

# STATISTICAL RESEARCH

## PATTERN RECOGNITION METHODS FOR HIGH DIMENSIONAL DATA:

### An application to FTIR spectral data

#### Background

Pattern recognition is a term which encompasses a wide range of techniques for classifying data. Given a collection of objects characterized by a set of measurements made on each object, the goal is to find and predict a property of the objects that is not directly measurable itself. Pattern recognition has a number of advantages for analyzing spectroscopic data. It provides an unbiased method of analysis, useful for both research and clinical applications. It provides ways of identifying features which differ between different classes of data together with methods for classifying unknown objects. Pattern recognition analysis is useful for both image processing and spectroscopy, but plays a different role in the two types of data. For image processing the main emphasis is on producing reproducible techniques which will assist the human analyst in interpretation of the image, whether it is to identify and typify structures or to quantify some property. With spectroscopy, the emphasis is more on discriminating between spectra from different classes of samples, and reducing the large numbers of spectral features in order to make the available information more accessible.

Fourier Transform Infrared (FTIR) Spectroscopic method has been used extensively to investigate biological samples—either by *in vivo* or *in vitro* analysis of tissue extracts—and their systemic effects. The common finding 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. With the onset of a disease, it is found that the relative content of bio-molecules changes, thereby producing a patho-physiological change in their functions. As blood serves as the primary metabolic transport system in the body, its composition is an excellent indicator with respect to the metabolic condition of the patient.

These biochemical changes of blood are particularly significant in the case of diseases such as diabetes mellitus. Using FTIR it has been demonstrated that glucose, cholesterol, albumin, total protein, triglycerides and urea can be assayed with dried serum.

Fourier – transform has been commonly used for spectroscopic analyses in the mid – infrared region due to following advantages: (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 wave-lengths are detected simultaneously and (3) high resolution can be obtained because its wave number precision is high.

### **Aim**

- To compare the multivariate based pattern classification methods for the discrimination of diabetic from normal serum using FTIR spectroscopic data

### **Methods**

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 spectrum, and the actual calculation of the Fourier transform of such systems is done by means of high-speed computers. The other commonly used methods are principal component analysis (PCA) and partial least square (PLS) analysis. In real samples, there are usually different sources of variation that make up the spectrum, such as the constituents in the sample matrix, inter constituents' interactions, instrumental variation such as detector noise, changing, of environments during sample collection that effect the baseline and absorbance, and differences in sample handling. 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 variations in the spectral data are determined, and these are used to construct the original spectrum by multiplying each one by a different constant scaling factor and adding the results factor. These 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. Reconstructed spectra data by PCA is obtained by using the goal of PCA to reduce the dimensionality of the spectra data and finally mean square sense is used to compare with original spectra data. The two step multivariate pattern recognition method of principal component regression is commonly used: in the first step, a 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 and MLR. PCA creates new orthogonal variables (latent variables) that are linear combinations of the original  $x$ -variables.

The PLS which includes indirect calibration modeling approach helps us to do multivariate calibration based on the least squares criterion. With respect to MLR, it 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. 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 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 with the data. In PLS method the relationship between the predictors' variance and the dependent variables is represented by principal components that follow a numeric sequence, depending upon the strength of the relationship. Predictors' variables are considered significant when they take part in the creation of a principal component and,

consequently, all principal components are modeled based on the influence of each variable. A set of a number of principal components 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 principal components 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, because the utilization of the responses data set is accomplished in an active way during the statistical calculations. By 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.

### **Data**

A state of high glucose level in the blood is recognized as diabetes. This state can be produced by 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 an anabolic protein hormone called “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. The IR spectrum of serum can provide qualitative and quantitative information on such biomolecules. The data consisted of 11 normal and 18 diabetes (non insulin dependent diabetes mellitus), and FTIR data was measured in the wave length of 400 – 4,000. The spectral region consists of three regions, which corresponds to the glucose region (925-1250  $\text{cm}^{-1}$ ); protein region (1500-1700 $\text{cm}^{-1}$ ) and lipids or fat region (2800-3400  $\text{cm}^{-1}$ ). Considerable spectral differences observed between the normal and diseased serum were considered for application of pattern recognition.

### **Results**

There were considerable variations between the patients and controls. Wave lengthwise comparison was made for the glucose region (1250 – 925  $\text{cm}^{-1}$ )

between the cases and controls. It was found that there was a significant difference in the FTIR values between diabetic and controls in the regions 1250-1206  $\text{cm}^{-1}$  and 1164-938  $\text{cm}^{-1}$ , whereas there was no significant difference in the other regions. The diabetic patients had consistently higher mean values compared to normals throughout the glucose region.

The principal component regression and PLS are the two most commonly used techniques in pattern recognition in high dimension databases. The principal component regression is a two step procedure; in the first step, PCA of the data matrix is performed, and the absorbance of the variables is measured at different wavelength and is computed 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. One of the major advantage of PLS is that principal components 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 is different from other multivariate calibration models such as principal component regression, because the utilization of the responses data set is accomplished in an active way during the statistical calculations.

Total wavelengths in the glucose region considered were 325. The PCA and PLS were applied to the diabetic data. The variations explained by the first few components are given in table 26. It was observed that about 99% of the variation was captured by the first 3 components in the predictors.

**Table 26:** PLS Model-variations explained by the components

PC	Variation explained for predictor(s)		Variation explained for response(s)	
	%	Cumulative %	%	Cumulative %
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

The discriminant analysis was carried out using the first three components, and the Jackknifing classification resulted in 97% correct classification as shown in table 27. Further studies are under progress for comparing these approaches with Machine learning approaches such as artificial neural networks, support vector machines and genetic algorithms. The study is in progress.

**Table 27:** Jackknifed Classification matrix

Observed	Predicted		
	Normal	Diabetic	% correct
Normal	10	1	91%
Diabetic	0	18	100%
Total	10	19	97%

[Contact person: Dr.P. Venkatesan (E-Mail ID: venkatesanp@trcchennai.in)]