EUROPEAN JOURNAL OF BIOMEDICAL AND PHARMACEUTICAL SCIENCES
http://www.ejbps.com

ISSN 2349-8870
Coden USA: EJBPBO
Volume: 12
Issue: 12
Year: 2025

A BRIEF DESCRIPTION OF CHEMOMETRIC TECHNIQUES AND THEIR TYPES

Pasala Lahari Sri Pratyusha[1], Agraharapu Venkateswararao[1*], Nagumalli Lakshmi Tulasi[1]

Department of Pharmaceutical Analysis, School of Pharmaceutical Sciences and Technologies, Jawaharlal Nehru Technological University - Kakinada (JNTUK), 533003, Andhra Pradesh, India.

***Corresponding Author: Agraharapu Venkateswararao**
Department of Pharmaceutical Analysis, School of Pharmaceutical Sciences and Technologies, Jawaharlal Nehru Technological University - Kakinada (JNTUK), 533003, Andhra Pradesh, India.

DOI: https://doi.org/10.5281/zenodo.17799493

How to cite this Article: Pasala Lahari Sri Pratyusha1, Agraharapu Venkateswararao1*, Nagumalli Lakshmi Tulasi1. (2025). A BRIEF DESCRIPTION OF CHEMOMETRIC TECHNIQUES AND THEIR TYPES. European Journal of Biomedical and Pharmaceutical Sciences, 12(12), 324–334.

Article Received on 05/10/2025 **Article Revised on 25/11/2025** **Article Published on 01/12/2025**

ABSTRACT

Chemometric techniques improve the interpretation of complicated datasets by analyzing chemical data using statistical and mathematical methodologies. These techniques are essential in many domains, such as medicines, pharmaceutical analysis, and environmental monitoring. Chemometrics makes it possible to take out useful information from high-dimensional data, like spectral or chromatographic data, by using multivariate analysis. In this, the chemometric techniques classification is presented and described the various classifications of the chemometric techniques. It involves the brief information and example graph structure of the each classified technique presented in this review. Chemometrics plays a major role in modern pharmacy by applying mathematical, statistical and computational methods to extract meaningful information from complex data. It enhances accuracy, reduces analysis time and supports automation in pharmaceutical quality control and research. Applications of chemometrics are also involved.

KEYWORDS: Chemometric techniques, interpretation, multivariate analysis, computational methods.

INTRODUCTION

By applying statistical as well as mathematical techniques into information analysis in order for maximizing the collection, also take out valuable details is known as chemometrics. This has been exhibit how chemometric perspectives can be used as instruments to enable various dimensional adjustment of specific spectroscopic, electrochemical, and chromatographic procedures. Both the identification, quantitative analysis of drug substances in complex mixtures has been facilitated by the use of this perspective, primarily for the elucidation of UV-Visible and near infrared spectra, also for data obtained through various instrumental methods. This is particularly applicable when examining pharmaceutical preparations that are available in market. The employment of sophisticated chemical instruments and data processing for such analytical work has created a demand for sophisticated techniques for experiment design, equipment calibration, and data analysis.[1] Three specialized journals emerged in the subject throughout the 1980s: Journal of Chemometrics, Chemometrics and Intelligent Laboratory Systems, and Journal of Chemical Information and Modeling. Both basic and methodological chemometrics studies are still covered in these publications. Currently, application-oriented journals frequently publish the majority of routine applications of current chemometric techniques.[2] The traditional method is reductionist and looks at one item at a time, separating the impacts as much as feasible. The multivariate approaches used in the chemometrics approach take into account all variables simultaneously. The model fits the data in this way. The conclusions drawn from developing a model to suit the data should be consistent with the information found in the data. This stands in stark contrast to the traditional method, which derives the model from theory and searches the data to demonstrate the model's validity.[3]

Chemometrics flow chart The chemometric techniques are divided mainly into two types as Regression techniques and Classification techniques. They are further divided into various sub types as shown in the figure 1 flow chart.

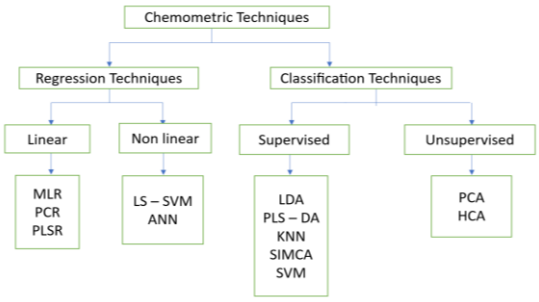

Fig. 1: Flow chart of chemometric techniques.[4]

Chemometric Techniques

1. Regression Algorithms

A kind of machine learning method that models the association allying a variable and single or several input variables for the purpose of predicting a continuous numerical result. In order to forecast things like home prices or stock market trends, the goal is to identify the "best fit" line, curve, or surface through the data that minimizes the difference between expected and actual values.

a. Linear

The dependent and independent variables are assumed to have a linear relationship.

i. Multiple Linear Regression (MLR)

A statistical method called multiple linear regression is used to forecast a variable's result based on the values of two or more variables. It is an extension of linear regression and is sometimes referred to simply as multiple regression. The dependent variable is the one we wish to predict, and the independent or explanatory variables are the ones we use to forecast the dependent variable's value.

Equation

$$Y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta pxp + \epsilon$$

Where,

- y = dependent variable
- β0 = y-intercept
- β1 & β2 = regression coefficients representing the change in y relative to x1 and x2, respectively.
- βp = slope coefficient for each independent variable
- ϵ = model's random error term.

Statisticians can use the information available about one variable to forecast the value of another by applying simple linear regression. The goal of linear regression is to determine a straight-line relationship with the two variables.

The dependent variable exhibits a linear relationship with two or more independent variables in multiple regression. The dependent and independent variables may not accompany a unbending line in a non-linear situation.

Both linear and non-linear regression use two or several variables to graphically track a particular response. However, because non-linear regression is based on trial-and-error assumptions, it is typically demanding to execute.

It is based on the following assumptions:

- Direct relation between the dependent and independent variables
- The independent variables are not highly related to one another
- The variance of the residuals do not same
- Observation independence
- Multivariate normality[5]

For the better understanding of MLR, the example structure of the MLR graph is shown in the figure 2 as given below.

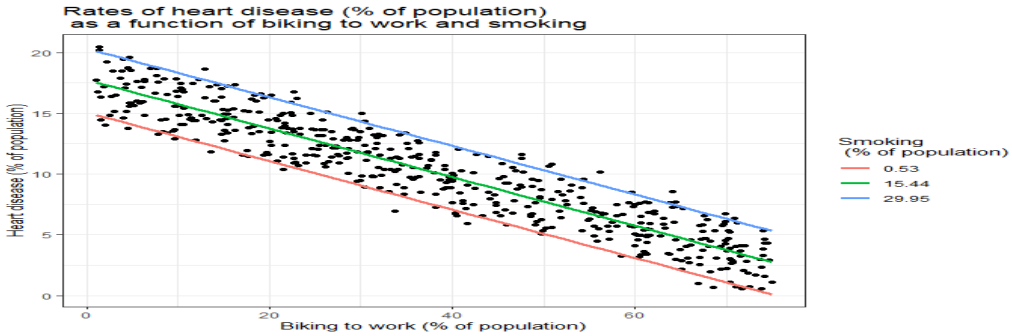

Figure 2: Graph of MLR.[6]

ii. PCR

It combines linear regression with principal component analysis. It is applied when there is a strong correlation or multicollinearity between independent variables. It is a different approach to the MLR.

Using a model of the form y = Xb, we relate the columns of two matrices (blocks) in multiple linear regression: X, a N × K matrix, to the single vector, y, a N × 1 vector. Solving b = (X'X)-1X'y yields the answer vector b. The calculated solution's variance is provided by.

We purposefully configured our procedure to produce a matrix X with independent columns in the factorial experiments phase. This produces a diagonal X'X matrix since every column is extraneous to another also couldn't be expressed in with reference to others.

However, the columns in X are connected on the majority of data sets. If a column is only slightly correlated, it is not very serious. However, the problem with strongly correlated variables is illustrated here; in this case, x1 and x2 have a strong positive correlation.

A prediction model for y is constructed using both variables. The residual error is minimized by finding the model plane, y^ = b0 + b1x1 + b2x2. The sum of squares of the residual error, or the goal function, has a single minimum. However, as the plane go around all over the axis of correlation, extremely slight changes in the raw x-data result in nearly no exchange in the design criterion, but they may cause significant oscillations in the explanation for b. This illustration helps to visualize this.

If we make little adjustments to the raw data, the plane will revolve around the dashed, axial line. The values of b1 and b2 will vary greatly with each successive rotation, even changing in sign (!), but the minimum value of the goal function remains relatively constant. Due to the huge off-diagonal elements in X'X, this phenomenon manifests in the least squares solution as wide confidence ranges for the coefficients. When attempting to learn about your process from this model, this has significant implications: you must use caution. The data strongly supports a model with low or uncorrelated variables, which cannot be freely rotated.

Returning to variable selection is the typical "solution" to this collinearity issue. The modeler would choose either x1 or x2 in the scenario above. Generally speaking, instead of using the entire matrix, the modeler must choose a subset of uncorrelated columns from the K columns in X. This creates a significant computing load when X is large. Furthermore, it is unclear how many columns should be in the subset and what the trade-offs are.

Another issue with MLR is the assumption that the variables in X are measured accurately, which is precisely what causes the spinning plane to become unstable and is known to be false in many real-world engineering scenarios. Moreover, missing data cannot be handled by MLR. In summary, multiple linear regression has the following drawbacks:

- It assumes X is noise-free, which is rarely the case in practice;
- It cannot handle strongly correlated columns in X; and
- It cannot manage missing values in X
- Variable selection is necessary to satisfy the N > K condition and provide independence between columns of X, but this process is not evident and may result in less-than-ideal predictions. MLR demands that N > K, which can be impracticable in many situations.

Principal component regression works by substituting the uncorrelated A score vectors from PCA for the K columns in X.[7]

Formula

EY = XB+F

Where, Y = response matrix (dependent variable)
X = matrix of predictors
B = multiplication of loadings matrix and the regression coefficients between scores and response.[8]

For the better understanding of the PCR, an example structure of PCR graph is shown in the figure 3 given below.

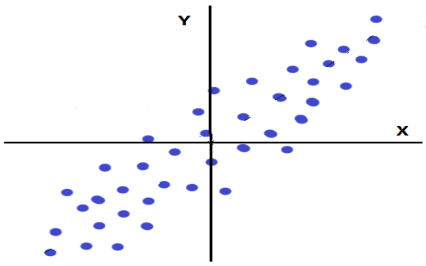

Figure 3: PCR Graph.[9]

iii. PLSR (Partial Least Squares Regression)

PCR and multiple regression features are combined and generalized in a new technique called PLS regression. When we need to forecast a set of dependent variables from a big set of independent variables, it is especially helpful. It first gained popularity in chemometrics, or computational chemistry, thanks in part to Herman's son Svante and in sensory evaluation. Originally, it has been applied in social sciences, notably economy. However, pls regression is also emerging as a preferred multivariate method in the social sciences for both investigational and non-investigational data. This has been quickly interpreted in a statistical framework after initially being presented as an algorithm similar to the power technique.

There are numerous options for selecting the latent vectors. Actually, in the earlier formulation, the role of T might be played by one group of extranous vectors spanning the vertical space of X. Further conditions must be met in order to specify T. This is equivalent to determining two groups of weights, w and c, for multivariate regression so that the verticals of X and Y can be linearly combined in a way that maximizes their covariance. In particular, the objective is to find a first pair of vectors, t = Xw and u = Yc, under the conditions that wTw = 1, t Tt = 1, and t Tu be maximal. The process is repeated until X becomes a null matrix after the first hidden vector is discovered and removed from both X and Y.

A schematic of the original approach can be used to study the characteristics of pls regression. Making two matrices, E = X and F = Y, is the initial step. After that, these matrices are normalized and column centered. These matrices' sum of squares are represented by the symbols SSX and SSY. The trajectory u is started with two random values before the iteration process begins.

Regression is clearly connected to both multiple factor analysis and canonical correlation. Tenenhaus (1998) & Pag`es & Tenenhaus (2001) go into great length on these relationships. The primary unique feature of pls regression is that, in contrast to these other methods, it maintains the asymmetry of the connection between predictors and dependent variables.[10]

Formula

The basic formula for multivariate PLS is as follows:

$$X = TP^t + E$$
$$Y = UQ^t + F$$

Where, X = n×m matrix of predictors
Y is an n×p matrix of responses
T and U = n×l matrices projections of X and Y respectively
P and Q = m×l and p×l orthogonal loading matrices and matrices respectively
E and F = error terms
The decays of X & Y are made with respect to maximise the covariance between T & U.[11]

For the better understanding of the PLSR, an exemplar structure regarding PLSR graph is shown in the figure 3 given below.

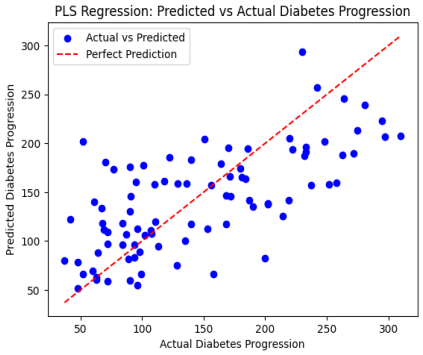

Figure 3: PLS Graph.[12]

b. Non Linear

When the data do not show a straight line relationship between variables, nonlinear regression algorithms are statistical or machine learning models. It fits the data to a curve or complicated function.

i. LS – SVM (Least Squares and Support Vector Machine)

- LS-SVM for classification and link with kernel Fisher discriminant analysis
- Links to normalization networks and Gaussian processes
- Bayesian inference for LS-SVMs
- Scatiness and robustness
- Big scale methods: Fixed Size LS-SVM Extensions to committee networks
- New formulations to kernel PCA, kernel CCA, kernel PLS
- Extensions of LS-SVM to recurrent networks and optimal control.[13]

Formula

$$y(x) = w^T \phi(x) + b$$

Where, $y(x)$ = predicted response
w = weight vector
(x) = non linear mapping of the input data into a higher dimensional feature space
b = bias term.[14]

For the better understanding of the LS-SVM, an example structure of the LS-SVM graph is shown in the figure 4 as given below.

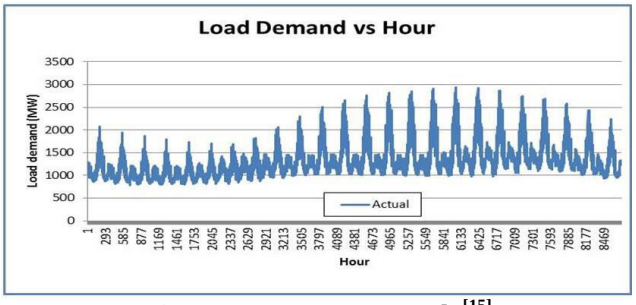

Figure 4: LS – SVM Graph.[15]

ii. ANN (Artificial Neural Network)

Computer programs also called Artificial Neural Networks (ANNs) are build to mimic how the human brain processes information. ANNs implement artificial neurons to analyse data, detect patterns, and produce prognostications, just as the brain utilises neurons to operate particulars and build decisions. These networks are prepared by layers of connected neurons that cooperate to track down solutions to challenging issues. The important premise is that, similar to how our brains learn from experience, ANNs may "grasp" from the information they produce. They are employed in many different applications, such as picture recognition and customized suggestions. We shall grasp more about ANNs, their operation, and other basic concepts in this study.

ANNs operate by using a procedure known as training to find patterns in data. By comparing its predictions with the actual outcomes, the network makes adjustments during training to increase its accuracy.

Let's examine how learning occurs:

- **Input Layer:** The input layer is used to feed data into the network, including text, images, and numbers.
- **Hidden Layers:** Every neuron in the hidden layers processes the information and transmits the outcome to the layer afterward. At every layer, the data is abstracted and modified.
- **Output Layer:** The network provides its final prediction, such as identifying an image as a dog or a cat, after going through each layer.

The weights between neurons are adjusted by the backpropagation process. The weights are adjusted when the network makes a mistake in order to minimize the error and enhance the subsequent prediction.[16]

For the better understanding of the ANN, an example structure of working of neural networks are shown in the figure 5 as given below.

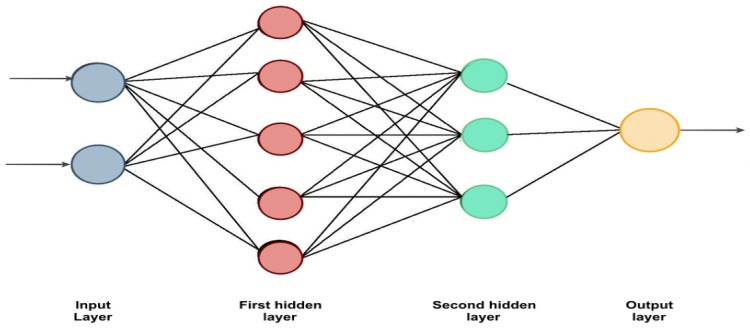

Figure 5: ANN Schematic Diagram.[17]

2. Classification Algorithms

With the aim of forecast the class of new, unseen data, classification algorithms learn from training data. They are divided into two categories: monitored and unsupervised.

a. Supervised

One kind of machine learning method used to categorize data into pre-established groups or classes is supervised classification.

In this case, the input data and the right output are known since the model is trained on a labelled data set. During training, the algorithm learns the relationship between inputs and outputs, which enables it to predict the class of fresh, untested data.

i. LDA (Linear Discriminant Analysis)

It is used to determine which linear feature combination best divides two or more classes. It seeks to reduce variation within each class and increase the distance between classes.

A supervised learning technique called LDA looks for a linear feature combination to create a decision boundary that successfully divides two or more classes in a dataset. Dimensionality reduction and linear classification are the two main steps. LDA minimizes variation within each class and maximizes the separation between classes by projecting high-dimensional data onto a lower-dimensional space. It can operate in the manner described below:

1. Manages the classification of several classes.
2. Stable model parameter estimation in contrast to logistic regression.
3. Like PCA, useful for dimensionality reduction in data pre-processing.[18]

For the better understanding of the LDA, an example structure of the LDA graph is shown in the figure 6 as given below.

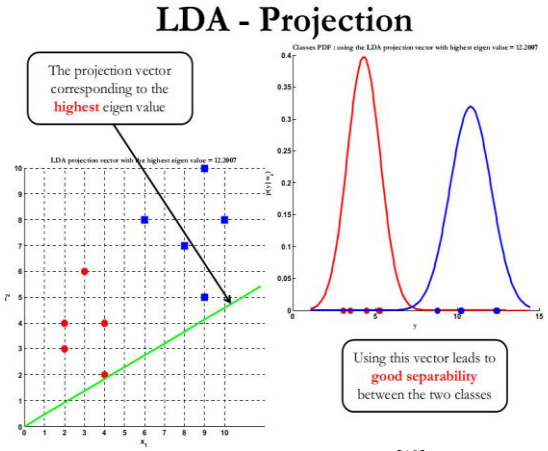

Figure 6: LDA Graph.[19]

ii. PLS-DA (Partial Least Squares-Discriminant Analysis)

Initially developed for regression applications, the partial least squares (PLS) algorithm later developed into the popular classification technique known as PLS-discriminant analysis (PLS-DA). The PLS-DA technique has been applied in practice for both discriminative variable selection and descriptive and predictive modeling. Here, we focus mostly on predictive modeling. In theory, PLS-DA is particularly useful for modeling high-dimensional (HD) data since it integrates discriminant analysis and dimensionality reduction into a single algorithm. Furthermore, PLS-DA is more adaptable than other discriminant algorithms like Fisher's linear discriminant analysis (LDA) since it does not presuppose that the data fit a specific distribution. Forensic examination, banking, health evaluation, food profiling, metabolic profiling, and topsoil research are just a few of the many domains in which PLS-DA modeling is used. According to Google Scholar, a formal explanation of the technique that was published in 2003 has actually been mentioned more than 1700 times since then. In actuality, however, many users still do not understand the fundamentals of building a valid and trustworthy PLS-DA model.

Even though there are many in-depth explanations of the PLS-DA algorithm, it is worthwhile to briefly summarize its main points, particularly for younger researchers who may not be familiar with it prior to reading this publication. PLS component construction, also known as dimension reduction, and prediction model development, also known as discriminant analysis, are the two primary processes that comprise PLS-DA predictive modelling In the past, PLS was suggested to deal with continuous variables, or regression tasks. However, the output variables in a classification problem will always be categorical. Therefore, recoding the categorical variables into continuous variables is the first stage in PLS-DA modeling. IR spectra datasets from two (binary, G = 2) and three (multi-class, G > 2) distinct pen sources, such as PILOT, STABILO, and ZEBRA, are examples.

It shows how the categorical variables (pen names) were recoded into continuous variables (dummy codes +1 and 0). Y is recoded to only include two numbers in a binary classification task. "Out-group" and "in-group" are typically denoted by the numbers 0 and +1, respectively. The symbol −1 is occasionally used to indicate "out-group." However, in a multi-class situation, y would have been changed to a dimwit Y. In actual use, the dimwit variables y and Y are used as output variables by the PLS1-DA and PLS2-DA algorithms, respectively.[20]

For the better understanding of PLS-DA, an example structure of PLS-DA graph is shown in the figure 7 as given below.

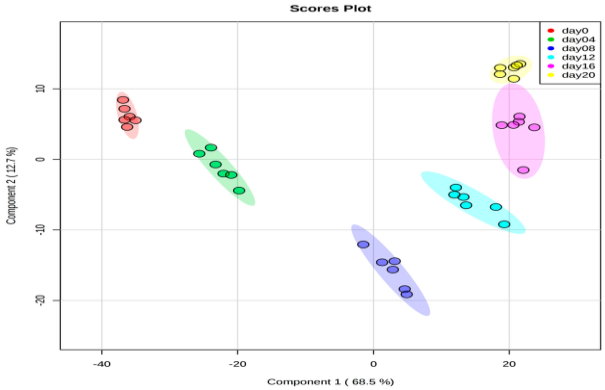

Figure 7: PLS – DA Graph.[21]

iii. KNN (K-Nearest Neighbours)

Data points are categorized according to the majority class of their K closest neighbors in the feature space. Although it can be applied to regression tasks, K-Nearest Neighbors (KNN) is a supervised machine learning technique that is mostly used for classification. It produces predictions based on the average value (for regression) or the majority class (for classification) after locating the "k" nearest data points (neighbors) to a given input. KNN is an instance-based and non-parametric learning technique since it does not assume anything about the distribution of the underlying data.

Because K-Nearest Neighbors maintains the complete dataset and only performs calculations during classification, it is also known as a lazy learner algorithm because it does not immediately learn from the training set.

Because the majority of its nearest neighbors are blue squares, the new location is categorized as Category 2. Based on the majority of neighboring points, KNN determines the category. The picture illustrates how KNN uses its closest neighbors to forecast a new data point's category.

- Category 1 is represented by red diamonds, and Category 2 is represented by blue squares.
- The nearest neighbors (circled dots) of the new data point are examined.
- KNN predicts that the new data point falls into Category 2 since most of its nearest neighbors are blue squares (Category 2).

KNN makes predictions by utilizing majority vote and proximity.

The number k in the k-Nearest Neighbors algorithm simply indicates how many neighboring points or neighbors the algorithm should consider before making a conclusion.[22]

For the better understanding of the KNN, an example structure of the KNN graph is shown in the figure 8 as given below.

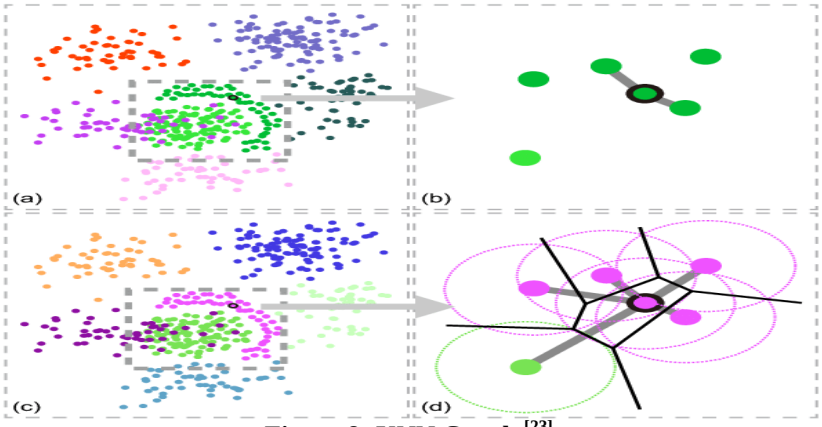

Figure 8: KNN Graph.[23]

iv. SIMCA (Soft Independent Modelling of Class Analogy)

One well-liked technique for creating authentication models is Soft Independent Modelling of Class Analogies (SIMCA). It can ascertain whether a fresh sample is consistent with a training set of recognized authentic class samples, just like Soft PLS-DA. SIMCA models, in contrast to PLS-DA, are trained on a single class and provide a binary True/False prediction on the consistency of a new sample with previously observed ones. Since World first introduced the approach in the late 1970s, it has undergone numerous changes. We will examine two distinct methods, one more traditional and the other more contemporary.

Additionally, there are variations in how each model might be optimized. "Compliant" models are trained utilizing different instances to achieve an overall optimal balance between specificity and sensitivity, whereas "rigorous" models use only examples of the target class and are designed to reach a given sensitivity (specificity cannot be measured). Even though the latter frequently seem to perform "better," the findings are skewed because of the training options, which are hard to measure.[24]

For the better understanding of the SIMCA, an example structure of the SIMCA graph is shown in the figure 9 as given below.

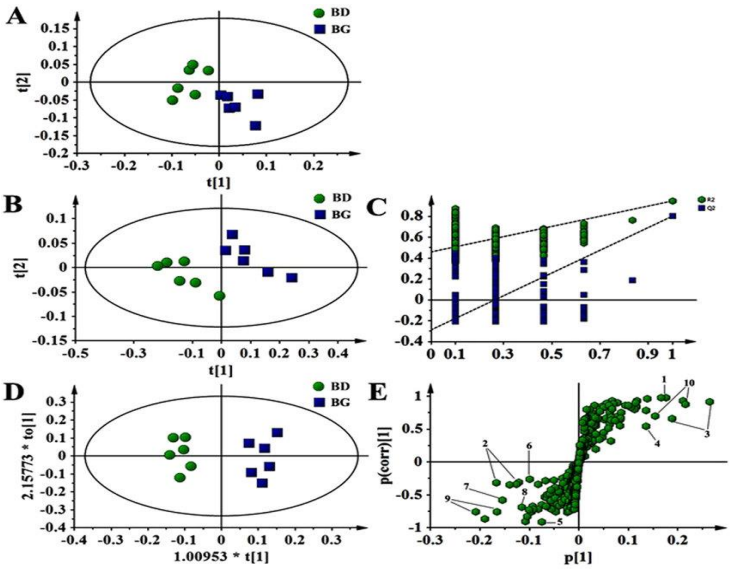

Figure 9: SIMCA Graph.[25]

v. SVM (Support Vector Machine)

Support vector machines (SVMs) are classification techniques that convert the learning algorithm to a high-dimensional feature space for efficient data categorization by optimally separating two groups using support vectors. It is used in many different fields, including pattern recognition and functional estimation, and can manage both linear and non-linear data-set segregation.

In the realm of ecology, SVMs are relatively new. Feature identification and categorization issues in remote observation are where it is most frequently used. Perhaps as a result of the quick advancement of deep information technologies and the associated lag in the enhancement of analytical techniques, researchers utilizing remote observation in eco-friendly and environmental utilization adopted SVMs earlier than other sectors. SVMs are used in fish age estimate, plant species and disease distribution mapping, and species identification. In fact, SVMs have a great possibility to address the ensuing challenges in data analysis whenever high-dimensional data are present together with a corresponding lack of understanding of the fundamental distribution. Apart from the fact that the information are independent and similarly dispersed—a problem that is increasingly recognized in contemporary science and statistics—SVMs also perform well in more manageable data groups of smaller dimensions and always have the advantage of not being restricted by distributional assumptions.[26]

For the better understanding of the SVM, an example structure of the SVM graph is shown in the figure 10 as given below.

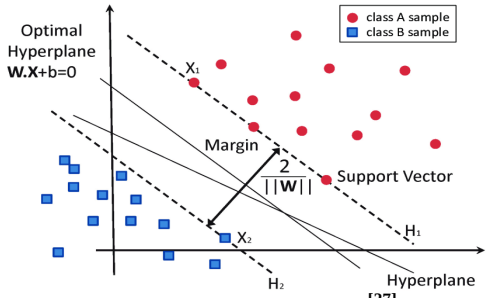

Figure 10: SVM Graph.[27]

b. Unsupervised

Another name for it is clustering algorithms. Without prior knowledge of class labels, it automatically categories data according to the natural relationships between samples.

i. PCA (Principal Component Analysis)

The process of creating a modern collection of non-associated factors from a set of potentially associated factors is known as principal component analysis. The newly generated factors are located in a modern correlated system, with the largest variation represented by the estimated information in the first collaborate, the second largest variation represented by the projected data in the second collaborate, and so on. The principal components (PC) are the newly created collaborates. In addition to altering the authentic data, the PCA offers two factors that aid in the chemometric analysis's interpretation: Eigen vectors and Eigen values. The Eigen vectors, also known as loading vectors, provide information about the weighting function used to get the PCA scores, while the Eigen values provide information about the variation in each PCA band. Although PCs with larger variance are chosen, the number of principle components acquired is typically equal to the number of original structures. When each new principle component is added, the condition of extraneous must be met with the left over principal components. Retaining the initial

principle components reduces the structurality of the data, which facilitates data visualization. Similarly, a two-dimensional scatter plot can be used to examine the data when the first and second principle components are kept. Therefore, before developing predictive models, PCA can be utilized as a resource tool for exploratory data research. The main benefit of PCA is that it protects the model from the curse of structurality by converting a large-structural space into a small-structural one while preserving variance as much as feasible. PCA merely uses the values of the learning features to make its projections; it need not require a actual data.

Researchers have only used the first five loading coefficients or score plots in their PCA investigations on hyperspectral data. These first five score photos contain every pertinent and helpful feature of interest. Furthermore, it has been claimed that the ML classifier's classification accuracy has been greatly enhanced by the use of PCA in the majority of situations.[28]

New variables known as principal components are created by combining or mixing the original variables in a linear fashion. These combinations are made so that the majority of the information included in the original variables is compressed or squeezed into the first components and the new variables are uncorrelated. The concept is that 10-dimensional data yields 10 principal components; however, PCA attempts to include as much information as possible in the first component, then as much information as possible in the second, and so on, until something is obtained.

Because the principle components are created as linear combinations of the original variables, it is crucial to understand that they are less interpretable and have no true significance.

In terms of geometry, principle components are the lines that encapsulate the majority of the data's information, or the directions of the data that explain the greatest amount of variance. Here, variance and information are related in that a line's dispersion of data points increases with the variation it carries, and a line's dispersion increases with the amount of information it contains. To put it simply, consider main components as additional axes that offer the optimum perspective for viewing and analyzing the data, making the contrasts between the observations more apparent.[29]

For the better understanding of the PCA, an example structure of the PCA graph is shown in the figure 11 as given below.

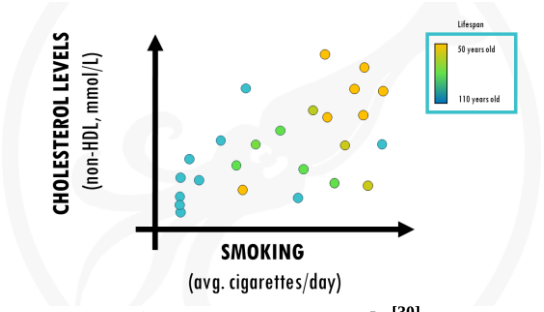

Figure 11: PCA Graph.[30]

ii. HCA (Hierarchical Cluster Analysis)

By creating a hierarchy (tree-like structure), hierarchical clustering is an unsupervised learning technique that groups related data points into clusters. Hierarchical clustering does not involve predetermining the number of clusters, in contrast to flat clustering such as k-means.

A dendrogram, which aids in understanding data similarity, is frequently used to visualize the process.

For clusters, a dendrogram is comparable to a family tree. It illustrates how separate data points or sets of data combine. Each data point is displayed as a separate group at the bottom, and comparable groups are joined as we go up. The groups are more comparable when the merge point is lower. It makes it easier for us to see how things are organized step by step.

- The closest points are combined into a single group as we ascend.
- The points are gradually combined depending on similarity, as indicated by the lines linking them.
- The similarity between the points is shown by the height at which they are joined; the shorter the line, the more similar they are.

There are two kinds of hierarchical testing. They are

1. **Agglomerative analysis:** Begin by treating each data point as a cluster.
2. **Divisive Testing:** All data points are initially grouped into a single, large cluster.[31]

For the better understanding of HCA, an example structure of the HCA graph is shown in the figure 12 as given below.

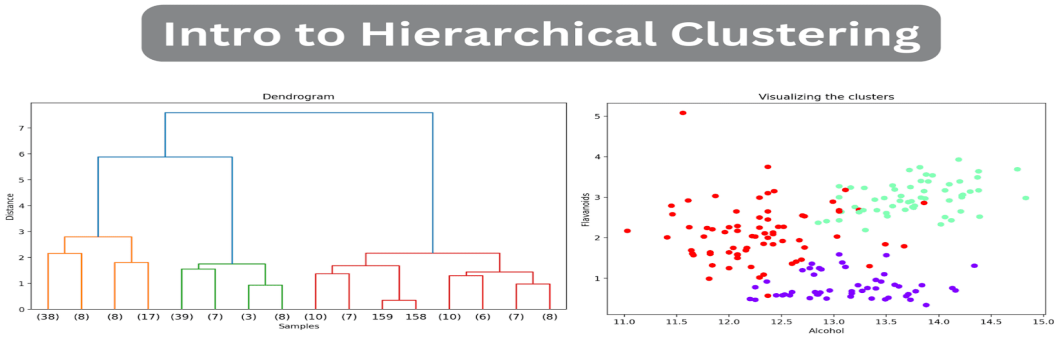

Figure 12: HCA Graph.[32]

Applications of Chemometrics

- Local Principal Component Analysis (PCA) has been used to evaluate non-synchronized batch process runs; two methods for defining specification regions for raw industrial materials have been investigated; and near-infrared (NIR) and engineering sensors have been combined to create MSPC control charts for monitoring polymerization reactions.[33]
- Pharmacology, food research, medicine, and bioinformatics all benefit from the use of PLS in chemometrics. Nonlinear calibrating techniques like nonlinear partial least squares (NPLS), locally weighted regression (LWR), alternating conditional expectations (ACE), and artificial neural networks (ANN) are useful in situations where PLS tends to produce substantial prediction errors.[34]
- Because chemometrics offers advanced techniques for obtaining, evaluating, and interpreting chemical data, it is crucial to analytical chemistry.
- This field integrates statistical and mathematical methods with chemical research to optimize analytical processes, improve data dependability, and extract insights. This is particularly crucial in sectors like medicines, ecological surveillance, and food safety where accurate data interpretation is essential.
- Chemometrics is crucial to modern analytical chemistry because it offers tools and techniques that increase data analysis capabilities, promote creative research, and strengthen quality control.[35]
- Although quantitative multivariate models might be difficult to implement across different locations due to equipment changes, chemometric approaches are essential for simulating spectroscopic characteristics of complicated mixtures.[36]

REFERENCES

1. Drashti N. Bhalodia, Dharmendrasinh A. Baria. A Review on Chemometrics in Pharmaceutical Analysis. International Journal of Pharmaceutical Sciences, 2024; 2(9): 58-78.
2. Chemometrics. The free Encyclopedia [Internet], 2025.
3. Karoly Vekey, Andras Telekes, Akos Vertes. Medical Appications of mass spectrometry. Elsevier science; 2008.
4. Sukhwinder Singh, Hanan Shakeel, Rakesh Sharma. Overview of chemometrics in forensic toxicology. Egyptian Journal of Forensic Sciences, 2025; 13(53).
5. Sebastian Taylor. Multiple Linear Regression. Corporate Finance Institute, 2025.
6. Rebecca Bevans. Multiple Linear Regression: A Quick Guide (Examples) Scribbr, 2023.
7. Kevin Dunn. Process Improvement Using Data (Principal Component Regression). Learnche.org, 2025.
8. Massart, D. L., Vandeginste, B. G. M., Buydens, L. M. C., De Jong, S., Lewi, P. J., & Smeyers-Verbeke, J. Handbook of Chemometrics and Qualimetrics: Part A. Elsevier, 1997.
9. Jim Frost, Principal component analysis: definition, examples & steps (Internet). Statistics by Jim, 2024.
10. Herve Abdi. Partial Least Squares (PLS) Regression. Encyclopedia of Research design, 2010.
11. Ali Sharifi. Partial Least Squares-Regression (PLS-Regression) In Chemometrics. (Internet), 2018.
12. Geeks for Geeks. Partial Least Regression (PLS Regression) using Sklearn [Internet], 2023.
13. Suykens JAK, Vandewalle J. Least Squares Support Vector machines [Internet]. Leuven (Belgium): KU Leuven, 1999.
14. Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. Least Squares Support Vector Machines. World Scientific Publishing, 2002.
15. Reddy SS, Panigrahi BK. Actual load demand for testing data [Internet], 2016.
16. Geeks for Geeks. Artificial Neural Networks and its Applications [Internet], 2023.
17. Mariam AL Harrach. Modeling of the sEMG/Force relationship by data analysis of high resolution sensor network. Research Gate, 2016.
18. Prasan N H. Linear Discriminant Analysis (LDA) in Classification. Medium (Internet), 2024.
19. Plot linear discriminant analysis in R. Stack Over flow, 2013.
20. Loong Chuen Lee, Choong-Yeun Liong, abdul Aziz Jemain. Partial least squares-discriminant analysis

(PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. Analyst, 2018; (15).
21. Florian Miehle. Gabriele Moller, Alexander Cecil, jutta Lintelmann, Martin Wabistch, Janina Tokarz, Jerzy Adamski, Mark Haid. Lipidomic Phenotyping Reveals Extensive Lipid Remodeling during Adipogenesis in Human Adipocytes. Researcg Gate, 2020; 10(6): 217.
22. K Karthik. K-Nearest Neighbor (KNN) Algorithm. Geeks for Geeks, 2025.
23. Kecheng Lu, Mi Feng, Xin Chen. Palettailor: Discriminable Colorization for Categorical Data. Research gate, 2020.
24. Nathan A. Mahynski. Soft Independent Modeling of Class Analogies (SIMCA). Read the Docs [Internet], 2023.
25. Xiaojian Yang, Fugui Yin, Yuhui Yang, Dion Lepp, Hai Yu, Zheng Ruan, Chengbo Yang, Yulong Yin, Yongging Hou, Steve Leeson. Dietary butyrate glycerides modulate intestinal microbiota composition and serum metabolites in broilers. Researcg Gate, 2018.
26. Hannah Kerner, Joseph campbell, Mark Strickland. Chapter 1 – Introduction to machine learning. Machine Learning for Planetary science. Elsevier, 2022; 1-24.
27. Esperanza Garcia-Gonzalo, Zulima Fernandez,-Muniz, Paulino Jose Garcia Nieto, Antonio Bernardo Sanchez, Marta Menendez. Hard-Rock stability analysis for span design in entry-type excavations with learning classifiers. Research Gate, 2016; 9(7): 531.
28. Dhritiman Saha, Annamalai Manickavasagan. Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review. Current research in food science, 2021; 4: 28-44.
29. Zkaiara Zadi written, upd by Brennan Whitfield, Review by Sadrach Pierre. Prinicipal Component Analysis. Built in [Internet], 2025.
30. Laura. A simple explanatiom of PCA. Biostatsquid [Internet], 2025.
31. Hierarachial Clustering. Geeksforgeeks [Internet], 2025.
32. Bala Priya C, Unveiling Hidden Patterns: An Introduction to Hierarchial Clustering. KD nuggets, 2023.
33. Alessandra Biancolillo, Angelo Antonio D'Archivio, Federico Marini, Raffele Vitale. Editorial: Novel applications of chemometrics in analytical chemistry and chemical process industry. National institutes of Health, 2022; 10: 926309.
34. Inderbir Singh, Prateek Juneja, Birendar Kaur, Pradeep Kumar. Pharmaceutical Applications of Chemometric Techniques. Research Gate [Internet], 2013.
35. Projesh saha, Bibhas Pandit, Soma Pramanik, Bhupendra Shrestha. A comprehensive review on the applicatons of chemometrics in analytical chemistry. Journal of Applied Pharmaceutical research, 2025; 13(3): 1-16.
36. Caroline Hroncich. Analytica 2024: The Latest Applications in Chemometrics. Spectroscopy; 2024.