Indian Journal of Science and Technology Vol. 4 No. 7 (July 2011) ISSN: 0974- 6846

A comparative study of principal component regression and partial least squares regression with application to FTIR diabetes data

P. Venkatesan[1], C. Dharuman[2] and S. Gunasekaran[3]
[1] *Department of Statistics, Tuberculosis Research Centre, ICMR, Chennai-600 031, India*
[2]*P. G. Department of Mathematics;* [3]*P. G. Department of Physics, Pachaiyappa's College, Chennai-600 030, India*
cdharuman55@gmail.com

Abstract

In recent years, Fourier Transform Infrared (FT-IR) spectroscopy has had an increasingly important role in the field of pathology and diagnosis of disease states. The principal component regression (PCR) and the partial least squares regression (PLS) are the often proposed methods and widely used in FTIR data analysis, when the number of explanatory variable is relatively large in comparison to the samples as the least squares estimator may fail in such situations. They provide biased estimators with the relatively smaller variation than the variance of the least squares estimators. In this paper, a FTIR diabetes dataset is used in order to examine the performance of the two biased regression models on prediction. The conclusion is that for prediction PCR and PLS provides similar results which require substantial verification for any claims as to the superiority of any of the two biased regression methods.

Keywords: Fourier Transform Infrared, Principal Component Regression, Partial Least Square, Diabetes Data.

Introduction

Fourier Transform Infrared (FTIR) spectroscopic method has been used extensively to investigate biological samples either *in vivo* or *in vitro* analysis of tissue extracts-or their systemic effects. The common findings in most studies, independent of the nucleus and the experimental parameters used, has been that the intensities of almost all infrared are altered with respect to normal tissue. Many reports about patterns of spectral changes associated with, for given disease types, cell types and disease states, have been reported (Bertacche *et al.*, 2006). With the onset of a disease, it is found that the relative content of bio-molecules changes, thereby producing a pathophysiological changes in their functions. Blood serves as the primary metabolic transport system in the body and so its composition is an excellent indicator with respect to the metabolic condition of the patient. These bio-chemical changes of blood are especially significant in the case of diseases like diabetes mellitus. Using FTIR it has been demonstrated that glucose, cholesterol, albumin, total protein, triglycerides and urea can be assayed with dried serum (Arnold & Small, 1990; Shaw *et al.*, 1998; Low-Ying *et al.*, 2002).

Diabetes is a syndrome of disordered metabolism which causes abnormal blood glucose levels. Generally, there are two different types of diabetes: type I which the body cannot produce sufficient amount of insulin and type II where insulin cannot be properly used. Therefore, multivariate techniques are generally required to separate target analytic information for the complex matrix components. In general, multivariate analysis uses information derived from multiple wave numbers or wavelengths instead of single one. Calibration is based on the relationship between the spectral variances at particular wave numbers or wavelengths and changes in the target analytic concentrations. This relationship can be established by two different approaches: projection analysis and correlation analysis. Several algorithms used in the literature including principal component analysis (PCA), partial least squares (PLS) analysis.

Infrared (IR) radiation is the term used to describe electromagnetic radiation with frequencies and energies somewhat lower than those associated with visible light are emitted as a range of frequencies from a heated object upon a collection of certain molecules, absorption of discrete frequencies by the molecules takes place, corresponding to the absorption of well-defined amounts of energy from the range of energies in the radiation. This is the basis of IR absorption spectroscopy. Two main applications of IR spectroscopy provide important structural information about molecules. The first is the study of simple molecules (di-atomic and tri-atomic) in the gas phase; the exact amounts of energy absorbed from the IR radiation are related to increase in the rotational and vibrational energy of the molecules. It is possible to determine bond lengths and force constants (a measure of the resistance to stretching). The second application of IR involves the recognition of the structures of more complicated molecules from their characteristic absorptions. IR can be used to indicate the nature of the functional groups in a molecule and by comparison with spectra from known compounds, to all identification of an unknown material.

Fourier- Transform Spectroscopy

Fourier-transform infrared (FTIR) spectroscopy (Griffiths & de Haseth, 1986) is based on the idea of the interference of radiation between two beams to yield an interferogram. The latter is a signal produced as a function of the change of path length between the two beams. The two domains of distance and frequency are inter-convertible by the mathematical method of Fourier-transformation. The basic components of an FTIR spectrometer are shown schematically in Fig.1. The radiation emerging from the source is passed through an interferometer to the sample before reaching a detector. Upon amplification of the signal, in which a filter has eliminated high frequency contributions, the data are converted to digital form by an analog-to-digital converter and transferred to the computer for Fourier-transformation.

Indian Journal of Science and Technology Vol. 4 No. 7 (July 2011) ISSN: 0974- 6846

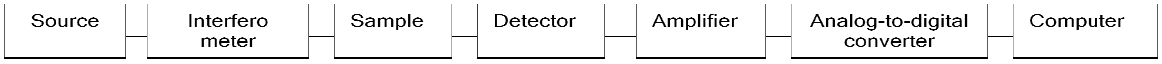
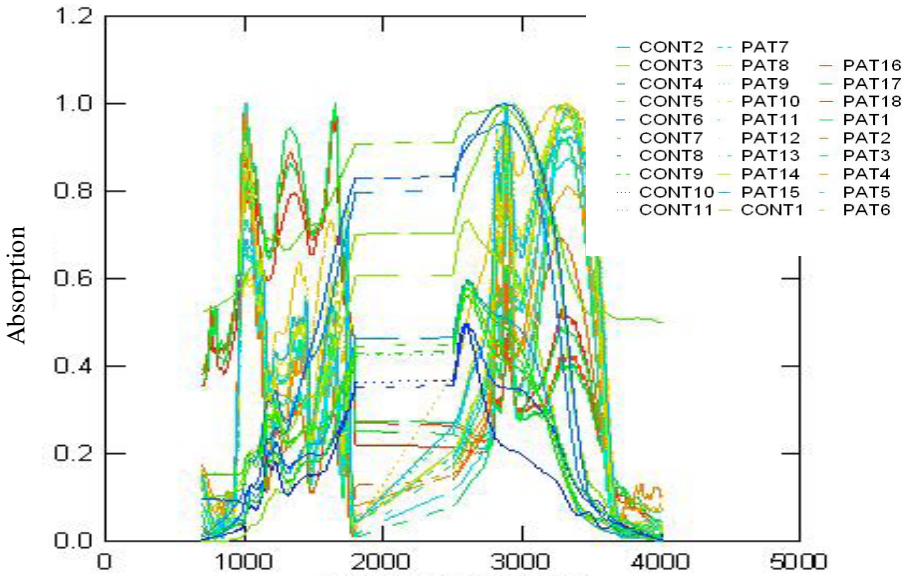


Fig. 1. Basic components of a FTIR spectrometer



Wave number

Fig. 2(a). Overlay plot of the FTIR Spectra of diabetic and non-diabetic individuals (700-4000 cm^{-1}).

Fourier-transform has been commonly used for spectroscopic analyses in the mid-infrared region. This is because of the following advantages with Fourier-transform spectroscopy: (1) the collection efficiency of photon fluxes is high because light from the light source or the sample with a wide area and a wide angle of radiation can be guided into the spectroscope efficiently (2) the detection efficiency of signals is high because all the wavelengths are detected simultaneously and (3) high resolution can be obtained because its wave number precision is high.

Fourier transform spectroscopy is a simple mathematical technique to resolve a complex wave into its frequency components. It is a different approach to radio frequency, microwave and infrared spectroscopy and has proved to be a powerful technique for the study of weakly found dimmer complexes. It is also used for the study of diatomic molecules both stable and transient. The conventional IR spectrometers are not of much use for the far IR region as the sources are weak and the detectors are insensitive. FTIR has made this energy-limited region more accessible. It has made the middle infrared also very useful. The conventional spectroscopy, called the frequency domain spectroscopy records the radiant power as a function of frequency where as in the time domain spectroscopy, the change in radiant power is recorded as a function of time. In a Fourier-Transform spectrometer, a time domain plot is converted into a frequency domain spectrum. Complicated time domain spectra could be transformed into frequency domain spectra, the actual calculation of the Fourier transform of such systems is done by means of high-speed computers (King *et al.*, 1998).

Due to the advent of FTIR techniques, a new awakening has taken place in the biological application of spectroscopy. Two basic aspects make FTIR spectroscopy potentially attractive to the biomedical research. First, the technique is truly molecular level in nature. One is able to probe directly the structure of molecule (Kartheek *et al.*, 2011) or spectrum under study and information is obtained about functional groups, bonding forms, confirmation and environmental, influences that affect the molecular frequencies. Second, IR spectroscopy needs relatively small quantity of the sample to be studied. Substance can be studied in the solid, liquid, or gaseous state. The analysis of FTIR spectroscopy data biological and biomedical fields possesses many methodological problems. Particularly each function group in a biological macromolecule contributes to the IR Spectrum, which is very complex. Overlapping and mixing of vibrational mode make the analysis very difficult. This paper considers a diabetes database to empirically evolve new methods.

FTIR spectral diabetes data

A state of high glucose level in the blood is recognized as diabetes. This state can be produced by many different factors. Basically, in diabetes there is a disturbance of metabolic function of all body cells and tissues. The cause for this metabolic disturbance relates to the deficiency of anabolic protein hormone insulin, which is an internal secretion of the pancreas. Lack of insulin affects the metabolism of carbohydrate, fat and protein and it causes a significant disturbance of water and electrolyte homeostasis. Insulin is an anabolic hormone and has profound effects on the metabolism of carbohydrates, fat and proteins in the body so that for diabetic patients, significant changes in serum composition with regard to the latter classes of substances can be expected. The IR spectrum of serum can provide qualitative and quantitative information on such bimolecular.

Methods and materials

The data consists of 11 normal which are considered as controls and 18 diabetes (NIDDM) FTIR data measured in the wavelength of 400 - 4000 cm^{-1}. The

Indian Journal of Science and Technology Vol. 4 No. 7 (July 2011) ISSN: 0974- 6846

spectral region consists of three region, which corresponds to the glucose region (925-1250 cm^{-1}); protein region (1500-1700 cm^{-1}) and lipids or fat region (2800-3400 cm^{-1}). Considerable spectral differences observed between the normal and diseased serum were consider for application of pattern recognition. Diagnostic classification for the discrimination of diabetic from normal serum based on pattern classifications methods using multivariate techniques are considered in this paper. The overlay plot of the normal's and NIDDM's are presented in Fig. 2(a) and 2(b) separately for the whole region (700-4000) and diabetes region (925-1250).

Factor based regression methods

The other commonly used factor based methods are principal component analysis (PCA) and partial least square (PLS) analysis. The advantage of full spectrum by least square method is a useful technique followed by PCA and PLS which are factor based data analysis techniques.

Principal Component Analysis (PCA)

In real samples, there are usually many different sources of variation that make up the spectrum, the constituents in the sample matrix, inter constituents interactions, instrumental variation such as detector noise, change of environments during sample collection that affect the baseline and absorbance, and differences in sample handing. All of these variations are presented in the collected spectral data at each wave length. The method used in the PCA statistical technique is that at characterized variation in the spectral data and these are then used to construct the original spectrum by multiplying each one by a different constant scaling factor and adding the results factor. Those variations are called principal components, eigenvectors, spectral loadings or loading vectors and they are orthogonal to each other. The scaling constants used to reconstruct the spectra are known as scores and they are unique to each separate principal component. The First Principal component accounts for the much of the variability in the data as possible, and each succeeding principal component accounts for as much of the remaining variability as possible. We obtain a reconstructed spectral data by PCA. It is done by using the goal of PCA to reduce the dimensionality of the spectral data and finally use mean square sense. Therefore a spectral data matrix can be decomposed as shown in eqn (2).

$$A=SP+E \quad (1)$$

where, A is an m ×n matrix of spectral data , S is an m × ℓ matrix of score values for all of the spectra, and P is an ℓ × n matrix of principal components. In PCA decomposition, the rows of P are eigenvectors of A^TA, and the columns of S are proportional in to the eigenvectors of AA^T. The E matrix contains the spectral residuals not fit by the optimal PCA and has the same dimension as A, m is the number of samples (spectra), n is the number of data points (wavelengths) and ℓ is the number is principal components used reconstruct the original spectral data, which is much less than m . There are two important benefits by the reduction of the matrix dimension, firstly there is a reduction of the random noise in the spectra and secondly, the amount of computational work for subsequent process is greatly reduced. The spectral data decomposition into product of principal components and scores occur only in PCA algorithm. A potential problem with this approach is that the principal components which best represent the variation in the spectral data might not be optimal for predicting concentration for the selected chemical constituent when more predictive information is placed in the initial factors than there is a necessity to derive principal components

Principal Component Regression (PCR)

The two step multivariate pattern recognition method of PCR is commonly used in the first step, Principal Component Analysis (PCA) of the data matrix X is performed. The measured variables (e.g., absorbance at different wavelengths) are converted into new ones (scores on latent variables). This is followed by a multiple linear regression step (MLR), between the scores obtained in the PCA step and the characteristic y to be modeled. PCA creates new orthogonal variables (latent variables) that are linear combinations of the original x-variables. This can be achieved by the method known as singular value decomposition of X :

$$X = U D P' = T P' \quad (2)$$

where X, U, D and P are matrices of order n×p, n×p, p×p, p×p, respectively. U is the weighted normalized score matrix and T is the weighted normalized matrix of order n×p and will represent the new co-ordinates for the n-objects in the new system and P is the loading matrix and the column vectors of P are called eigenvectors or loading-PCs. The elements of P are the loadings (weights) of the original variables on each eigenvector. High loadings for certain original variables on a particular eigenvector mean that these variables are important in the construction of the new variable or score on that principal component (PC). D is a diagonal matrix which means that all off-diagonal elements are equal to zero. These elements, the singular values λ_i are the squared roots of the eigen values of what in this context is often called the covariance matrix (X'X). λ_1 is associated with the score on the first principal component, PC_1, for each object and is related to the amount of variance explained by PC_1. By definition $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_\rho$ so that the principal components can be said to describe D. Although in majority of applications we assume that n > p, in spectroscopy however there are fewer samples than wavelengths measured so that n < p and eqn.(2) has to be rewritten as:

$$X = U D P' = T P' \quad (3)$$

Two main advantages arise from this decomposition. The first being the orthogonal (U'U=I) whose matrix inverse is not a problem, as it was in correlation which has been the usual in spectroscopy. Moreover, we assume that the first new variables or PCs, accounting for the majority of the variance of the original data, contain

Indian Journal of Science and Technology Vol. 4 No. 7 (July 2011) ISSN: 0974- 6846

meaningful information, while the last ones, which account for a small amount of variance, only contain noise and can be deleted. Therefore only r PCs are retained and r < min (n, p). This simplifies the data examination. After performing PCA on X, the second step in PCR consists of the linear regression of the scores and the y property of interest. The linear model between Y and T is of the form:

$$Y_{nxl} = T_{nxr} b_{rxl} + e_{nxl} \quad (4)$$

with the solution: $$b_{rxl} = (T'_{rxn} T_{nxr})^{-1} T'_{rxn} Y_{nxl} \quad (5)$$

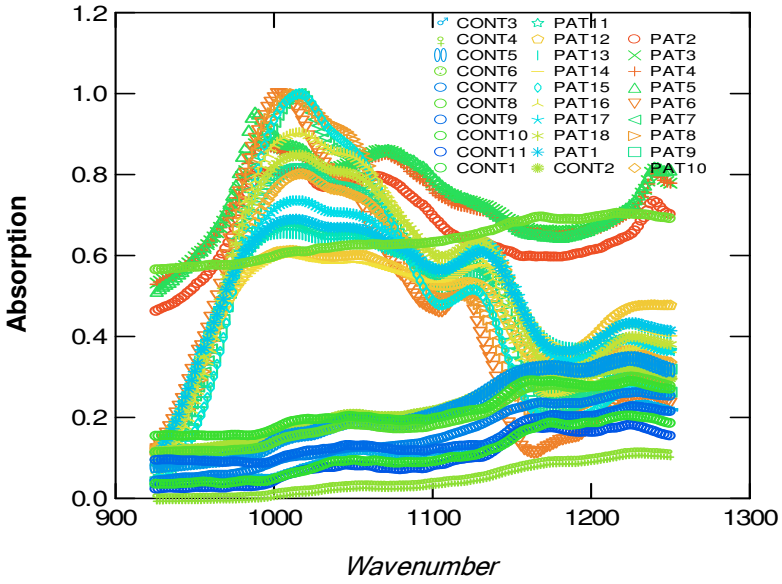


Fig. 2(b): Overlay plot of the FTIR Spectra of diabetic and non-diabetic individuals (925-1250 cm^{-1})

Partial Least Squares Regression (PLS)

The PLS which is included indirect calibration modeling approach helps us to do multivariate calibration based on the least squares criterion (Trygg, 2004). Respect to multiple linear regression (MLR), of which it is the generalization and which has been traditionally used for the modeling of matrix Y by means of X, PLS possesses the distinct advantage of being more adaptable to modern measuring instrumentation, such as FTIR spectroscopy, which provide a large number of strongly correlated X-variables, also called predictors (Wold *et al.*, 2001). In PLS projection method the scores are linear combinations of the original variables X_k, and hence, these scores have the characteristic of weighed averages, being nicely normally distributed and precise. Consequently, by selecting and combining the variables to few groups called scores, PLS may be useful to analysts to better interpret the large number of variables associated to data. In PLS method the relationship between the predictor's variance and the dependent variables is represented by principal components (PCs) that follow a numeric sequence, depending on how strong this relationship is. Predictor variables are considered significant when they take part in the creation of a PC and, consequently, all PCs are modeled on the influence of each variable. The set of a number of PCs sufficient to give an exhaustive description of the Y-matrix is called model. If the model includes all the samples it is termed a calibration model. One of the advantages in using PLS method is that PCs are modeled not only on the predictors set, but also on the responses, so that it is possible to maximize the variance of both X and Y coordinates of the model. PLS is different from other multivariate calibrations, such as principal component regression (PCR), because the utilization of the responses data set is accomplished in an active way during the statistical calculations. In this way, the information contained in X and Y coordinates are well balanced, and the effect of heavy but irrelevant variations in the predictors set is reduced.

The PLS algorithm finds principal components from spectral data that are also relevant to analytic concentration (Geladi & Kowlaski, 1986). Specifically, PLS regression searches for a set of factors (latent variables) that performs a simultaneous decomposition of spectral data and analytic concentration with the constraint that these factors explain as much as possible the covariance between spectral information and analyte concentrations. Simultaneously the latent variables in PLS are developed along with calibration model, which will ensure maximum correlation with the one information provided by analyze concentration as each latent variable is a linear combination of the spectral information rotated. Hence interference form the chemical matrix is usually less of problem in PLS compared to PCA. The power of PLS is both the response and measurement variables are used iteratively to determine the latent variables for the analysis (Beebe & Kowlaski, 1987; Haaland & Thomas, 1988a,b).

The PLS regression algorithm has become one of the dominant practices of multivariate calibration because of the quality of the calibration model. PLS has been used in various, disciplines such as chemistry (Marquardt *et al.*, 1993), economics, psychology (Nestor *et al.*, 2002) and pharmaceutical science (Zhou *et al.*, 1998) where predictive linear modeling especially with a large number of target analyte, is necessary (Marquardt *et al.*, 1993; Zhou *et al.*, 1998; Nestor *et al.*, 2002). The PLS regression algorithm has become standard toll for modeling linear relations for multivariate measurements.

Results

The FTIR spectra were pre-processed using the following three stages: (a) Normalization of each FTIR spectrum, (b) identifying characteristic spectral features of each experiment group and control group and (c) development of classification algorithm using PCR and PLS extracted features. Each spectrum was normalized by peak absorption intensity of the spectrum. Normalization removes absolute intensity information.

Indian Journal of Science and Technology Vol. 4 No. 7 (July 2011) ISSN: 0974- 6846

Table 1. Mean (± SE) of the normalized FTIR (whole spectrum) spectral values for patients and controls

Patient				
No	Mean	SE	Mini	Maxi
1	0.463	0.007	0.065	1.000
2	0.293	0.006	0.008	1.000
3	0.324	0.008	0.004	1.000
4	0.397	0.008	0.043	1.000
5	0.408	0.009	0.001	1.000
6	0.384	0.007	0.057	1.000
7	0.427	0.008	0.014	1.000
8	0.531	0.009	0.020	1.000
9	0.466	0.009	0.014	1.000
10	0.471	0.009	0.016	1.000
11	0.039	0.008	0.011	1.000
12	0.0438	0.009	0.020	1.000
13	0.143	0.008	0.031	1.000
14	0.443	0.008	0.027	1.000
15	0.322	0.005	0.024	1.000
16	0.425	0.007	0.046	1.000
17	0.433	0.007	0.046	1.000
18	0.437	0.007	0.037	1.000
Control				
1	0.234	0.004	0.000	1.000
2	0.135	0.003	0.000	0.704
3	0.236	0.005	0.000	1.000
4	0.549	0.005	0.001	1.000
5	0.340	0.006	0.000	0.994
6	0.297	0.005	0.000	1.000
7	0.207	0.003	0.000	1.000
8	0.206	0.003	0.000	1.000
9	0.050	0.003	0.000	1.000
10	0.119	0.003	0.000	1.000
11	0.207	0.005	0.013	0.592

Table 2. Mean (±SE) of the normalized FTIR (glucose region) spectral values for patients and controls

Patient				
No	Arithmetic Mean	SE	Minimum	Maximum
1	0.675	0.008	0.461	1.000
2	0.745	0.007	0.532	1.000
3	0.736	0.007	0.529	1.000
4	0.750	0.008	0.512	1.000
5	0.471	0.022	0.111	1.000
6	0.521	0.020	0.082	1.000
7	0.511	0.022	0.119	1.000
8	0.480	0.018	0.041	1.000
9	0.525	0.015	0.124	1.000
10	0.499	0.012	0.091	1.000
11	0.478	0.010	0.103	1.000
12	0.512	0.011	0.113	1.000
13	0.480	0.010	0.095	1.000
14	0.487	0.023	0.047	1.000
15	0.535	0.019	0.068	1.000
16	0.519	0.013	0.111	1.000
17	0.549	0.017	0.096	1.000
18	0.515	0.012	0.114	1.000
Control				
1	0.224	0.006	0.000	0.325
2	0.121	0.005	0.000	0.229
3	0.051	0.003	0.000	0.118
4	0.212	0.007	0.000	0.344
5	0.633	0.005	0.000	0.703
6	0.151	0.006	0.000	0.260
7	0.196	0.005	0.000	0.299
8	0.141	0.004	0.000	0.226
9	0.215	0.004	0.000	0.290
10	0.095	0.004	0.000	0.180
11	0.107	0.005	0.000	0.203

Ratio variable quantify the results and estimate the spectral diagnostic potentiality. The ratio variables were introduced which represents characteristic spectral features of diabetic and non-diabetic groups. The overlay plots of diabetic control spectra for full range and glucose range are given in Fig.2(a) and Fig.2 (b). The figures show considerable variations among patients of diabetes and control. Mean and standard deviation were calculated for each group. A two tailed t-test was used to determine the level of significance. Table1 and Table2 give the descriptive statistics of the diabetic patients and controls for the full spectrum as well as for the glucose region. From the tables it can be seen that there are considerable variations between the patients and controls. Wave number wise comparisons were also made for the glucose region (1250 - 925 cm^{-1}) between the cases and controls and the selected wavenumber comparison are presented in Table 3. It is found that there is a significant difference in the mean FTIR values between diabetic and controls in the region (1250-1206 cm^{-1}) and (1207-1163 cm^{-1}) where as there is no significant difference in the region (1164-938 cm^{-1}). The diabetic patients have higher mean values compared to control in glucose region. From Table 4 it can be seen that the second component distinguishes the significant and non-significant spectral regions giving three patterns. The components also give three patterns. From Table 5, it is seen that the PLS explains more than 92.7 percent variation of the prediction by first three factors of the predictors and more than 99 percent variation of the responses using seven factors. The PCA based factor analysis is applied for the spectral data and it is found that the first four components explain more than 99.59 percent of variance as shown in the Table 6. PCR and PLS are applied to the diabetic data and the results are presented in Table 4 and Table 5. The Hierarchical clustering was carried out and the dendrogram is given in Fig.3. From the dendrogram it is seen that all the controls except one are clustered into one cluster and the diabetes cases are clustered into other cluster. Over all the two groups controls and diabetes had clustered into two except one control (No.23) is clustered along with diabetes.

Discussion and conclusions

The PCR and PLS are the two most commonly used techniques in pattern recognition. The PCR is a two step procedure; the first step, PCA of the data matrix is performed the measured variables absorbance in different wavelength are commuted in the latent variable. This is followed by a MLR between the scores in the PCA step and the characteristic of y to be modeled. The PCA creates new orthogonal variables that are linear combination of the wavelengths which are highly correlated. The PLS is a generalization of MLR and PLS possesses the distinct advantage of being more acceptable to modern measuring instruments such as FTIR Spectroscopy which provides a large number of strongly correlated X-variables. In PLS projection the scores are linear combination of X's and weighted averages. In PLS the relationship between predictor variables and dependant variables is represented by

Indian Journal of Science and Technology Vol. 4 No. 7 (July 2011) ISSN: 0974- 6846

Table 3 Mean (SD) of FTIR Values for select wave numbers for the glucose region

Wave number	Controls (11) Mean± SD	Diabetic (18) Mean± SD	p value
1250	0.270 ± 0.152	0.442 ± 0.190	0.013
1240	0.279 ± 0.152	0.454 ± 0.196	0.018
1230	0.287 ± 0.152	0.445 ± 0.164	0.016
1220	0.285 ± 0.154	0.435 ± 0.151	0.016
1210	0.277 ± 0.155	0.412 ± 0.154	0.031
1200	0.269 ± 0.155	0.386 ± 0.160	0.064
1190	0.267 ± 0.155	0.372 ± 0.162	0.096
1180	0.270 ± 0.154	0.368 ± 0.163	0.119
1170	0.269 ± 0.157	0.376 ± 0.165	0.097
1160	0.263 ± 0.157	0.413 ± 0.158	0.019
1150	0.246 ± 0.160	0.481 ± 0.138	0.001
1140	0.224 ± 0.161	0.557 ± 0.108	0.001
1130	0.207 ± 0.162	0.597 ± 0.083	0.001
1120	0.198 ± 0.160	0.599 ± 0.073	0.001
1110	0.190 ± 0.159	0.591 ± 0.089	0.001
1100	0.182 ± 0.159	0.591 ± 0.100	0.001
1090	0.177 ± 0.159	0.627 ± 0.099	0.001
1080	0.174 ± 0.159	0.667 ± 0.096	0.001
1070	0.175 ± 0.158	0.705 ± 0.093	0.001
1060	0.179 ± 0.158	0.737 ± 0.090	0.001
1050	0.180 ± 0.157	0.759 ± 0.095	0.001
1040	0.175 ± 0.157	0.762 ± 0.100	0.001
1030	0.167 ± 0.158	0.775 ± 0.110	0.001
1020	0.159 ± 0.158	0.806 ± 0.128	0.001
1010	0.152 ± 0.160	0.815 ± 0.132	0.001
1000	0.39 ± 0.160	0.800 ± 0.125	0.001
990	0.128 ± 0.159	0.760 ± 0.117	0.001
980	0.124 ± 0.158	0.670 ± 0.091	0.001
970	0.121 ± 0.158	0.550 ± 0.081	0.001
960	0.120 ± 0.158	0.437 ± 0.109	0.001
950	0.119 ± 0.156	0.342 ± 0.137	0.001
940	0.117 ± 0.156	0.261 ± 0.160	0.025
930	0.115 ± 0.156	0.202 ± 0.177	0.194
925	0.114 ± 0.156	0.180 ± 0.181	0.329

Table 4. Component loading for selected wave numbers

Wave number	Principal components		
	PC_1	PC_2	PC_3
1250	0.878	-0.452	0.071
1226	0.891	-0.444	0.004
1224	0.891	-0.444	-0.002
1210	0.871	-0.487	-0.011
1144	0.956	-0.049	-0.281
1142	0.960	0.000	-0.272
1140	0.962	0.044	-0.264
1162	0.914	0.399	-0.002*
1160	0.914	0.399	0.003
1058	0.908	0.414	0.009
960	0.986	0.017	0.058
958	0.985	-0.031*	0.069
956	0.981	-0.078	0.086
926	0.737	-0.575	0.329

Table 5. Percent variation explained by factors for predictors and responses

Factors	Variation explained for predictor(s)		Variation explained for response(s)	
	%	Cum. %	%	Cum. %
1	79.744	79.744	70.980	70.980
2	17.988	97.732	19.165	90.145
3	1.448	99.180	2.580	92.724
4	0.358	99.538	3.091	95.815
5	0.071	99.609	2.112	97.927
6	0.295	99.905	0.333	98.260
7	0.062	99.966	0.760	99.021
8	0.016	99.982	0.283	99.304
9	0.005	99.986	0.337	99.641
10	0.005	99.991	0.068	99.709
11	0.002	99.993	0.077	99.786
12	0.002	99.995	0.046	99.879
13	0.002	99.997	0.047	99.908
14	0.001	99.998	0.029	99.944
15	0.000	99.998	0.036	99.958
16	0.001	99.999	0.014	99.971
17	0.000	99.999	0.013	99.975
18	0.001	100.000	0.004	99.979
19	0.000	100.000	0.003	99.983
20	0.000	100.000	0.004	99.983
21	0.000	100.000	0.001	99.984
22	0.000	100.000	0.001	99.985
23	0.000	100.000	0.001	99.986
24	0.000	100.000	0.000	99.986
25	0.000	100.000	0.000	99.987
26	0.000	100.000	0.001	99.989
27	0.000	100.000	0.002	

Table 6. FTIR diabetic data- Factor analysis

Latent Roots	Variance Value	Total Variance (%)	Cumulative (%)
1	133.07	81.14	81.14
2	27.25	16.62	97.76
3	2.44	1.49	99.25
4	0.56	0.34	99.59

Indian Journal of Science and Technology Vol. 4 No. 7 (July 2011) ISSN: 0974- 6846

PCs that follow a numeric sequence, depending on how strong this relationship is. One of the major advantages of PLS is that PCs are modeled not only on the predictors, but also on the responses, so that it is possible to minimize the variance of both X and Y co-ordinates of the model. PLS differs from other multivariate calibration models such as PCR, because the utilization of the responses data set is accomplished in an active way during the statistical calculations. There are also modification and extensions of partial least square the SIMPLS algorithm (de Jong, 1993). de Jong and Kiers (1992) described a technique called principal covariates regression. On any case PLS has become an established tool in Chemo-metrics modeling. PLS is still evolving as a statistical modeling tool. Further studies are needed to standardize PCR and PLS for the FTIR spectral data analysis. It is concluded that for prediction, PCR and PLS provides similar results which require substantial verification that may claim superior to any of the two biased regression methods.

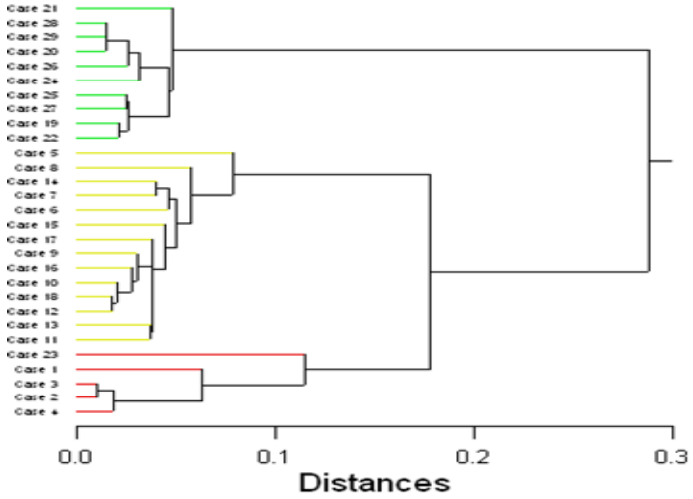


Fig. 3. Dendrogram of diabetic data

References

1. Arnold MA and Small GW (1990) Determination of physiological levels of glucose in an aqueous matrix with digitally filtered Fourier Transform Near-Infrared Spectra. *Anal. Chem.* 62, 1457-1464.
2. Beebe KR and Kowalski BR (1987) Comparison of multivariate calibration and analysis. *Anal. Chem.* 59. 1007A-1017A.
3. Bertacche V, Pini E, Stradi R and Stratta F (2006) Quantitative determination of amorphous cyclosporine in crystalline cyclosporine samples by fourier transform infrared spectroscopy. *J. Pharma. Sci.* 95, 158-166.
4. De Jong S (1993) SIMPLS An alternate approach to PLS regression. *Chemometrics & Intelligent Lab. Systems.* 18: 251-263.
5. De Jong S and Kiers H (1992) Principal convent regression chemometrics and intelligent laboratory systems. 14, 155-164.
6. Geladi P and Kowlakski BR (1986) Partial least squares methods regression: A *Tutorial Anal. Chemica Acta*. 185, 1-17.
7. Griffiths PR and de Haseth JA (1986) Fourier transform infrared spectrometry. Wiley, NY.
8. Haaland DM and Thomas EV (1988a) Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem.* 60, 1193-1202.
9. Haaland DM and Thomas EV (1988b) Partial least-squares methods for spectral analyses. 2. Application to simulated and glass spectral data. *Anal. Chem.* 60, 1202-1208.
10. Kartheek M, Anton Smith A, Kottai Muthu A and Manavalan R (2011) Determination of adulterants in food: A review. *J. Chem. & Pharma. Res*. 3(2), 629-636.
11. King H, Aubert RE and Herman WH (1998) Global burden of diabetes,1995-2025. Prevalence, numerical estimates and projections. *Diabetes Care.* 21, 1413-1414.
12. Low-Ying S, Shaw R A, Leroux M and Mantsch HH (2002) Quantification of glucose and urea in the whole blood by mid-infrared spectroscope of dry films. *Vibr. Spectr.* 28,111-116.
13. Marquardt LA, Arnold MA and Small GW (1993) Near-infrared spectroscopic measurement of glucose in a P-protein matrix. *Anal. Chem.* 62, 3271-3278.
14. Nestor PG, O'Donnell BF, McCarley RW, Niznikiewiez M, Barnard J, Jen S Z and Bookstein FL (2002) A new statistical method for testing hypotheses of neuropsychological/ MRI relationships in schizophrenia; Partial least squares analysis. *Schizophrenia Res*. 53, 57-66.
15. Shaw RA, Kotowich S, Leroux M and Mantsch HH (1998) Multianalyte serum analysis using mid-infra spectroscopy. *Ann. Clinical Biochem.* 35, 624-632.
16. Trygg J (2004) Prediction and spectral profile estimation in multivariate calibration. *J. Chemometrics*. 18,166-172.
17. Wold S, Sjo¨stro¨m M and Eriksson L (2001) PLS-regression: A basic tool of chemometrics. *Chemometrics & Intelligent lab. Sys*. 58, 109-130.
18. Zhou X, Hines P and Borer (1998) Moisture determination in hygrosckoic drug substances by near infrared spectroscopy. *J. Pharma & Biomed. Anal.* 17(2), 219-225.