Partial Least Square (PLS) Analysis*

Most Favorite Tool in Chemometrics to Build a Calibration Model

Keshav Kumar

Partial least square (PLS) analysis is the most favourite tool in chemometrics to develop calibration models. PLS technique allows us to decipher even the complex systems by analysing all the variables instead of looking at them one at a time. PLS technique not only capture the maximum variation associated with predictor (i.e. spectra) and predicted (i.e. concentration) variables but also maximises the correlation between them. The present article describes the working scheme of the PLS algorithm. It also describes important technical details that need to be considered for developing a parsimonious and robust calibration model.

Keshav Kumar is working as a postdoc scientist at Geisenheim University of Applied Sciences, Germany. His research mainly focuses on chemometrics and its application in various fields. He obtained his PhD from the Department of Chemistry, Indian Institute of Technology, Madras, India, under the guidance of Professor A K Mishra.

1. Introduction

Chemometrics is a relatively new branch of science that has been defined as “the chemical discipline that uses mathematical, statistical, and other methods employing formal logic (a) to design or select optimal measurement procedures and experiments and (b) to provide maximum relevant chemical information by analyzing chemical data” [1–4]. The Swedish organic chemist Svante Wold coined the term chemometrics in 1971 [1–4]. International chemometrics society was formed in the 1970s, and the first international chemometrics meeting was held in Cosenza (Italy) in 1983 [1–4]. Chemometrics society sponsored two international journals—*Chemometrics and Intelligent Laboratory System* and *Journal of Chemometrics*—in 1986 and 1987, respectively [1–4]. These two journals publish theoretical developments as well as novel applications of chemometrics. Chemometrics is a highly

Keywords
Analytical chemistry, partial least square analysis, PLS, chemometrics, calibration, cross-validation.

*Vol.26, No.3, DOI: https://doi.org/10.1007/s12045-021-1140-1

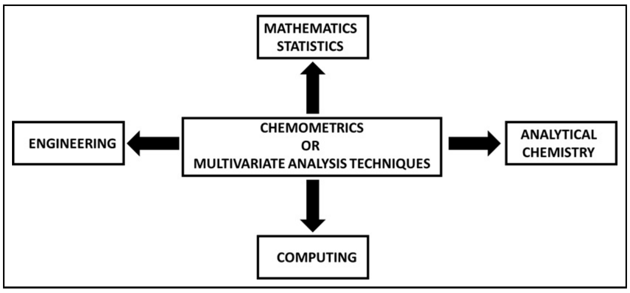

Figure 1. Relation of chemometrics with different disciplines.

interdisciplinary science with applications in various fields. Its relation with various disciplines can be summarized using *Figure* 1.

Basic methods of chemometrics were originally developed for the analyses of complex and correlated data sets obtained from the fields such as economics and psychology to predict the trends of economics using certain indicators or to predict the behaviour or to measure intelligence [1–3]. Chemometrics has been considered a part of analytical chemistry since the beginning of the 1970s. It is generally used for (i) process control analysis, (ii) food analysis, (iii) forensic science, (iv) metabolomics, (v) clinical diagnosis, (vi) environmental monitoring, (vii) reaction monitoring, and (viii) synthesis optimization [1–4]. One of the major reasons for the increase in popularity of chemometric techniques among analytical chemists was the introduction of personal computers [1–4]. The combination of chemometrics and computers enabled analysts to analyse the large volume of data sets obtained from various instruments (single or hyphenated) efficiently. This exercise otherwise is cumbersome and time-consuming.

The chemometrics techniques involve simultaneous analysis of more than one variable at a time while considering the correlation among the data set variables.

The chemometrics techniques involve simultaneous analysis of more than one variable at a time while considering the correlation among the data set variables [1–5]. Chemometric techniques make an approximation of the given data set by finding a set of orthogonal factors that can describe the maximum variance of the data, thereby reducing the collinearity and simplifying the data set for further analysis [1–5]. An analytical chemist often uses chemometric techniques to develop a calibration model that could further be applied for analyzing the analytes of interest present in a given set of samples. Principal component analysis (PCA) [1–5] is one of the most used chemometric technique to reduce the dimension and simplify the data sets. Application of PCA on a given data sets generates two matrices of a smaller dimension known as score and loading matrices. The score matrix that describes the relationship among the samples could be regressed against the concentration of the analytes to develop a calibration model, commonly known as the principal component regression (PCR) approach [1–3]. Multiple linear regression (MLR) is another technique that can be used to develop the calibration model. It essentially finds a single factor that maximizes the correlation between the predictor and predicted variables [1–3]. The MLR approach has a few limitations, for example (1) the relative abundance of response variables relative to the number of available calibration samples (for the typical spectral calibration problem), which causes an underdetermined situation, and (2) the possibility of collinearity of the predictor variables leading to unstable matrix inversions and regression outcomes [1–3]. The PCR address the issue of collinearity by performing PCA. However, it does not attempt to maximize the correlation between the predictor and predicted variables [1–3]. The limitations of MLR and PCR techniques are required to be overcome with a suitable approach that not only addresses the issue of collinearity but also maximizes the correlation between the predictor and predicted variables.

Partial least square (PLS) algorithm is the most favourite tool of chemometrician to develop a calibration model for analyzing a large volume of data sets in a fast and swift manner [1–3, 6–8]. PLS algorithm essentially provides a means for regressing the predictor and predicted variables against each other [1–3, 6–8]. Herman Wold introduced the PLS algorithm as an econometric technique [1–3, 6–8]. Ever since it was introduced, chemical engineers and chemometricians have been using it. Professor Svante Wold (son of Herman Wold), an avid proponent of the PLS

Partial least square (PLS) algorithm is the most favourite tool of chemometrician to develop a calibration model for analyzing a large volume of data sets in a fast and swift manner.

algorithm, successfully introduced this technique to the community of analytical chemists. PLS algorithm has been successfully integrated with various spectroscopic and chromatographic techniques [3, 9–11]. The integration of PLS algorithm with these techniques have been quite successful. The PLS algorithm successfully finds the underlying low-rank structure present in the spectroscopic and chromatographic data sets that otherwise are highly correlated, multivariate in nature, and large in volume. The PLS algorithm essentially processes the information associated with both predictor and predicted variables and finds a set of factors that not only can explain the maximum variance associated with them but also provide maximum correlation among them. The great advantage associated with the PLS algorithm is that it gives equal importance to both predictors and predicted variables, often not the case in traditional approaches used for creating the calibration models. The theory, various technical details, and an application of the PLS algorithm are given below.

2. Theory

2.1 *Calibration and Validation*

Before proceeding further, it is also important to explain what we mean by calibration and validation [2, 3, 9–11]. Usually, the job of an analytical chemist consist of two steps. In the first step, a calibration model is developed by studying the characteristics of an analytical method or analytical instrument and subsequently establishing a relationship among the predictor and predicted variables. In the context of spectroscopy, it will be establishing the relationship between the spectral intensity and the concentration of an analyte of interest. In the second step, the validation is carried out, wherein the acquired spectral data sets for one or more external sample containing the analyte of interest are subjected to the calibration model. Ideally, the actual and predicted concentrations of the analyte in both calibration and validation step must be as close to each other.

Usually, the job of an analytical chemist consist of two steps. In the first step, a calibration model is developed by studying the characteristics of an analytical method or analytical instrument and subsequently establishing a relationship among the predictor and predicted variables.

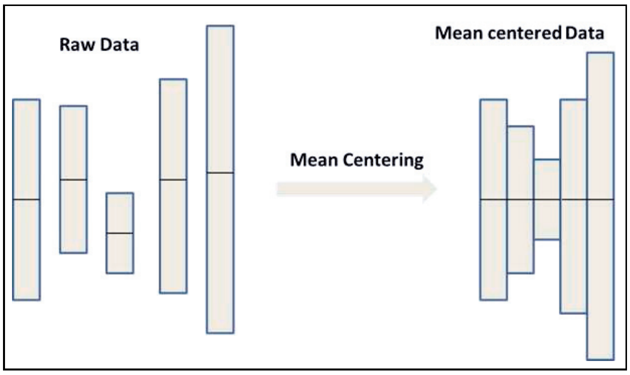

Figure 2. Effect of mean centering on the raw data sets.

2.2 *Data Pre-processing*

Before processing the data with the PLS algorithm, it is important to pre-process the calibration data set to simplify the calculation [2, 3, 9–11]. The mean centering of the data set is the commonly used pre-processing technique. In fact, in the context of spectroscopy, mean centering is an inevitable step. Herein, the mean spectrum of the calibration set is subtracted from each spectrum. Mean centering helps in removing any constant offset and noise, if present in the data sets [2, 3, 9–11]. Moreover, it places the origin of the data to the center. The mean centering is also applied to the concentration values. The effect of the mean centering on the data sets is summarized in *Figure* 2.

Mean centering helps in removing any constant offset and noise, if present in the data sets

2.3 *PLS Algorithm*

As discussed earlier, the PLS model capture maximum variances of spectral and concentration data matrices and ensures maximization of the correlation between them. PLS achieves this using the non-iterative partial least squares (NIPALS) approach [3, 6, 9–12]. The NIPALS approach can be summarized used the scheme given below:

PLS model captures maximum variances of spectral and concentration data matrices and ensures the maximization of the correlation between them. PLS achieves this using the non-iterative partial least squares (NIPALS) approach.

- The acquired spectral data set of the calibration set is stored in a matrix X of dimension $(I \times J)$ where I is the number of samples and J is the number of spectral variables over which spectra are acquired. The concentration-related information is stored in a matrix Y of dimension $(I \times N)$ where I is the number of samples and N is the number of analytes under investigation. The acquired spectral data set of the validation (or set of unknown samples) are stored in a matrix X_{test} of dimension $(I_{\text{test}} \times J)$ where I_{test} is the number of validation set samples.

- Both X and Y data sets are mean centered by subtracting the mean spectrum and mean concentration values, respectively.
- The PLS decomposition is carried out by selecting one of the column of matrix Y (usually the one that has maximum variation). The selected column is designated as u and it is the first column of the score matrix U (of dimension $I \times L$, where L is the number of latent variables) related to the Y block matrix.
- First vector w_1 of weight matrix W (of dimension $J \times L$,) is calculated by projecting the X data set on space spanned by $u_1 : w_1 = X^T u_1/(u_1^T u_1)$.
- Normalize w_1 to unit length: $w_1 = w_1/||w_1||$, where $||w_1||$ is the norm of w_1.
- First score vector t_1 of score matrix T of dimension $I \times L$ is calculated by projecting X in the space spanned by $w_1 : t_1 = Xw_1$.
- First loading vector q_1 of loading matrix Q (of dimension $L \times N$) related to Y data set is calculated by projecting Y matrix in the space spanned by $t_1 : q_1 = Y^T t_1/||t_1^T t_1||$.
- Normalize q_1 to unit length: $q_1 = q_1/||q_1||$, where $||q_1||$ is the norm of q_1.
- First loading vector p_1 of loading matrix P (of dimension $J \times L$) related to X block is calculated by projecting X on the space spanned by $t_1 : p_1 = X^T t_1/||t_1^T t_1||$.
- Normalize p_1 to unit length: $p_1 = p_1/||p_1||$, where $||p_1||$ is the norm of p_1.
- Using normalized loading vector p_1, score (t_1) and weight (w_1) vectors are rescaled: $t_1 = t_1||p_1||$ and $w_1 = w_1||p_1||$.
- Regression coefficient b_1, an element of regression matrix B of dimension $L \times L$ that relates the score vector t_1 and u_1 of X and Y block, respectively, is calculated: $b_1 = u_1^T t_1/(t_1^T t_1)$.

- Next the residuals matrices E_f and F_f for the X and Y data sets respectively are calculated:

$$E_f = E_{f-1}–t_f pf^T andF_f = F_{f-1} - u_f qf^T, \text{ where } X = E_0 \text{ and } Y = F_0.$$

From here, one goes to Step 3 to repeat the procedure for the next latent variable. Please note that these steps are repeated by replacing X and Y with their corresponding residual matrices E and F, respectively.

PLS model that was fitted using a sequential approach can easily be summarized using (1)–(3):

$$X = TP^T + E, \tag{1}$$

$$Y = UQ^T + F, \tag{2}$$

$$U = TB. \tag{3}$$

(1) and (2) describes the decomposition of X and Y data matrices, whereas (3) describes the inner relationship between their score matrices T and U, respectively.

The score matrix for the validation set samples can be calculated as, $T_{\text{test}} = X_{\text{test}} W(P^T W)^-1$. By substituting the score and regression matrices in (2) and (3), the concentrations of analytes of interest could easily be calculated.

Before proceeding further, it is necessary some of the above terminologies be explained

(i) Latent variables: The set of new variables (i.e. t_1 and u_1) obtained from the linear combinations of original variables (X and Y matrices) are defined as the latent variables. The first latent variable explains the maximum variation followed by the second latent variable, and so on.

(ii) Loading vectors: They are a set of orthonormal vectors that form the basis for projecting the original data sets and subsequently simplifying further analysis.

(iii) Weight vectors: They ensure the required orthogonality to the loading vectors.

(iv) Score and loading values: The numerical values associated with latent variables and loading vectors are defined as the score and loading values, respectively. The score values could be used to explain how the samples are related, whereas loading values could be used to study the relationship among the variables.

2.4 *Selecting the Optimum Number of Latent Variables for PLS Analysis with Cross Validation Approach*

As discussed above, one of the key points while developing the PLS model is the selection of an optimum number of latent variables. It is typically achieved using the cross-validation approach that essentially solves two main purposes in PLS analysis. It allows the easy assessment of the optimal complexity of the PLS model and provides a measure for evaluating the PLS model performance when applied to a validation data set. The cross-validation step involves the removal of some of the samples (test set) from the calibration set and constructing the PLS model with the remaining samples (model building set) [2, 3, 9–11]. The developed model is subsequently used for predicting the concentration of the analytes of interest in the test set samples. The same procedure is repeated by varying the latent variables, and a statistical parameter called the root mean square error of cross-validation (RMSECV) is estimated for each of the developed PLS models. The RMSECV values are plotted against the latent variable index [2, 3, 9–11]. From the curve, the number of latent variables that minimize the RMSECV value can be used for developing the PLS model [2, 3, 9–11]. The RMSECV value for a developed PLS model can be calculated using (4).

The cross-validation step involves the removal of some of the samples (test set) from the calibration set and constructing the PLS model with the remaining samples (model building set).

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{I_{\text{test}}} (Y_{\text{actual,i}} - Y_{\text{predicted,i}})^2}{I_{\text{test}}}}. \tag{4}$$

In the above equation, $Y_{\text{actual,i}}$ and $Y_{\text{predicted,i}}$ are the actual and predicted concentrations of the ith sample belonging to the test set and I_{test} is the total number of samples in the test set.

There are two commonly used approaches for creating the test set from a given calibration set (i) leave one out and (ii) Venetian blind [2, 3]. In the first approach, every single sample is used as the test set, whereas in the second approach, a test set is created by selecting every i th sample (starting at sample numbered 1 through i) from the calibration set. The leave one out approach generates I number of test sets, whereas the Venetian blind approach generates i number of test sets with (I/i) number of samples. It is advised that for a smaller set of the sample (< 25), one should use the leave one out approach, whereas for the larger number of samples, one should use the Venetian blind approach.

2.5 *Statistical Parameters to Assess the Performance of Developed PLS Model*

2.5.1. Root mean square error of calibration (RMSEC): RMSEC is a measure of the error in predicting the properties of the samples of the calibration set; it is calculated using (5):

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{I}(Y_{\text{actual,i}} - Y_{\text{predicted,i}})^2}{I}}. \tag{5}$$

$Y_{\text{actual,i}}$ and $Y_{\text{predicted,i}}$ are the actual and predicted concentrations of the i th sample, and I is the total number of samples used to create the calibration model [2].

2.5.2. Root mean square error of prediction (RMSEP): RMSEP is a measure of the error in predicting the properties of the samples of the testing set; it can be calculated using (6):

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{I}(Y_{\text{actual,i}} - Y_{\text{predicted,i}})^2}{I}}. \tag{6}$$

$Y_{\text{actual,i}}$ and $Y_{\text{predicted,i}}$ are the actual and predicted concentrations of the i th sample, and I is the total number of samples used to create the validation set [2].

2.5.3. Square of correlation coefficient (R^2): Square of correlation coefficient (R^2) is a measure of the correlation between actual and predicted concentration for the samples of the calibration set, and it can be calculated using (7):

$$R^2 = 1 - \frac{\sum_{i=1}^{I} (Y_{\text{actual,i}} - Y_{\text{predicted,i}})^2}{\sum_{i=1}^{I} (Y_{\text{actual,i}} - Y_{\text{mean}})^2}. \tag{7}$$

$Y_{\text{actual,i}}$ and $Y_{\text{predicted,i}}$ are the actual and predicted concentrations of the ith sample, Y_{mean} is the mean value for the actual concentration values [13].

In principle, a calibration model should have R^2 of unity, and RMSEC and RMSEP values must be zero.

3. Developing PLS Model: An Example Using Synchronous Fluorescence Spectral (SFS) Data Sets of Gasoline-ethanol Blends

3.1 *Data Used*

The synchronous fluorescence spectral (SFS) data set acquired for a calibration set consisting of 21 samples containing varying ethanol volumes (0–100% in a step of 5%) in gasoline-ethanol blends are taken as an example. The SFS data was acquired at 40 nm wavelength offset over the excitation wavelength range of 250–500 nm. SFS data with the same instrumental parameters were also acquired for a validation set consisting of six gasoline-ethanol blends containing 2, 5, 7, 10, 20, and 40% of ethanol.

3.2 *Software Used*

PLS modelling and other required calculations were carried out on MATLAB platform.

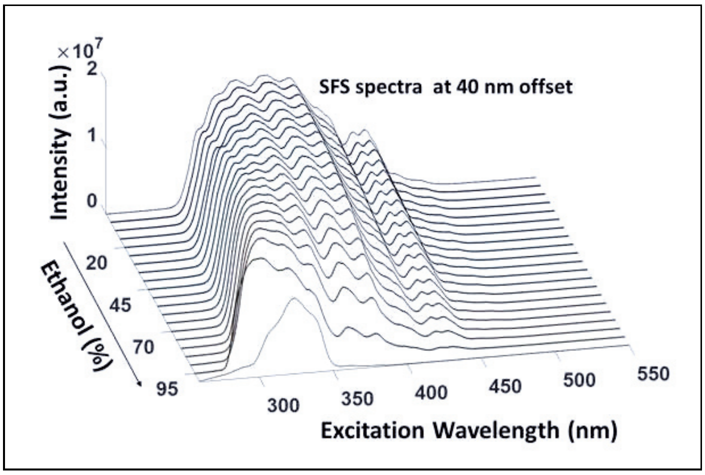

Figure 3. SFS spectra of gasoline-ethanol blends acquired at 40 nm offsets.

3.3 *3.3. Results and Discussion*

The synchronous fluorescence spectra of ethanol-petrol blends are shown in *Figure* 3. From the spectra, a blue shift in the spectra along with spectral intensity reduction could be observed with an increase in ethanol concentration. However, there was no linear relationship between the spectral intensity (at any excitation wavelength) and ethanol concentration. Consequently, the analysis of ethanol in gasoline-ethanol blends demands the application of the PLS technique that improves the correlation between the spectral and concentration data matrices.

Before proceeding further, SFS data of calibration and validation sets were arranged in matrices X and X_{test} of dimension 21×251 (sample × excitation wavelength) and $6{\times}251$ (sample × excitation wavelength), respectively. Ethanol concentration in all the sample of calibrations set are summarized in matrix Y of dimension $21{\times}1$ (sample × number of analytes of interest)

The arranged spectral and concentration data matrices were mean-centered prior to subjecting them to PLS analysis. In the next step, we have to find the optimum number of latent variables required for developing the PLS model. To achieve it, leave one cross-validation approach was used. The RMSECV versus latent variable curve, shown in *Figure* 4, suggested that 5 latent variable

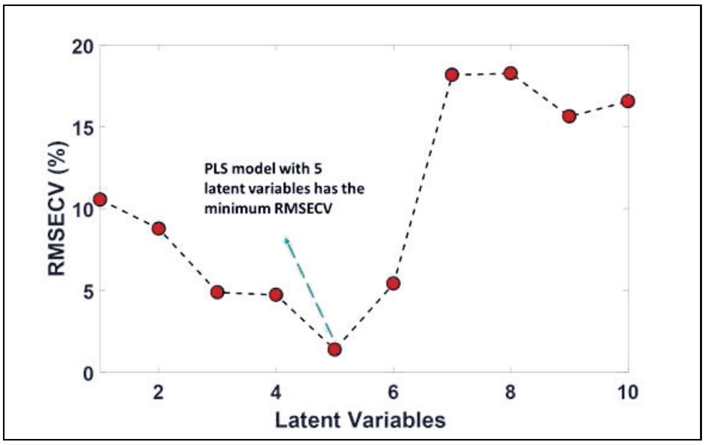

Figure 4. RMSECV versus latent variables plot. It suggest that PLS model with 5 latent variables must be preferred for analysing the ethanol content in gasoline-ethanol blends.

Table 1. Various statistical parameters for assessing the quality of developed PLS model. The statistical parameters clearly suggest that the developed model is robust and can make precise estimation of ethanol content in gasoline-ethanol blends.

Statistical Parameter	**Value**
RMSECV	1.55%
RMSEC	0.55%
R^2	0.999
RMSEP	0.60%

PLS model must be preferred over the others. The developed PLS model was found to explain more than 99.99% variance of both spectral and concentration data sets.

The RMSECV, RMSEC, and R^2 parameters summarized in *Table* 1 clearly suggest that the developed PLS model is highly robust in quantifying ethanol in low as well as in high concentration ranges.

A linear relationship between the actual and PLS predicted ethanol concentration can be seen in *Figure* 5. The RMSECV, RMSEC, and R^2 parameters summarized in *Table* 1 clearly suggest that the developed PLS model is highly robust in quantifying ethanol in low as well as in high concentration ranges. The developed PLS model is tested using the validation set of six samples, the predicted ethanol concentration are summarized in *Table* 2. The RMSEP summarized in *Table* 1, clearly suggest that PLS made precise and accurate predictions for the ethanol concentration in the validation set.

As demonstrated above by taking gasoline-ethanol blends, the behavior of complicated system can be easily captured by analyzing a series of variables than relying on a single variable measure-

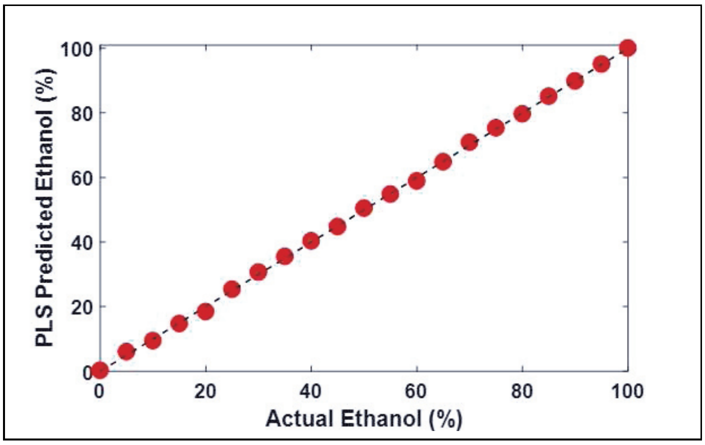

Figure 5. Actual and PLS predicted ethanol (%) in gasoline ethanol blends. The actual and predicted concentration are linearly related, and are in close correspondence. It shows that PLS algorithm can be useful for analysing even a complex system such as gasoline-ethanol blends.

Actual ethanol (%) in gasoline-ethanol blend	**Predicted ethanol (%) in gasoline-ethanol blend**
2	2.22
5	5.87
7	7.81
10	10.05
20	20.14
40	40.85

Table 2. Actual and PLS predicted ethanol (%) in gasoline-ethanol blends. The PLS predicted ethanol concentration for the six samples of validation set are in close correspondence to their actual concentrations.

ment. PLS algorithm can really serve as a useful tool to achieve this in a swift manner without really demanding too much of technical and computational skills.

4. Conclusions

PLS technique is the favorite tool in chemometrics to develop a calibration model for quantifying the analyte of interest. PLS technique not only captures the maximum variations associated with predictor and predicted variables but also maximizes the correlation between them. The application of PLS was successfully demonstrated by analyzing the SFS data of gasoline ethanol blends.

Acknowledgment

The data used in the current article was collected during my PhD at Department of Chemistry, Indian Institute of Technology-Madras

Suggested Reading

[1] D L Massart, B G M Vandeginste, L M.C Buydens, S de Jong, P J Lewi, V J S Verbeke, *Handbook of chemometrics and qualimetrics*, Elsevier, New York , 1997.

[2] R Kramer, *Chemometric techniques for quantitative analysis*, Marcel Dekker, New York, 1998.

[3] R Brereton, *Chemometrics for Pattern Recognition*, John Wiley & Sons, Ltd, U.K., 2009.

[4] S Wold, Chemometrics, What do we mean with it, and what do want from it, *Chemometrics and Intelligent Laboratory Systems*, Vol.30, pp.109–115, 1995.

[5] K Kumar, Principal Component Analysis: Most Favourite Tool in Chemometrics, *Resonance*, Vol.22, pp.747–759, 2017.

[6] S Wold, PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, Vol.58, pp.109–130, 2001.

[7] A Höskuldsson, PLS regression methods, *Journal of Chemometrics*, Vol.2, pp.211–228, 1988.

[8] G G Dumancas, S Ramasahayam, G Bello, J Hughes, R Kramer, Chemometric regression techniques as emerging, powerful tools in genetic association studies, *Trends in Analytical Chemistry*, Vol.74, pp.79–88, 2015.

[9] R G Brereton, *Applied Chemometrics for Scientists*, John Wiley & Sons, Ltd, U.K., 2007.

[10] R G Brereton, *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, John Wiley & Sons, Ltd, U.K., 2003.

[11] R G Brereton, *Chemometrics: Data Driven Extraction for Science*, John Wiley and sons, Ltd, U.K., 2018.

[12] K Varmuza, P Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*, Taylor & Francis Group, Boca Raton, FL, 2008.

[13] R G D Steel, J H Torrie, *Principles and Procedures of Statistics*, McGraw-Hill, New York, 1960.

Address for Correspondence
Keshav Kumar
Geisenheim University of
Applied Sciences
Germany
Email:
keshavkuma29@gmail.com