# Principal Component Analysis: Most Favourite Tool in Chemometrics

*Keshav Kumar*

**Principal component analysis (PCA) is the most commonly used chemometric technique. It is an unsupervised pattern recognition technique. PCA has found applications in chemistry, biology, medicine and economics. The present work attempts to understand how PCA work and how can we interpret its results.**

**Keshav Kumar did his PhD from Department of Chemistry, Indian Institute of Technology-Madras, India, under the guidance of Professor A K Mishra. Currently he is working as a Postdoc at the Institute for Wine Analysis and Beverage Research, Hochschule Geisenheim University, Germany. His research mainly focus on chemometrics and its application in various fields.**

## 1. Introduction

Chemometrics is a discipline that combines mathematics, statistics, and logic to design or select optimal measurement procedures and experiments. It allows the extraction of maximum relevant chemical information by analysing chemical data and helps in understanding chemical systems [1]. In recent years, chemometrics has emerged as an important part of analytical chemistry. Chemometric techniques have enabled the analysis of large volumes of data obtained from various instruments (single or hyphenated) efficiently. The analyses of such large data sets are otherwise a time consuming process and might end up with no meaningful interpretation or conclusions.

Among various chemometric techniques, principal component analysis (PCA) [1, 2] is considered the 'most favourite'. PCA has found applications in various fields. For example, Singh *et al.,* have successfully used PCA for stellar spectral classification [3]. Kumar *et al.,* have applied PCA for (i) classifying aqueous herbal drugs [4] and (ii) diagnosis and therapeutic prognosis of oral submucous fibrosis [5]. Kowalski *et al.,* have used PCA for the classification of archaeological artefacts [6]. Kowalkowaski has applied PCA for river water classification [7], while Ragot and coworkers have used PCA for air quality monitoring [8].

**Keywords**
Chemometrics, principal component analysis, classification, pattern recognition, chromatography.

Data compression by PCA involves finding a new space spanned by fewer number of dimensions over which original data set is projected. The dimensions of the new space are orthogonal to each other simplifying the data sets for further analysis.

It can be realised that PCA is capable of providing a fast and effective way of analysing data sets from various disciplines *viz* physics, biology, chemistry, archaeology, etc. PCA essentially reduces the dimensions of the data set while retaining most of the variation [1, 2]. Data compression by PCA involves finding a new space spanned by fewer number of dimensions over which original data set is projected. The dimensions of the new space are orthogonal to each other simplifying the data sets for further analysis. Theoretical and various technical aspects of PCA are discussed below.

## 2. Theory

### 2.1 *Geometrical Representation of PCA*

In order to understand PCA geometrically, let us consider a two dimensional data set $I \times J$, where $I$ is the number of samples and $J$ is the number of variables. In the present case, for convenience, we have set the number of variables (*i.e.,* $J$) to two – $J_1$ and $J_2$. As shown in *Figure* 1, these samples can be presented in a two dimensional space spanned by $J_1$ and $J_2$. The two axes $J_1$ and $J_1$ are orthogonal to each other. The data set acquired for the samples have considerable amount of variation along $J_1$ and $J_2$ axes. In other words, both the dimensions are significantly important to have the complete information about the sample set.

An anti-clock wise rotation of the $J_1$ and $J_2$ axes by an angle $\theta$ (= $45^{\text{o}}$ in the present case) generates another pair of orthogonal axes $T_1$ and $T_2$. Mathematically, it could be shown using (1):

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} J_1 \\ J_2 \end{pmatrix}$$

The new variables (or dimensions) $T_1$ and $T_2$ are the linear combinations of $J_1$ and $J_2$ variables with sine and cosine as coefficients

$$T_1 = J_1 \cos\theta + J_2 \sin\theta \qquad (1)$$

$$T_2 = -J_1 \sin\theta + J_2 \cos\theta \qquad (2)$$

Projection of the data set in space, spanned by the new variables $T_1$ and $T_2$ is shown in *Figure* 2. The data set has most of the variations along $T_1$ axis and is literally invariant along $T_2$ axis.

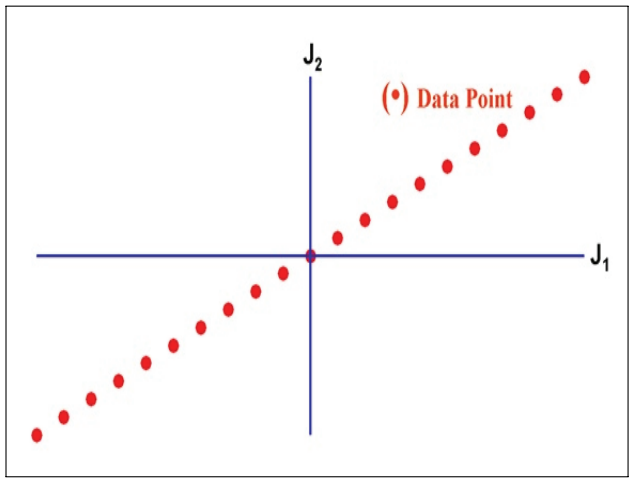


**Figure 1.** Representation of a data set in the space spanned by $J_1$ and $J_2$. Data has significant variation along both the axes.

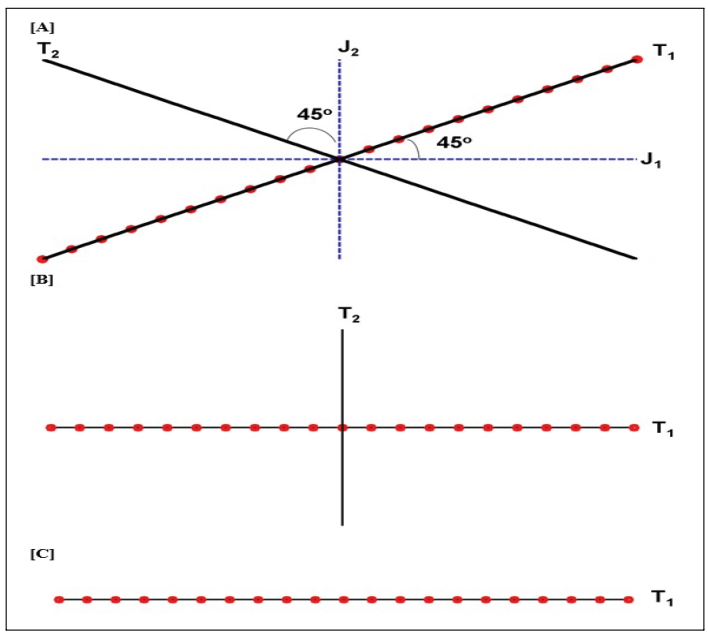


**Figure 2.** **(a)** Rotation of axes $J_1$ and $J_2$ by 45° to generate another pair of orthogonal axes $T_1$ and $T_2$. **(b)** Representation of data set in the new space spanned by $T_1$ and $T_2$ axes. The data set has variation along $T_1$ and no variation along $T_2$. **(c)** Reduction of dimensions. $T_2$ is unimportant and hence could be removed, and $T_1$ can be taken as the approximation of data spanned in the two dimensional space spanned by $J_1$ and $J_2$.

While the score value explains how the samples are related to each other, the loading value explains how the variables are related to each other.

In principle, variation along $T_1$ axis can be taken as a good approximation of the original two-dimensional data set, and one can easily ignore the $T_2$ axis. Thus, by projecting the data set in a suitable space, it is possible to reduce the dimensions of the data sets while retaining all the information.

### 2.2 *Commonly Used Terminologies in PCA*

Before proceeding further, it is necessary that we briefly describe some commonly used terminologies.

(i) *Principal Components*: The set of new variables (*i.e.,* $T_1$ and $T_2$) obtained from the linear combinations of old variables ($J_1$ and $J_2$) are called principal components. The variable that explains the maximum variation is called the first principal component. Second principal component explains the second highest variation from the unexplained variance of the data set and so on. In the above given example, $T_1$ is the first principal component and explains all the variations of the data set, and $T_2$ is the second principal component that explains the remaining variance of the data set.

(ii) *Loading Vectors*: They essentially form the basis for projecting the original data set to obtain the principal components. In the above example, $[\cos\theta \sin\theta]^T$ and $[-\sin\theta \cos\theta]^T$, transpose of first and second row of the matrix, respectively, given in (1) represents the first and second loading vectors corresponding to first ($T_1$) and second ($T_2$) principal components. The loading vectors in more generic sense are known as Eigen vectors of the data set.

(iii) *Score and Loading Value*: The numerical values associated with principal components of each sample are called the score values. The numerical values associated with the elements of loading vectors are called the loading values. The score values explain how the samples are related to each other, and the loading values explain how the variables are related to each other.

### 2.3 *Fitting PCA Model*

PCA model can be fitted using Eigen value decomposition method as summarized below.

(1) In the first step, the data set is mean centered $X = X - mean(X)$.

(2) In the second step, the covariance matrix for mean centred $X$ is calculated,

$$\mathrm{Cov}(X) = \frac{X^{\mathrm{T}}X}{I-1}$$

(3) In the next step, the covariance matrix is diagonalized to obtain the $\wedge$ (Eigen values) and $P$ (Eigen matrix containing the Eigen vectors): $\mathrm{Cov}(X)P = \wedge P$.

(4) In this step, the diagonal elements in the matrix $\wedge$ are arranged in the decreasing order (*i.e.,* $\wedge_1 > \wedge_2 > \wedge_3 ... > \wedge_k$), and the corresponding arrangement is made in the loading matrix $P$. For example if the positions of $\wedge_1$ and $\wedge_3$ are interchanged in the matrix $\wedge$, then the first and third rows of $P$ are interchanged.

(5) Score matrix $T$ can be calculated by projecting the data set $X$ in the space spanned by the Eigen vectors of matrix $P$ : $T = XP$. The score matrix $T$ and loading matrix $P$ are orthogonal and orthonormal, respectively. $T^T T$ = diagonal matrix, and $P^T P$ = identity matrix.

(6) The approximation of data set $X$ by PCA model can be represented as: $X = TP^T + E$.

$E$ is the residual matrix of dimension $I \times J$, $T$ is the score matrix of dimension $I \times K$, and $P$ is loading matrix of dimension $J \times K$, where $K$ is the number of significant factors of PCA model that explain majority of the variation in the data set and it is always $\leq$ min( $I$ and $J$).

(7) Score value of any new sample can be calculated by projecting $X_{\mathrm{new}}$ new data on $P$: $T_{\mathrm{new}} = \mathrm{X}_{\mathrm{new}}\, P$.

We can also use the autoscaled data in step 1 to perform PCA analysis. It is very useful when the variables are in different scales or in different magnitudes. The data set is autoscaled

Autoscaled data is very useful when the variables are in different scales or in different magnitudes. The data set is autoscaled by subtracting the mean from each column, followed by division with the standard deviation.

The choice of number of factors can only affect the extent to which one can retrieve different pieces of orthogonal information. Thus, a PCA model can be created with any number of factors, and each model would provide a true piece of information available in the data set.

by subtracting the mean from each column, followed by division with the standard deviation:

$X = [X - \text{mean}\ (X)]$ /standard deviation $(X)$.

Steps 2–7 can be performed on autoscaled $X$ to create the PCA model. It is to be noted that square matrix obtained in step 2 is defined as the correlation matrix.

### 2.4 *Finding Optimum Number of Factors for PCA Model*

One of the significant advantages of PCA is that it is essentially sequential in nature. In other words, $K$ factor PCA model is always a subset of $K + 1$ factor PCA model. The choice of number of factors can only affect the extent to which one can retrieve different pieces of orthogonal information. Thus, a PCA model can be created with any number of factors, and each model would provide a true piece of information available in the data set. In order to ensure that we capture the complete information of the data set without overfitting the model, a general thumb rule is that one has to first select $k$ factors from available $K$ factors that could capture at least 80% of the data set.

$$\text{Amount of variance} = \frac{\sum_{i=1}^{k} \wedge_i}{\sum_{i=1}^{K} \wedge_i}, > 80\%; k \leq K.$$

One has to keep adding the number of factors if the amount of variance captured by the model increases by more than 3–4%.

### 2.5 *Some Statistical Parameters Involved in PCA*

*Lack of Fit Parameter (Q)*: It is a measure of the difference between the actual data set and the approximation made by the PCA model. The lack of fit parameter $(Q)$ of PCA model can be calculated by taking the outer product of the residual matrix $E$.

$Q = EE^T = X(I - PP^T)X^T$.

In an ideal case, the diagonal elements of $Q$ (a matrix of dimension $I \times I$) should be zero.

*Hotelling's $T^2$ Statistic*: This parameter measures the variation of each sample within the PCA model. It indicates the spread of samples from the origin in the model. It could be calculated as, $T^2 = T\lambda^{-1}T^T$.

*Leverage*: It measures the influence of a sample in PCA model. A sample with high leverage reinforces PCA model and may cause the rotation of principal components. Leverage of a sample can be calculated using the formula: Leverage = $T(T^TT)^{-1}T$.

### 2.6 *Detection of Outliers*

Principal component analysis can be used to find the outlier in a data set, provided we study the leverage and residual of the samples. A sample that is not well described by the model will have unusually high residual, and the samples that have high influence on the model will have unusually high leverage [9]. In an ideal case, all the samples of a data set should have low leverage and low residuals. The samples having high residual are classified as outliers and need to be analyzed carefully. The samples having high leverage may have high residual or low residual. The former is called as bad leverage samples, and the latter is called as good leverage samples. The bad leverage samples need to be analysed very carefully because they tend to bias the model significantly. The good leverage samples also have to be analysed carefully as they indicate unusual variation of the variable, though the data set of those samples are well fitted by PCA model.

## 3. Performing PCA: An Example

### 3.1 *Data Used*

The chromatographic data set reported in literature [10, 11] has been used to carry out the present work. The chromatographic data set consist of 120 oil samples. Of these samples, 68 belongs to the class of olive oils and 52 belongs to the class of non-olive vegetable oils, and oil blends (*i.e.*, the vegetable and olive oil blends).

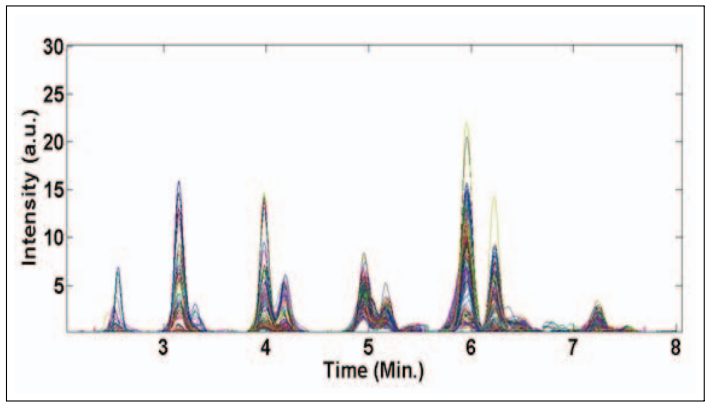


**Figure 3.** The chromatograms of 120 oil samples. Of this, 68 belongs to the class of olive oils and the remaining belongs to the class of non-olive oils and blended oils.

### 3.2 *Software Used*

All the analysis and data plotting was carried out on MATLAB-2014 platform. However, there are other platforms such as R, python, *etc*., that can be used for the analysis.

### 3.3 *Results and Discussion*

The chromatographic data sets acquired for the olive and non-olive oils are shown in *Figure* 3. It can be seen that based on the visual analysis of the chromatographic profiles, it is difficult to differentiate olive oils from non-olive vegetable oils. Moreover, manual analysis of such a large volume of data is laborious and time consuming, and may not provide any meaningful interpre-

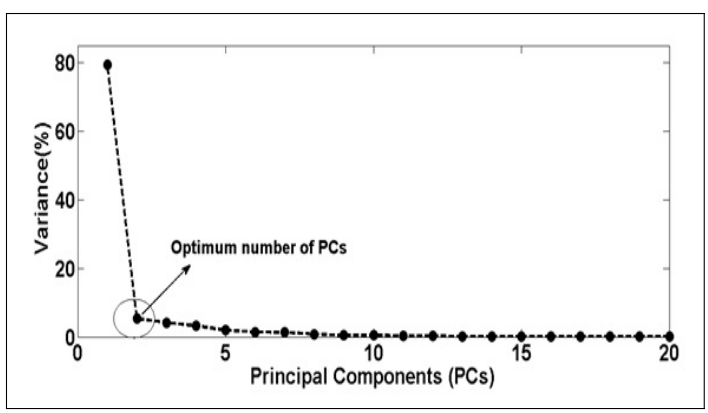


**Figure 4.** Amount of variance captured by different principal components (PCs). The plot indicates that first two PCs are sufficient to explain most of the variance (more than 85%) of the data set without overfitting the model.

tations. However, PCA essentially simplifies and reduces the dimensions of the data set, and provides a fast and efficient way of analysing the complex chromatographic data of the selected oil samples. The chromatographic data sets are arranged in a matrix of dimensions $126 \times 4001$, where 126 is the number of samples and 4001 is the number of variables (*i.e.*, retention time points). The chromatographic data sets are normalized to unit area and mean-centered prior to PCA. The optimum number of principal components required for fitting PCA model is obtained from the variance captured by different principal components against the principal component numbers, as shown in *Figure* 4. It can be seen that with the addition of principal components, there is a substantial improvement in the cumulative percentage of variance captured by the PCA model. However, beyond 2 principal components, the addition of extra factors do not bring any substantial improvement in the cumulative variance captured by the PCA model. Thus, one can conclude that PCA model of two principal components that explains more than 80% variance is optimum to capture all the important information buried in the data set. PC1 and PC2 individually explains 78% and 5% variances of the data sets. The PC1 versus PC2 score plot is shown in *Figure* 5; PCA model clearly separates the samples in two groups. It is found that all the samples belonging to the class of olive oils have negative PC1 score values and the samples belonging to the non-olive oil class have positive PC1 score values. The blended oils

The optimum number of principal components required for fitting PCA model is obtained from the variance captured by different principal components against the principal component numbers.

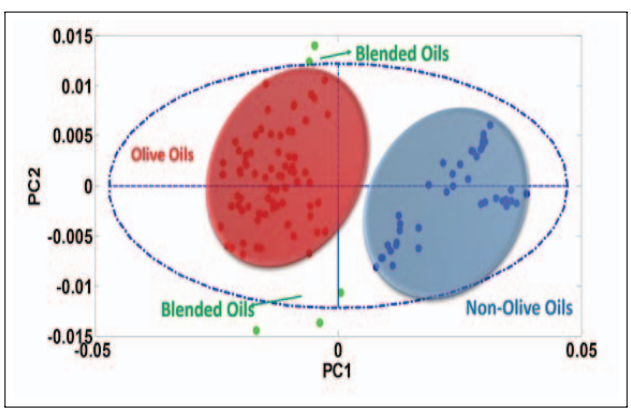


**Figure 5.** PC1 versus PC2 score plot classifying the olive oil samples from non-olive oils and blended oil samples.

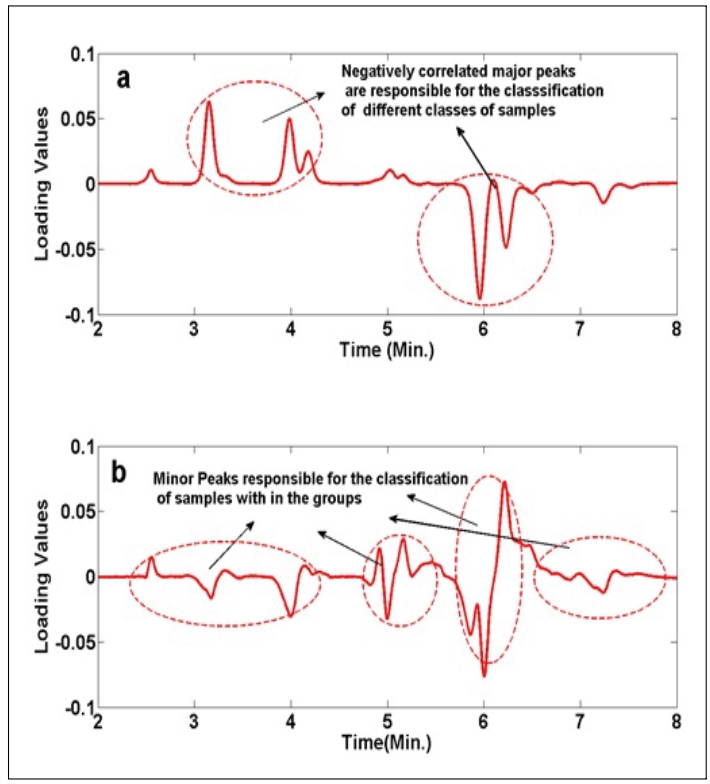


**Figure 6.** **(a)** Loading vectors corresponding to PC1 that mainly contains negatively correlated major peaks can be used to rationalize the classification of olive oil, non-olive oil, and blended oil samples in the PC1 versus PC2 score plot. **(b)** Loading vectors corresponding to PC2 mainly explains the minor peaks that can be used for the classification of the samples within the groups.

in the score plot appear near the edges of the ellipse. The blended oil containing more of olive oils have negative PC1 score values, whereas the blended oils containing less of olive oils have positive PC1 score values. The loading vectors that explain how the variables are related to each other are shown in *Figure* 6.

The analysis of loading vector plot can be really helpful in finding the set of variables that are really helpful in characterizing the samples. The loading vector corresponding to PC1 mainly explains the variations of the major peaks that can be used to characterize the classes of olive oils and non-olive oils. The loading vector corresponding to PC2 mainly explains the variation of the minor peaks and can be used to differentiate the samples within the olive oil and non-olive oil groups. Based on the loading vector profiles, the appearance of blended oil samples on the extreme edges of PC2 axes can be attributed to the fact that blended oils

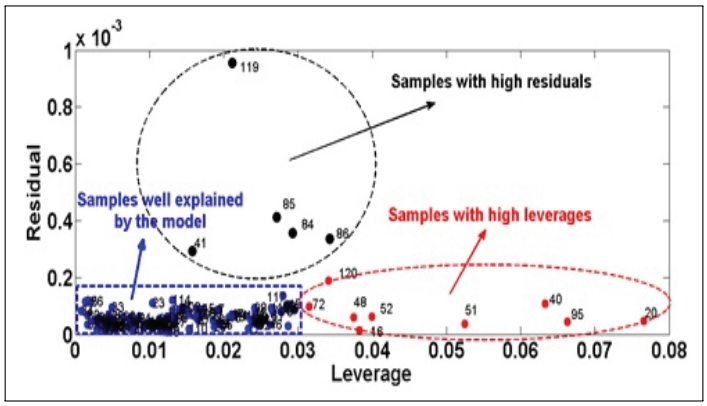


**Figure 7.** The outlier diagnostic plot explaining the leverage and residual values in two component PCA model. Nine samples (16, 20, 40, 48, 51, 52, 72, 95 and 120) are found to have high leverage values. Of this, 5 samples (16, 20, 40, 51 and 95) are the blended oils. Five samples (41, 84, 85, 86 and 119) are found to have high residual values indicating that their composition is very different from others or something went wrong at the sample preparation or data acquisition stages and need further attention.

contain different types of olive and non-olive oils. The outlier diagnostic plot created by plotting the leverage versus residual values can be used to find the really unusual samples called the outliers. The outlier diagnostic plot is shown in *Figure* 7. All the 5 blended samples (16, 20, 40, 51, and 95) are found to have unusually high leverage values that correlate well with the fact that they contain constituents of both olive and non-olive oils. There are some samples with high residual values indicating that these samples need a careful analysis. These samples might have unusual compositions or something might have gone wrong at the sample preparation or data acquisition stages. In summary, the obtained PCA model is found to be highly specific and sensitive in classifying the oil samples.

In most cases, PCA is well capable of classifying the samples. Though, in some cases due to the complexity of the data sets, the output of PCA such as score matrix needs to be further processed with some other chemometric techniques such as linear discriminant analysis (LDA) [12], soft independent modelling of class analogy (SIMCA) [13,14], neural network analysis (NNA) [3,14], *etc*., for achieving meaningful interpretation of the data sets.

### 3.4 *Conclusions*

PCA is the most favourite tool in chemometrics. It reduces the dimensions of the data sets and simplifies the data for easy and meaningful interpretations. Using the chromatographic data set of olive and non-olive oil samples, it has been clearly shown that PCA can be used as an unsupervised pattern recognition technique. PCA successfully differentiated olive oil from non-olive oil samples. It is also shown that PCA can be used for detecting the outlier samples in the data set.

## Suggested Reading


[1] D L Massart, B G M Vandeginste, L M C Buydens, S de Jong, P J Lewi and 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] H P Singh, R K Gulati and Ranjan Gupta, Stellar Spectral Classification Using Principal Component Analysis and Artificial Neural Networks, *Monthly Notices of the Royal Astronomical Society*, Vol.295, pp.312–318, 1998.

[4] K Kumar, P Bairi, K Ghosh, K K Mishra and A K Mishra, Classification of Aqueous-based Ayurvedic Preparations Using Synchronous Fluorescence Spectroscopy and Chemometric Techniques, *Current Science*, Vol.107, No.3, p.107, 470–477, 2014.

[5] K Kumar, S Sivabalan, S Ganesan, and A K Mishra, Discrimination of Oral Submucous Fibrosis (OSF) Affected Oral Tissues From Healthy Oral Tissues Using Multivariate Analysis of In-vivo Fluorescence Spectroscopic Data: A Simple and Fast Procedure for OSF Diagnosis, *Analytical Methods*, Vol.5, pp.3482–3489, 2013.

[6] B R Kowalski, T F Schatzki and F H Stross, Classification of Archaeological Artifacts by Applying Pattern Recognition to Trace Element Data, *Analytical Chemistry*, Vol.44, pp.2176–2180, 1972.

[7] T Kowalkowski, R Zbytniewski, J Szpejna and B Buszewski, Application of Chemometrics in River Water Classification, *Water Research*, Vol.40, pp.744–752, 2006.

[8] M F Harkat, G Mourot and J Ragot, An Improved PCA Scheme for Sensor FDI: Application to An Air Quality Monitoring Network, *Journal of Process Control*, Vol.16, pp.625–634, 2006.

[9] S Wold, K Esbensen and P Geladi, Principal Component Analysis, *Chemometrics and Intelligent Laboratory Systems*, Vol.2, pp.37–52, 1987.

[10] de la P Mata-Espinosa, J M Bosque-Sendra, R Bro and L Cuadros-Rodriguez, Olive Oil Quantification of Edible Vegetable Oil Blends Using Triacylglyc-

**erols Chromatographic Fingerprints and Chemometric Tools, *Talanta*, Vol.85, pp.177–182, 2011.**

**[11] http://www.models.life.ku.dk/oliveoil**

**[12] S Balakrishnama and A Ganapathiraju, Linear Discriminant Analysis A Brief Tutorial, Institute for Signal and Information Processing, March 2, 1998. https://www.isip.piconepress.com/publications/reports/1998/isip/lda/lda_theory.pdf**

**[13] S Wold, Pattern Recognition by Means of Disjoint principal component models, *Patt. Recog*., Vol.8, pp.127–139, 1976.**

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

*Address for Correspondence*
Keshav Kumar
Institute for Wine Analysis and Beverage Research,
Hochschule, Geisenheim University, Geisenheim 653 66
Germany
Email:
keshavkumar29@gmail.com