory reference values (>99%), and can be used to measure protein content in tofu.

3.3. NIR analysis of soybean milk

3.3.1. Sampling and experiments.

FT-NIRS measurements were carried out in quadruplicate for 27 soymilk samples with different protein and water contents. The liquid soymilk samples were made from a commercial soymilk powder product Mount Elephant Soybean Drink (Guangxi Cereal and Oil Product Company, Wuzhou City, Guangxi Province, China), with 10% protein and 69% carbohydrates. After mixing the soymilk powder with different portions of water, liquid soymilk samples were prepared for different concentrations. FT-NIR spectra were collected over a spectral range from 4000 to 12000 cm^{-1} (833 to 2500 nm) at a resolution of 8 cm^{-1} with Spectrum One NTS. The beam size was set to be 8.94 mm. The number of scans was 64 for each spectral accumulation. Due to the fact that water is the dominant component in soymilk, protein bands on the soymilk spectra are overlapped by huge water bands. In order to get as much chemical information of the other components except water as possible, a specially designed metal reflector was used to obtain the transflectance spectra. Only 5 μl of liquid sample was put onto the instrument each time, with the reflector covered on top of the liquid layer, in order not to lose diffuse reflectance signals.

3.3.2. Calibration results.

The TQ Analyst software was employed to process NIR spectra and develop calibration files. A total of 108 FT-NIR spectra were recorded as shown in **Figure 3.6,** and the spectra were preprocessed by applying a suitable Multiplicative Scattering Correction (MSC). Partial Least Squares Type 1 (PLS-1) multivariate regression analyses were employed to develop high-quality calibration models. The composition ranges for calibration development were: protein 0.5% to 10%, water 1.7% to 100%, and carbohydrates 3.5% to 69.1%. These are quite wide concentration ranges and cover almost all of the soymilk and even tofu contents. The optimized parameters for the calibration result are listed in **Table 3.4.**

Table 3.4. Optimized SECV, R^2 values and number of factors for the calibration of soymilk developed on Spectrum One NTS, wavelength range 4080 to 11500 cm^{-1}.

	Protein %	Fat %	Moisture %	Carbohydrates%
SECV	0.03	0.02	0.34	0.23
R^2	0.999	0.999	0.999	0.999
# Factors	9	9	9	9
SEP	0.08	0.04	0.73	0.53

Our calibration for soy milk has achieved low standard errors, especially for protein and fat (**<0.1%**), as well as high degrees of correlation between the NIR calculated values and the laboratory reference values (~99%) obtained with the primary methods. It is suitable for measuring soymilk within regular concentration ranges and beyond. The results were reported as (see Appendix A) "Rapid Determinations of Soybean Isoflavones, Soy and Other Health Foods Composition by Fourier Transform Near Infrared Reflectance Spectroscopy", Jun Guo and Ion C. Baianu, *Proceedings of the China and International Soy Conference and Exhibition 2002* (*CISCE 2002*), November 6-9, 2002, 391-392.

Figure 3.6. Overlay of 108 FT-NIR Transflectance spectra for soymilk obtained with Spectrum One NTS .

3.4. NIR analysis of soybean isoflavones.

3.4.1. Sampling and experiments.

In order to develop NIR calibrations on such instruments for soybean composition analysis, soybean standard samples were selected from the USDA Soybean Germplasm Collection (Urbana, IL, USA). The selection of standard samples was based on their protein, oil, moisture, and isoflavone contents, to ensure that the ranges of standard sample constituent contents covered the full range of possible constituent variations of samples. To minimize screening effects of the soybean seed coat (especially black and brown coat) on isoflavones, soybean seeds were ground for preparation of standard samples. Twenty eight ground soybean samples from isoflavone standards plus one isoflavone tablet sample (NovaSoy tablet) were utilized in the calibration development, with isoflavones range from 0.04% to 0.9% (HPLC data), protein range from 34% to 47.1% (ZX-50 data), oil range from 12.8% to 23% (ZX-50 data), and moisture range from 5.6% to 11.0% (ZX-50 data). Laboratory reference values of isoflavone composition were obtained by HPLC analyses of soybeans, which were kindly provided by Dr. J. Widholm's laboratory at UIUC. The TQ Analyst software was employed to process NIR spectra and develop calibration files. Totally 116 FT-NIR spectra were preprocessed by applying a suitable Multiplicative Scattering Correction (MSC). Partial Least Squares Type 1 (PLS-1) multivariate regression analyses were employed for high-quality calibration model developments. The samples were ground with a Braun KSM2B Grinder. The average grinding time is 25 seconds, producing a powder sample with a particle size ranging from 100 um to 200 um. Quadruplet FT-NIRS measurements were carried out for the 29 isoflavone samples with a weight of 300 mg each (two soybean seeds). FT-NIR spectra were collected over a spectral range from 4000 to 12000 cm^{-1} (833 to 2500 nm) at a resolution of 8 cm^{-1} with Spectrum One NTS. The beam size was set at 8.94 mm. The number of scans was 32 for each accumulated spectrum.

3.4.2. Calibration results.

We present in **Figure 3.7** an overlay of FT-NIR spectra of ground soybeans for isoflavones standards. They are baseline corrected and normalized. A calibration was developed based on these spectra. Standard composition values were obtained with ZX-50 instrument for protein, oil, moisture and HPLC data for isoflavones. The optimized parameters for the calibration result are listed below in **Table 3.5**.

Figure 3.8 presents the calibration plot for calculated isoflavones% vs. actual (or reference) isoflavones%, with 9 factors. The correlation coefficient R and SECV (RMSEC) values are also listed in the figure.

Table 3.5. Optimized SECV, R^2 values and number of factors for the calibration of soybean isoflavones developed on Spectrum One NTS, wavelength range 4100 to 10625 cm^{-1}.

	Protein%	**Oil%**	**H_2O %**	**Isoflavones%**
SECV	0.67	0.28	0.12	0.0146
R^2	0.989	0.994	0.995	0.997
Number of Factors	6	9	9	9
SEP	0.43	0.32	0.26	0.0172

Our first calibrations for soybean isoflavones are characterized by outstandingly low standard errors (**<0.02%**), as well as high degrees of correlation between NIR calculated values and laboratory reference values (>99% in most cases). For soybean samples containing a normal isoflavone content, i.e. 0.2% to 0.9%, the calibration is accurately applicable. For soybean samples containing a low isoflavone content, i.e. 0.04% to 0.2%, the calibration can roughly predict the isoflavone concentration. The accuracy of this calibration is comparable with that of a recently published calibration for soybean isoflavones developed with single half soybean seeds (You et al., 2002). The results were reported as (see Appendix) (1) "Rapid Determinations of Soybean Isoflavones, Soy and Other Health Foods Composition by Fourier Transform Near Infrared Reflectance Spectroscopy", Jun Guo and Ion C. Baianu, Proceedings of the China and International Soy Conference and Exhibition 2002 (CISCE 2002), November 6-9, 2002, 391-392.

Figure 3.7. Overlay of FT-NIR Reflectance spectra for soybean isoflavone standards obtained with Spectrum One NTS

Figure 3.8. The calibration plot for calculated isoflavones percentage vs. actual (or reference) isoflavones percentage, with 9 factors.

4. Conclusions

Perkin Elmer's model FT-NIR spectrometer can be utilized for accurate measurements of food protein, oil (fat), carbohydrate and fiber contents for both solid and liquid samples, as well as isoflavones contents in soybean powder samples. It can be also employed to obtain detailed characterization of foods and to investigate the interactions between major food components such as: protein, oil, water and carbohydrates. Moreover, fast and economical measurements of food composition allow online quality control and chemical analysis in food production. Calibration transfers are also possible between different instruments of the same model SpectrumOne NTS.