Determination of LC-PUFAs and Protein Content in Premature infant formula powders by ATR-FTIR Spectroscopy and Chemometrics

Mahyar Sanadgol Nezami†, Alireza Feizbakhsh*†, Amir Bagheri Garmarudi‡

† Department of Chemistry, Faculty of Science, Central Tehran Branch, Islamic Azad University, Tehran, Iran;
‡ Department of Chemistry, Faculty of Science, Imam Khomeini International University, Qazvin, Iran;
*e-mail: ali.fb1970@yahoo.com

Recieved: May 04, 2021; Accepted: November 02, 2021
DOI: 10.17721/moca.2022.17-27

Premature infant formula powder (PIFP) contains essential nutrients, such as protein and LC-PUFAs in which inadequate quantities of them may restrict the development of premature infants. The aim of this study was to propose a quantitative analysis for evaluating these nutrients in PIFP. To achieve this purpose, an integrated methodology was developed by combining attenuated total reflectance Fourier-transform infrared spectroscopy (ATR-FTIR) with the chemometrics. ATR-FTIR was utilized to obtain the PIFP spectra in the wavenumber range of 4000-400 cm^{-1}. Principal component analysis (PCA) was employed to perform exploratory analysis of the spectra. Partial least square regression (PLSR) and principal component regression (PCR) models were established for normal spectra, first, and second derivatives to quantitatively determine the protein and LC-PUFAs. The results showed that the PLSR model provided a better prediction than the PCR model with the coefficient of determinations (R^2) of 0.992 and 0.983 for training and validation, respectively. It also offers the best prediction for the PLSR model of protein considering the parameters such as R^2 (normal spectra) (Cal: 0.995, Val: 0.981). The results revealed that a combination of the ATR-FTIR spectroscopy and the chemometrics models creates a rapid, inexpensive, and non-destructive method for the quantitative analysis of PIFP.

Keywords: ATR-FTIR spectroscopy, chemometrics, LC-PUFAs, premature infant formula powder, protein

According to the definition of the international organizations (e.g., the American Academy of Pediatrics (AAP)), babies born before 37 weeks of pregnancy are called premature infants [1]. These infants need special nutrition in comparison with other healthy infants [2]. This issue is due to the lack of development of the immune system, gastrointestinal tract, and respiratory system so that their malnutrition can occasionally lead to their death [3]. Previous investigations reported that the annual mortality rate of premature infants is significantly higher in Iran country than global statistics [4]. If the infants are fully fed, (i.e., receiving all the required micronutrients), the growth rate will increase, and they become a healthy infant [5]. Breast milk has been suggested as the most complete food source, and it is rich in all micronutrients needed by premature infants [6]. The World Health Organization (WHO) offers the premature infant formula powder (PIFP) as the best alternative in the lack of breastfeeding [7]. PIFP contains a variety of micronutrients, including protein, fatty acids, lactose, as well as amino acids, and urea [8]. Among them, the protein and fatty acids are the most prominent micronutrients [9]. A premature infant can have adequate growth and quickly become a term infant by receiving 3.2-3.6 g/100 kcal of protein per day [10, 11]. As counted, long-chain polyunsaturated fatty acids (LC-PUFAs) are taken into consideration as the most important micronutrients in PIFP. The most significant LC-PUFAs are comprised of Arachidonic acid (AA) (20:4n-6) and Docosahexaenoic acid (DHA) (22:6n-3). AA and DHA play a key role in the biological systems of infants. These are realized as the main fats in the biological membrane of their cells. These compounds are known as the rich sources of energy so that the presence of appropriate amounts of them in the diet of premature infants will lead to an increase in the cognitive development and visual acuity [12]. Several PIFP manufacturers attempted to supply AA, DHA, and protein products in recent decades that adhered to international standards. Thus, it is essential to employ quality control standards to accurately measure available micronutrients in these products in PIFPs. PIFP production processes can be monitored by simple, inexpensive, and rapid analytical methods. Nowadays, there are several available methods for the quantitative analysis of minerals and micronutrients in PIFP [13]. The Kjeldahl method is considered as one of the methods, which is defined as a reference method for measuring the protein content in food industries. In this approach, the total nitrogen content is measured without identifying its sources [14]. However, the most important disadvantage of this method refers to the issue that it takes a long time to be

performed [15]. There are other quantitative methods such as atomic absorption (AA), gas chromatography (GC), and inductively coupled plasma-optical emission spectrometry (ICP-OES), which can be utilized for measuring the amounts of AA and DHA in PIFP [16,17]. These methods have the capability of measuring the elements at ultra-trace levels, but their shortcomings are involved in the high cost, the complexity of the sample preparation process, and its time-consuming process. These issues have been led to the limited application of these methods.

Recently, many researchers have focused on developing vibrational spectroscopy methods such as Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) [18, 19]. These methods are considered as a powerful analytical technique for monitoring and controlling processes in the food industry [20]. This technique can analyse samples in any physical state (i.e., solid, liquid, and gas) with the shortest time and the least cost, and without the demand for sample preparation steps. Furthermore, a combination of the ATR-FTIR method and multivariate analysis techniques can be used for measuring the nutrients in food [21-23]. For example, FT-IR spectroscopy and chemometrics methods were widely employed for evaluating the composition of fatty acids and their characterizations in vegetable oils [24]. Also, some investigations have analysed the DHA levels in pork fat, and the total fatty acids have been predicted by integrating the ATR-FTIR spectroscopy and multivariate analysis methods [25]. Another study utilized the FT-IR spectroscopy to predict a major source of mineral (i.e., calcium) in milk with moderate accuracy (R^2: 0.58-0.68, RMSECVs: 120-153 mg/kg) [26]. Several studies have utilized the vibration spectroscopy to quantitatively measure the LC-PUFAs and protein content in food. Nevertheless, we have not found any research for the quantitative analysis of AA, DHA, and protein content in PIFP by combining the ATR-FTIR method with chemometrics. As mentioned above, the ATR-FTIR method alone cannot directly detect and measure the amounts of AA, DHA, and protein in PIFP. In this research, partial least squares regression (PLSR) and principal component regression (PCR) were established as two widely applied regression modelling techniques for normal and transformed (first and second derivatives) spectra. Then, the results of the two techniques were compared to each other. Therefore, the ATR-FTIR spectroscopy was combined with the mathematical models of chemometrics to quantitatively determine the AA, DHA, and protein content in PIFP. The ATR-FTIR spectroscopy is taken into account as a rapid, simple, inexpensive, and environmentally friendly method without demand for sample preparation.

Experimental part

Sample preparation. In this study, all PIFP samples were purchased from pharmacies in three cities of Iran (Tehran, Zahedan, and Mashhad). Two well-known and widely used brands of PIFP (brand A: high amounts of AA, DHA, and protein, brand B: low amounts of AA, DHA, and protein) in the Iranian market having the analytes we wanted to analyse were selected and 40 samples of brand A and 50 samples of brand B were supplied. To select the samples as randomly as possible, attempts were made to purchase all 90 samples with different production dates Two-thirds of the samples were included in the training set and the rest of them were used in the validation set. To be at the same temperature as the environment, the samples were all stored in the laboratory at 20 °C for 24 h before testing. To observe the effect of instrument conditions all samples were divided into 2 groups and their spectra were obtained at different times. To develop a quantitative model, 60 samples were randomly selected as the training set and 30 remaining samples served as the validation set. Samples were assigned by labels PIFP(A) and PIFP(B).

Reference methods. In the first step, the Kjeldahl method was employed to analyse the protein content by three times the analysis of a protein from PIFP samples. The factor value of 6.38 was applied to transform the values of nitrogen into protein using 0.1 g PIFP with a 2.0 g catalyst. The digestion process was carried out for 150 minutes at 350°C with 4 ml of sulfuric acid (H_2SO_4). Then, the distillation process was conducted by adding 15 ml sodium hydroxide (NaOH) to the sample after titration with the hydrochloric acid (HCl). Content values are in the form of weight per 100 grams of PIFP (wt%). The official method of the *Association of Official Analytical Chemists (AOAC)* is commonly considered as a reference method for measuring LC-PUFAs in infant formula samples. This official method is based on the GC method. In this study, this method was also employed to quantitatively measure the amounts of AA and DHA in PIFP samples [27]. Table 1 gives the range of actual values for the LC-PUFAs and protein content in PIFP samples.

Table 1. Actual values of nutritional parameters of PIFP.

Nutritional parameters	Actual values range (per 100g powder)
Protein	(22- 32%) or (13-16g)
LC-PUFAs (AA and DHA)	(0.5- 1.0%) or (0.2- 0.4g)

Instrument and measurement. Mid-FTIR spectra were collected from each PIFP sample using a Bruker© Tensor Ⅱ FTIR spectrometer (Germany) with a fiber optic sampling module and then supported by OPUS software. The spectrometer is equipped with the highest sensitivity and stability of an ADC 24Bit DigiTech™ detector. The spectra were scanned in the wavenumber range of 4000-400 cm^{-1} with 270 scans and a resolution of 4 cm^{-1} used for chemometrics

analysis. Collecting the PIFP sample spectrum took 2 minutes. Air black reference was used for conducting background calibration before each measurement. Ethanol (99%) was used for cleaning after and before each measurement to clear the ATR crystal to remove residues between measurements.

Spectral pretreatment. Figure 1 depicts the raw Mid-FTIR spectra for all PIFP samples in the range of 4000-400 cm^{-1}. As shown in the figure, the spectra are noisy and exhibit the systematic variations in the spectral baseline. Here, spectral pretreatments were used for increasing the accuracy of the prediction of chemometrics models, as well as reducing the noise, the effect of powder size, and baseline drift. For this purpose, several spectral pretreatment methods were employed, including derivatives, Savitzky-Golay smoothing, and standard normal variate (SNV). Then, the results of these methods were compared to each other. SNV is a transformation, which is normally applied to the spectral data. In this transformation, the scattering effects are removed by scaling and centering each spectrum [28]. Spectral smoothening was accomplished to reduce the noise. The spectral data are baseline adjusted to avoid the baseline offset and drift generated by light scattering particles in the instrument. In this study, Savitzky-Golay first and second derivatives (fifth-degree polynomial, and seven-point smoothening) were used to improve the baseline shifts and enhance the spectral resolution [29].

Chemometrics analysis. Unscrambler® X 10.4 (CAMO AS, Oslo, Norway) was used for the chemometrics analysis in Windows 10. To evaluate the calibration model, we split the PIFP data into validation and training sample sets to evaluate the calibration model. The training set contains 60 samples and will be used to build the calibration model. The validation sample set contains 30 samples and will be used to assess the effectiveness of the created calibration model. The descriptive data for the PIFP samples utilized in this research are shown in Table 2 [30]. Specific fingerprint regions were identified by exploring different peaks in the raw spectra plotted. Spectral regions dominated by noise were removed.

Table 2. Descriptive statistics of protein and LC-PUFAs content in PIFP samples.

	Item	Training sample set	Validation sample set
% Protein	Number of samples	60	30
	Mean	26.31	27.12
	Standard deviation	4.92	5.31
% LCPUFAs	Number of samples	60	30
	Mean	0.74	0.81
	Standard deviation	0.29	0.37

Principal component analysis (PCA). PCA is an unsupervised exploratory method which is generally used for the discrimination of different groups of samples and is usually performed before other complex classification or prediction methods. It also is an exploratory technique with the ability to convert large data sets of possibly correlated variables into a few new uncorrelated variables (principal components) which could retain maximum variance in the data set [31]. The results of PCA can be introduced as score plots, representing loading plots and principal groupings, indicating the relationship between the observations [32].

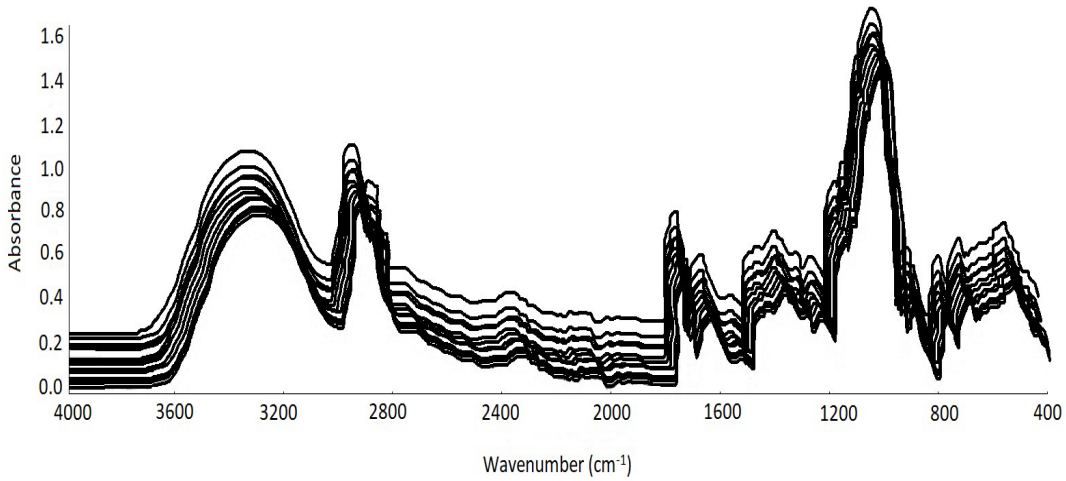

Fig. 1. Mid-FTIR spectra for all PIFP samples.

Quantitative models. Two popular methods of multivariate linear calibration are comprised of partial least squares regression (PLSR) and principal component regression (PCR) in chemometrics generally and spectroscopy particularly. Principal component regression is a linear regression model based on principal component analysis. The target is regressed using the main components of the feature matrix, i.e., altered features, which may subsequently be used to recreate the original data. A model of PCR is expressed as follows:

$$X = S_X \cdot F_X + E_X, \quad (1)$$

$$Y = S_X \cdot C + E, \quad (2)$$

where Equation (1) denotes the decomposition of PCA; X denotes the original spectrum matrix; and S_X and F_X denote the score matrix and spectral loadings, respectively, of X. Equation (2) denotes the final regression model, where Y is the concentration matrix in gas mixtures, and C denotes the regression coefficient. E_X and E indicate residual errors in the two equations, respectively.

Another helpful model for the quantitative analysis of complex spectra is the PLSR model. Unlike PCR, PLSR extracts latent variables from the target Y and the original spectrum X and then builds a regression model between the latent variables. The following procedures are used to implement the PLSR models:

$$Y = S_Y \cdot F_Y + E_Y, \quad (3)$$

The latent variables for X are still retrieved using Equation (1), and those for Y using Equation (3), where F_Y and S_Y denote the loadings and scoring matrices, respectively, and E_Y denotes the residual error matrix. Regression modeling Assuming that the latent variable matrices (S_Y and S_X) are positively linked, a regression model explaining this connection can be constructed as follows:

$$S_Y = S_X \cdot C + E, \quad (4)$$

whereby C is the matrix reflecting the regression coefficients between S_X and S_Y, and E is the residual error.

From the description above, it is evident that the aim of PLSR modelling is to decompose both X and Y into two loadings and scores matrices, and build a regression model between the score matrices. of X and Y with maximum covariance. Based on the PCR and PLSR described in Equations (2)–(4), it is evident that these two models are linear in describing the relationship between concentration and spectral signals, and that by using the obtained regression coefficients one may predict the LCPUFA and protein content within an unknown sample.

In this study, both methods are used for quantitative analysis and prediction of AA, DHA, and protein content in PIFP samples. In many cases, PLSR and PCR methods are similar and there is always a theoretical relationship between them. PLSR and PCR decompose data into loading matrices and spectral scores before model development by new variables. PLSR utilizes concentration and spectral data, whereas PCR decomposes data using only spectral information.

The performance of the model was evaluated based on the statistical parameters, which allow for quantifying the model. These parameters were low relative predictive error (RE%), root mean standard error (RMSE), and high coefficient of determination (R^2). RMSE is the residuals' standard deviation. The term "residuals" refers to the distance between the data points on the regression line, and the term "RMSE" refers to the spread of the residuals. RSME is often used to evaluate experimental findings in regression analysis, forecasting, and climatology. R^2 is a parameter which determines and assesses the ability of a statistical model to explain and predict future outcomes. In other words, if we have dependent variable (Y) and independent variable (X) in a model, then R^2 helps in determining the variation in (Y) by variation (X). The value of R^2 lies between 0 and 1 and higher the value of R^2, better will be the prediction and strength of the model. The R^2 equation can also be found with the following formula (equation 5), where (RSS) is the residuals sum of squares and (TSS) is the total sum of squares.

$$R^2 = 1 - RSS/TSS, \quad (5)$$

The accuracy of the proposed model is affected by factors, such as bias, standard error of prediction (SEP, which should be equally low), and residual predictive deviation (RPD) [33]. The SEP evaluates the performance of the prediction at the validation step, while the bias is the difference between the measured and predicted values [34]. The SEP describes the deviation in observations made around the regression line that has been calculated. Similar to the standard deviation, which indicates the variation in a collection of data relative to its mean, the SEP indicates the variation in the real values of (Y) relative to the predicted values of (Y) on the regression line, and it is expressed as a standard deviation, where the deviations represent the vertical distance between each dot and the mean relationship line. It determines the accuracy of regression line predictions. The RPD represents the predicted model's precise behavior. Williams interpreted six levels of RPD, including (a) very poor prediction (less than 2.3), (b) poor prediction (2.4-3.0), (c) fair prediction (3.1-4.9), (d) good prediction (5.0-6.4), (e) very good prediction (6.5-8.0), and (f) excellent prediction (more than 8.0) [35]. RMSE and RE% were measured by using Equations (6) and (7). Here, n represents the number of samples in the dataset, yi is the reference values, and ŷi characterizes the model-predicted values [36, 37]. Equations (8) - (10) express the bias, SEP, and RPD parameters. In these equations, $\check{y}$ is the mean of the predicted values, $\bar{y}$ is the mean of the measured values, and SD indicates the standard deviation from the reference values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (6)$$

$$RE\% = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i)^2}} \times 100 \quad (7)$$

$$bias = (\check{y} - \bar{y}) \quad (8)$$

$$SEP = \sqrt{RMSE^2_{Validation} - bias^2} \quad (9)$$

$$RPD = SD/SEP \quad (10)$$

On the other hand, there are other statistical parameters to evaluate the performance of chemometric models, such as root mean square error of prediction (RMSEP), root mean square error of calibration (RMSEC), and root mean square error of cross validation (RMSECV). RMSEP is widely used as a criterion for judging the performance of a multivariate calibration model, often it is even the sole criterion. RMSEC measures the average difference between the predicted and the reference values and thus, gives an overall view of the fit of the model (how well the model predicts the same samples that were used to calculate the model). In RMSECV, errors are calculated on training set splits using a cross validation scheme for the splitting. If the splitting of the data is done correctly, this gives a good prediction on how the model built on the data set at hand performs for unknown cases.

Results and Discussion

ATR-FTIR Spectral analysis. The ATR-FTIR spectra of PIFP samples with the corresponding bands are shown in Figure 2. The N-H stretching band of 3500-3400 cm^{-1} forms the primary amino acids. Moreover, the region at band 3000-2800 cm^{-1} is due to the stretching of CH_2 fatty acids in milk. As we know, protein has amino acid chains. These chains are connected by amide bonds. Hence, there are three important areas for identifying functional groups of proteins in their IR spectrum. The peaks at 1450-1200 cm^{-1}, 1650-1600 cm^{-1}, and 1750-1690 cm^{-1} can be attributed to the functional groups of C-N stretching, amide I and C=O, and triglycerides (C=O), respectively. Mono and polysaccharides dominate the region at band 1100-750 cm^{-1} (sugar related C-O-C stretch as lactose) [38]. Based on mentioned matters, regions used for chemometrics analysis of the desired analytes were in a way that 2800-3000 cm^{-1} region was used for analyzing LCPUFA and 1200-1650 cm^{-1} region that encovered amide (I) group was used to study protein in PIFP samples.

Chemometrics

Principal component analysis (PCA). PCA is considered as one of the standard tools for chemometrics analysis. It is used to compress data, noise reduction and evaluate the potential of the discrimination of samples based on their quality. In this study, the PCA method is employed for exploratory analysis to achieve information about the possible trends in different samples. It is taken into account as a bilinear decomposition method that can compact the vast quantities of data into several parameters, which are referred to as principal components (PCs). The PCs capture the differences and similarities between the samples and variables constituting the modeled data. This task is performed by a linear transformation under certain constraints such as preserving data variance. In PCA, to construct new dimensions of spectral data, a new set of data resulted from decreasing the total number of parameters to two principal components that show the highest data variance in another space is applied. To determine the number of PCs that is required for principal components' regression, data were separated into two training set and validation set and then, using the singular value decomposition (SVD) algorithm, the eigenvalue for all engaged PCs in the method was obtained in a way that sum of two first and second components described the highest variances of training set data (Table 3). In the end, these two components were used to apply main component regression (Figure 3).

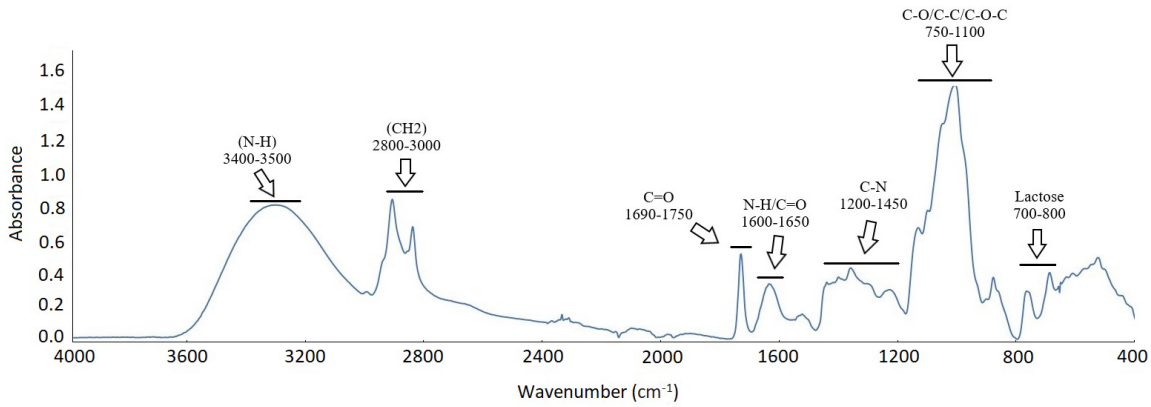

Fig. 2. ATR-FTIR spectra of PIFP sample in mid-infrared range (4000-400 cm^{-1}).

Table 3. Eigenvalues and variance percentage for principal component analysis.

Number of principal components Total	Eigenvalues	% Variance
1	3.9	68.2%
2	1.1	18.2%
3	0.5	7.4%
4	0.4	6.2%

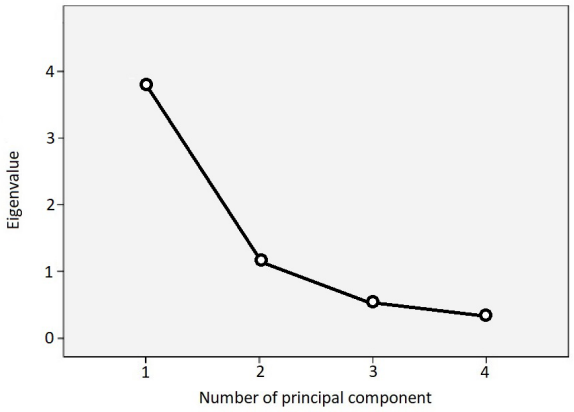

Fig. 3. Scree plot of PCA components and their eigenvalues.

To assess PCA in separating spectrums obtained from PIFP samples, their 2d plot were plotted for PC1 and PC2. the principal components that were applied for optimizing parameters in PCA were employed using the leave on out cross validation (LOO-CV) method. This method was conducted by removing a sample in each calibration stage, and then modelling was performed using the remaining samples. This procedure occurred for every sample in the training set. Figure 4 demonstrates the PC2×PC1 score plot derived from applying PCA to derivative spectra. It can be concluded that the two classes of PIFP are not overlapped, implying the transfer of suitable information for classification by the Mid-FTIR spectrum. A good separation was acquired between the samples with significant differences between brand A and brand B samples. PC1 and PC2 provide the most discriminating separation of the two groups of PIFP samples. Brand (B) scores highly on PC1, while brand (A) scores highly on PC2. This implies that the left-hand group of brands (A) and the right-hand group of brands (B) have distinct spectrum patterns.

PC1 represented a variance of 68.2 %, while PC2 showed a variance of 18.2 %, and therefore the total variances described by PC1 and PC2 were found to be 86.4 %.

PCA and clustering methods are unsupervised methods that classify samples into classes that are not identified beforehand. The aim of the analysis of multivariable is to decrease data. In most cases, there is a correlation between parameters and subsequently, some of the information is additional. PCA is an approach for decreasing data size during correlation. In this research, PC1 is selected in a way that the most changes occurred in data class and then, PC2 justified the most changes and then it continued in this way. As a result, when there is a meaningful correlation, the number of principal components will be a lot less than the number of main variables. After correcting of baseline and noise suppression, the PCA operation is

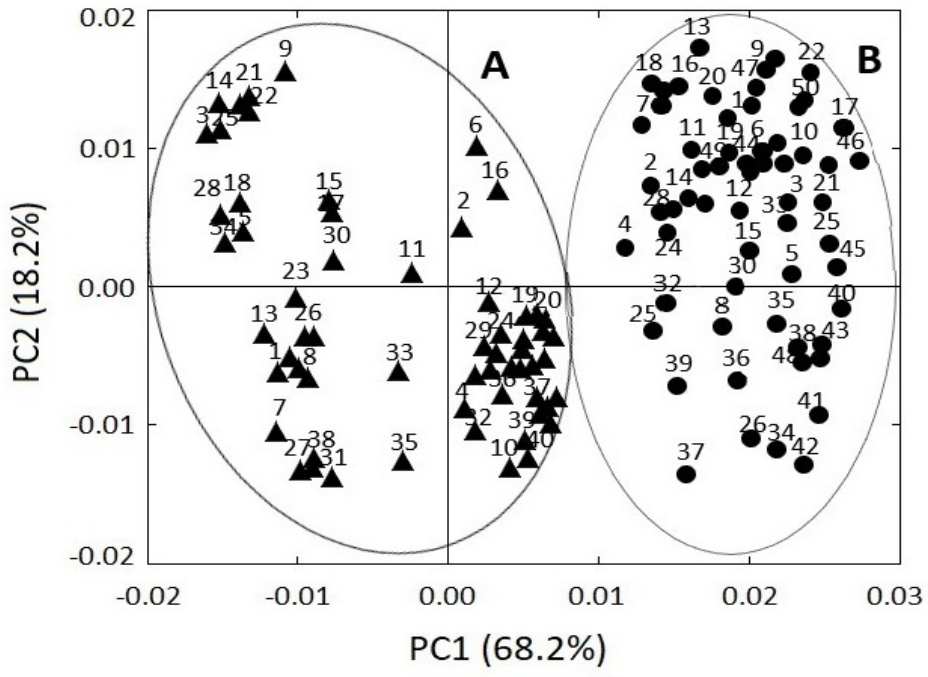

Fig. 4. PC2×PC1 biplot illustrating clusters of 90 PIFP samples (▲: brand A; ●: brand B).

conducted on data and data variances for PCs were obtained. After conducting preprocessing methods for decreasing data noise, we performed PCA once more that the 2d diagrams for PC1 and PC2 are shown. As shown in Figure 4, about 68.2 % and 18.2 % of data were allocated to PC1 and PC2, respectively. Thus, there was a high correlation between data. PCA showed acceptable results in demonstrating possible clusters in samples and could perform properly.

Training and validation of the quantitative models. The PLSR and PCR models include a training step in which the relationship between component concentration and spectra (Mid-FTIR) is estimated from a prediction step and a group of reference samples. The concentration of each component is approximated from an unknown sample spectrum with the reference to the calibration results. In this research, to create an ideal model, we need to estimate parameters and the number of them. This means that by optimizing the number of effective parameters in model construction, we evaluate the model performance for out-of-sample data. In this study, we used the LOO-CV method for optimizing the number of effective parameters. In this method, we utilized samples in the training set, which included 60 randomly selected samples of PIFB sample. From available samples in training, one of them was removed and the model was constructed using other samples and the parameters were evaluated. Then, the model error for the removed sample was calculated and assessed. Since in the LOO-CV method, in each stage, only one of the samples is removed. The number of LOO-CV iterations is equal to the data size in the training set. Then, we only had a total number of 60 samples that we constructed the model using samples 1-59 and predicted the sample 60. In this manner, we constructed the model from samples 1-58 and sample 60 and predicted sample 59.

PLSR creates the regression equation between the independent variables (x, FT-IR spectrum) and the dependent variable (y, analyte concentrations). At the same time, the PCR performs several inverse regressions for the predictor variables versus scores instead of using the raw data [39]. The PCR and PLSR models were applied to the spectra for each of the training sets, including two types of PIFP samples and respective concentration data.

The predictability was evaluated for the models in terms of the high R^2 values and the lowest values for RMSE and RE%. Several factors have also been optimized to prevent overfitting problems in the model. This phenomenon usually occurs when either the model closely fits into a set of data points or the model successfully predicts training samples, while it does not show a good performance in dealing with new samples. The previous studies revealed that the overfitting can be prevented by adjusting the number of factors to less than half of the number of sample sets [40]. Table S1 (Supplementary material) gives several statistical parameters, including bias, RPD, R^2, RE%, RMSEC, RMSECV, RMSEP, and SEP for the normal and transformed data. R^2, RE, and RMSE values were estimated for both validation and calibration models, whereas bias, RPD, and SEP values were only estimated for the validation model.

In the case of protein content, all PCR and PLSR regression models prepared the value of R^2 for the training and validation sets with the normal and transformed data in the range of 0.961 to 0.995 and 0.951 to 0.981, respectively. Whereas, RMSE values extend from 0.142 to 0.602 and 0.203 to 0.596 for the training and validation sets, respectively. The most satisfactory results were related to the normal spectra of PLSR models with the highest R^2 (0.995), lowest RMSEC, and RE values (0.142 and 4.5), respectively. The precision of the predicted model was evaluated with low bias values (-0.02) and low SEP (0.16). The values of R^2, RE%, RMSEP, and RPD were obtained 0.981, 5.9, 0.545, and 13.8, respectively. These values were the most satisfactory results for the validation set, indicating an acceptable accuracy for the prediction model [41]. Similarly, the results of LC-PUFAs (DHA and AA) revealed that the PLSR model assigned the value of R^2 for the training and validation sets (with normal spectra, first derivative, and second derivative) in the range of 0.969 to 0.992 and 0.957 to 0.983, respectively. For the PCR model, the R^2 values for the training and validation sets were 0.967 to 0.989 and 0.951 to 0.981, respectively. Additionally, the RMSE values for the training set for both the PLSR and PCR models were 0.204 to 0.582 and 0.218 to 0.695, respectively, whereas they were 0.187 to 0.438 and 0.199 to 0.451 for the validation set. The PLSR model offers the best results for the training set with normal spectra based on high R^2 (0.992), low RMSEC (0.204), and RE (4.1) values. The precision of the predicted model was evaluated with low bias (-0.03) and low SEP (0.19) values. The results accompanied by R^2 (prediction) value of 0.983, RE% (5.1), RMSEP (0.438), and RPD (13.2). These results indicated an excellent accuracy for the developed model [42]. Hence, the PLSR regression model defined excellent predictions with normal spectra, and it very well fits into the experimental measurements. Figure 5 illustrates the scatter plots of the correlation between actual values (x-axis) and predicted values (y-axis) of LC-PUFAs (AA and DHA), and protein content in PIFP samples using PLSR model for normal spectra. Many points of the plots fall on or very close to the equal concentration line, showing the good prediction capability of the model.

Similarly, according to Figure 6 that shows the scatter plot of actual values versus predicted values by PCR, it can be noted that predicted values resulted in a proper accuracy in comparison to actual values. Although the distance between points with line is higher than in the PLSR method, the model could predict protein and LCPUFA values.

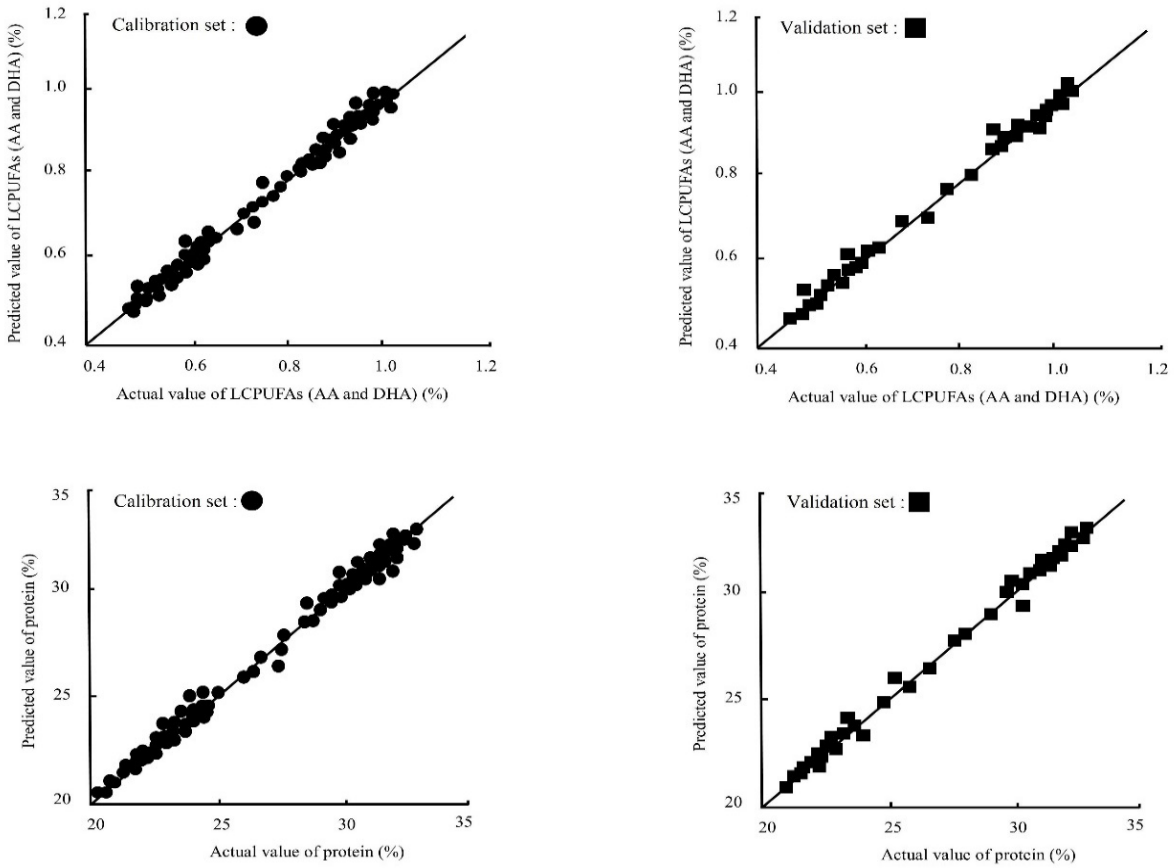

Fig. 5. Plots of predicted values versus actual values of training and validation samples determined by the PLSR model.

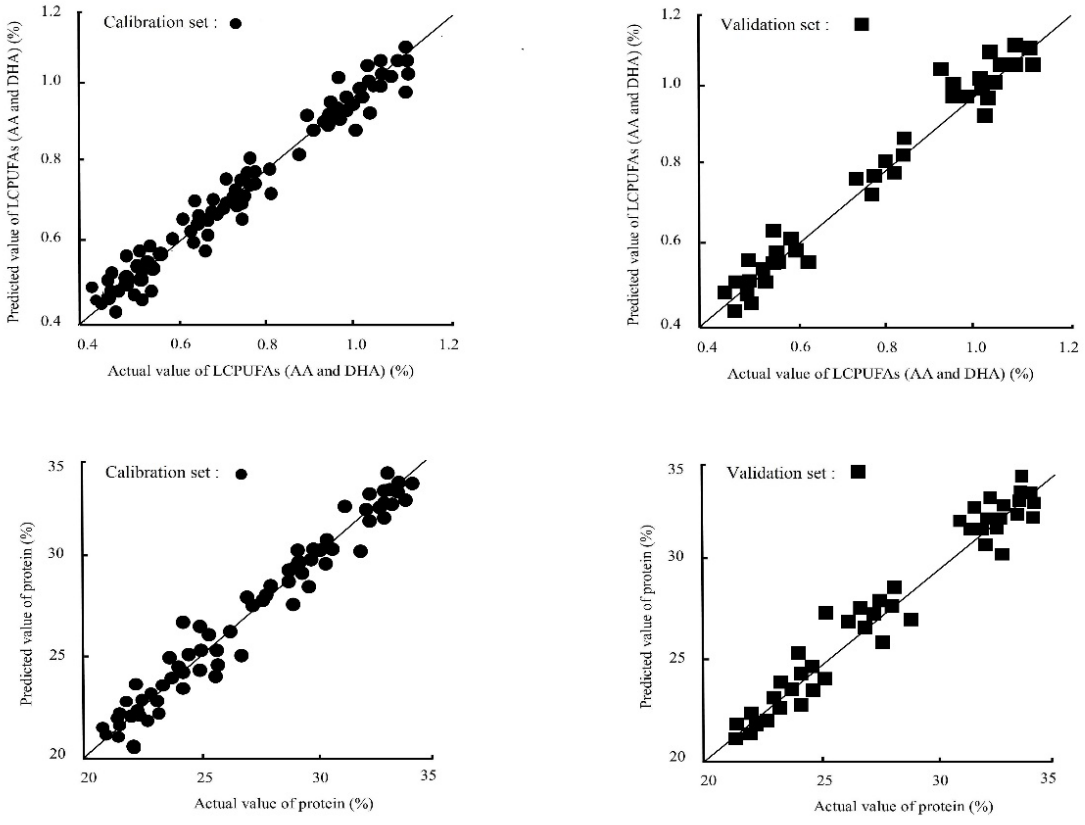

Fig. 6. Plots of predicted values versus actual values of training and validation samples determined by the PCR model.

Conclusions

This study established that ATR-FTIR spectroscopy could be used in conjunction with chemometric methods (PCR and PLSR) to quantify the AA, DHA, and protein content of PIFP samples. A comparative study was conducted using PCR and PLSR models (normal spectra, first and second derivatives). Among the models, the PLSR model predicted the normal spectra the best. R^2 values of 0.992 and 0.983 were calculated for the training and validation sets, respectively. Thus, the PLSR model was used to predict DHA and AA quantitatively. Here, the RMSEP and RMSEC values were 0.438 and 0.204, respectively. The prediction set had a low relative error and a high RPD value (13.2), indicating the developed model's high precision. Similarly, the PLSR model predicted the protein content with an R^2 value of 0.995 for the training and 0.981 for the validation set. The RMSEP and RMSEC values, in this case, were 0.545 and 0.142, respectively. The prediction set displayed a high PRD (13.8) and a low relative error, indicating that the developed model was highly accurate. This study demonstrated that combining ATR-FTIR spectroscopy and chemometrics resulted in a non-destructive, simple, and rapid analytical method for the quantitative analysis of PIFP samples that did not require sample preparation. Therefore, this method is a viable alternative for validating a wide variety of infant formula products.

Acknowledgments

The authors are grateful to the director and staff of central laboratory of university of Sistan & baluchestan for providing all the facilities to carry out this research.

Disclosure Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Lien R. Neurocritical care of premature infants. *Biomedical journal*. **2020** Jun 1;43(3):259-67.
2. Hay WW. Nutritional support strategies for the preterm infant in the neonatal intensive care unit. *Pediatric gastroenterology, hepatology & nutrition*. **2018** Oct 1;21(4):234-47.
3. van Goudoever JB. Nutrition for preterm infants: 75 years of history. Annals of Nutrition and Metabolism. **2018**;72(3):25-31.
4. Azami M, Jasemi S, Khalifpur Y, Badfar G. Causes of mortality in a neonatal intensive care unit in Iran: one year data. *Medical Journal of Indonesia*. **2020** Jun 30;29(2):143-8.
5. Ceratto S, Savino F, Vannelli S, De Sanctis L, Giuliani F. Growth Assessment in Preterm Children from Birth to Preschool Age. *Nutrients*. **2020** Jul;12(7):1941.
6. Crippa BL, Colombo L, Morniroli D, Consonni D, Bettinelli ME, Spreafico I, Vercesi G, Sannino P, Mauri PA, Zanotta L, Canziani A. Do a few weeks matter? late preterm infants and breastfeeding issues. *Nutrients*. **2019** Feb;11(2):312.
7. World Health Organization. Guidelines on optimal feeding of low birth-weight infants in low-and middle-income countries. World Health Organization; **2011**.
8. Ahern GJ, Hennessy AA, Ryan CA, Ross RP, Stanton C. Advances in infant formula science. Annual review of food science and technology. **2019** Mar 25; 10:75-102.
9. Dipasquale V, Serra G, Corsello G, Romano C. Standard and specialized infant formulas in europe: making, marketing, and health outcomes. *Nutrition in Clinical Practice*. **2020** Apr;35(2):273-81.
10. Embleton ND, van den Akker CH. Protein intakes to optimize outcomes for preterm infants. InSeminars in perinatology **2019** Nov 1 (Vol. 43, No. 7, p. 151154). WB Saunders.
11. Fenton TR, Al-Wassia H, Premji SS, Sauve RS. Higher versus lower protein intake in formula-fed low birth weight infants. Cochrane Database of Systematic Reviews. **2020**(6).
12. Koletzko B, Bergmann K, Brenna JT, Calder PC, Campoy C, Clandinin MT, Colombo J, Daly M, Decsi T, Demmelmair H, Domellöf M. Should formula for infants provide arachidonic acid along with DHA? A position paper of the European Academy of Paediatrics and the Child Health Foundation. *The American journal of clinical nutrition*. **2020** Jan 1;111(1):10-6.
13. Poitevin E. Official methods for the determination of minerals and trace elements in infant formula and milk products: a review. *Journal of AOAC International*. **2016** Jan 1;99(1):42-52.
14. Simonne AH, Simonne EH, Eitenmiller RR, Mills HA, Cresman Iii CP. Could the Dumas method replace the Kjeldahl digestion for nitrogen and crude protein determinations in foods? *Journal of the Science of Food and Agriculture*. **1997** Jan;73(1):39-45.
15. Helrich K. Official methods of analysis of the Association of Official Analytical Chemists. Association of official analytical chemists; **1990**.
16. Tian H, Wang J, Zhang Y, Li S, Jiang J, Tao D, Zheng N. Quantitative multiresidue analysis of antibiotics in milk and milk powder by ultra-performance liquid chromatography coupled to tandem quadrupole mass spectrometry. *Journal of chromatography B*. **2016** Oct 15;1033: 172-9.
17. Sokol E, Ulven T, Færgeman NJ, Ejsing CS. Comprehensive and quantitative profiling of lipid species in human milk, cow milk and a phospholipid-

enriched milk formula by GC and MS/MSALL. *European Journal of Lipid Science and Technology*. **2015** Jun;117(6):751-9.

18. Limm W, Karunathilaka SR, Yakes BJ, Mossoba MM. A portable mid-infrared spectrometer and a non-targeted chemometric approach for the rapid screening of economically motivated adulteration of milk powder. *International Dairy Journal.* **2018** Oct 1;85: 177-83.

19. Mabood F, Jabeen F, Ahmed M, Hussain J, Al Mashaykhi SA, Al Rubaiey ZM, Farooq S, Boqué R, Ali L, Hussain Z, Al-Harrasi A. Development of new NIR-spectroscopy method combined with multivariate analysis for detection of adulteration in camel milk with goat milk. *Food Chemistry.* **2017** Apr 15;221: 746-50.

20. Hernández-Martínez M, Gallardo-Velázquez T, Osorio-Revilla G, Almaraz-Abarca N, Ponce-Mendoza A, Vásquez-Murrieta MS. Prediction of total fat, fatty acid composition and nutritional parameters in fish fillets using MID-FTIR spectroscopy and chemometrics. *LWT-Food Science and Technology*. **2013** Jun 1;52(1):12-20.

21. Cebi N, Yilmaz MT, Sagdic O, Yuce H, Yelboga E. Prediction of peroxide value in omega-3 rich microalgae oil by ATR-FTIR spectroscopy combined with chemometrics. *Food chemistry*. **2017** Jun 15;225: 188-96.

22. Ozulku G, Yildirim RM, Toker OS, Karasu S, Durak MZ. Rapid detection of adulteration of cold pressed sesame oil adultered with hazelnut, canola, and sunflower oils using ATR-FTIR spectroscopy combined with chemometric. *Food Control*. **2017** Dec 1;82: 212-6.

23. Condurso C, Cincotta F, Tripodi G, Verzera A. Characterization and ageing monitoring of Marsala dessert wines by a rapid FTIR-ATR method coupled with multivariate analysis. *European Food Research and Technology*. **2018** Jun;244(6):1073-81.

24. Wang Y, Huang HY, Zuo ZT, Wang YZ. Comprehensive quality assessment of Dendrubium officinale using ATR-FTIR spectroscopy combined with random forest and support vector machine regression. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*. **2018** Dec 5;205: 637-48.

25. Bekhit MY, Grung B, Mjøs SA. Determination of omega-3 fatty acids in fish oil supplements using vibrational spectroscopy and chemometric methods. *Applied spectroscopy*. **2014** Oct 1;68(10):1190-200.

26. Visentin G, Penasa M, Gottardo P, Cassandro M, De Marchi M. Predictive ability of mid-infrared spectroscopy for major mineral composition and coagulation traits of bovine milk by using the uninformative variable selection algorithm. *Journal of dairy science.* **2016** Oct 1;99(10):8137-45.

27. Golay PA, Moulin J, Collaborators: Alewijn M Braun U Choo LF Cruijsen H Delmonte P Fontecha J Holroyd S Hostetler G Lacoste F Lehmann C Nagelholt L Phillips S Ritvanen T Rizzo A Shimelis O Srigley C Sullivan D Trossat P. Determination of labeled fatty acids content in milk products, infant formula, and adult/pediatric nutritional formula by capillary gas chromatography: Collaborative study, final action 2012.13. *Journal of AOAC International*. **2016** Jan 1;99(1):210-22.

28. Stordrange L, Libnau FO, Malthe-Sørenssen D, Kvalheim OM. Feasibility study of NIR for surveillance of a pharmaceutical process, including a study of different preprocessing techniques. *Journal of Chemometrics: A Journal of the Chemometrics Society*. **2002** Aug;16(8-10):529-41.

29. Hadjikinova R, Petkova N, Hadjikinov D, Denev P, Hrusavov D. Development and validation of HPLC-RID method for determination of sugars and polyols. *Journal of Pharmaceutical Sciences and Research.* **2017** Aug 1;9(8):1263.

30. Mabood F, Hussain J, Al Nabhani MM, Gilani SA. Detection and quantification of formalin adulteration in cow milk using near infrared spectroscopy combined with multivariate analysis. *J Adv Dairy Res.* **2017**;5(2).

31. Cozzolino D, Power A, Chapman JJ. Interpreting and reporting principal component analysis in food science analysis and beyond. *Food Analytical Methods.* **2019** Nov;12(11):2469-73.

32. Smirnova A, Konoplev G, Mukhin N, Stepanova O, Steinmann U. Milk as a Complex Multiphase Polydisperse System: Approaches for the Quantitative and Qualitative Analysis. *Journal of Composites Science.* **2020** Dec;4(4):151.

33. Jamwal R, Kumari S, Kelly S, Cannavan A, Singh DK. Rapid detection of pure coconut oil adulteration with fried coconut oil using ATR-FTIR spectroscopy coupled with multivariate regression modelling. LWT. **2020** May 1;125: 109250.

34. Carbas B, Machado N, Oppolzer D, Ferreira L, Brites C, Rosa EA, Barros AI. Comparison of near-infrared (NIR) and mid-infrared (MIR) spectroscopy for the determination of nutritional and antinutritional parameters in common beans. *Food chemistry*. **2020** Feb 15;306: 125509.

35. Márquez C, Ruisánchez I, Callao MP. Qualitative and quantitative multivariate strategies for determining paprika adulteration with SUDAN I and II dyes. Microchemical Journal. **2019** Mar 1;145: 686-92.

36. Bassbasi M, Platikanov S, Tauler R, Oussama A. FTIR-ATR determination of solid non-fat (SNF) in raw milk using PLS and SVM chemometric methods. *Food chemistry*. **2014** Mar 1;146: 250-4.

37. de Souza Gondim C, Junqueira RG, de Souza SV, Ruisánchez I, Callao MP. Detection of several common adulterants in raw milk by MID-infrared spectroscopy and one-class and multi-class multivariate strategies. *Food chemistry*. **2017** Sep 1;230: 68-75.

38. Botelho BG, Reis N, Oliveira LS, Sena MM. Development and analytical validation of a screening method for simultaneous detection of five adulterants

in raw milk using mid-infrared spectroscopy and PLS-DA. *Food chemistry*. **2015** Aug 15;181: 31-7.

39. Elmas ŞN, Arslan FN, Akin G, Kenar A, Janssen HG, Yilmaz I. Synchronous fluorescence spectroscopy combined with chemometrics for rapid assessment of cold–pressed grape seed oil adulteration: Qualitative and quantitative study. *Talanta.* **2019** May 1;196: 22-31.

40. Beebe KR. Pell, RJ: Seasholtz, MB. Chemometrics: A Practical Guide. **1998**;1.

41. Pereira CG, Leite AI, Andrade J, Bell MJ, Anjos V. Evaluation of butter oil adulteration with soybean oil by FT-MIR and FT-NIR spectroscopies and multivariate analyses. Lwt. **2019** Jun 1;107: 1-8.

42. Jamwal R, Kumari S, Dhaulaniya AS, Balan B, Singh DK. Application of ATR-FTIR spectroscopy along with regression modelling for the detection of adulteration of virgin coconut oil with paraffin oil. Lwt. **2020** Jan 1;118: 108754.