Chapter 30
Visualization Tools for Big Data Analytics in Quantitative Chemical Analysis:
A Tutorial in Chemometrics

Gerard G. Dumancas
Louisiana State University – Alexandria, USA

Lakshmi Chockalingam Kasi Viswanath
Oklahoma Baptist University, USA

Ghalib A. Bello
Icahn School of Medicine at Mount Sinai, USA

Casey O'Neal Orndorff
Louisiana State University – Alexandria, USA

Jeff Hughes
RMIT University, Australia

Glenda Fe Dumancas
Louisiana State University – Alexandria, USA

Renita Murimi
Oklahoma Baptist University, USA

Jacy D. O'Dell
Oklahoma Baptist University, USA

ABSTRACT

Modern instruments have the capacity to generate and store enormous volumes of data and the challenges involved in processing, analyzing and visualizing this data are well recognized. The field of Chemometrics (a subspecialty of Analytical Chemistry) grew out of efforts to develop a toolbox of statistical and computer applications for data processing and analysis. This chapter will discuss key concepts of Big Data Analytics within the context of Analytical Chemistry. The chapter will devote particular emphasis on preprocessing techniques, statistical and Machine Learning methodology for data mining and analysis, tools for big data visualization and state-of-the-art applications for data storage. Various statistical techniques used for the analysis of Big Data in Chemometrics are introduced. This chapter also gives an overview of computational tools for Big Data Analytics for Analytical Chemistry. The chapter concludes with the discussion of latest platforms and programming tools for Big Data storage like Hadoop, Apache Hive, Spark, Google Bigtable, and more.

DOI: 10.4018/978-1-5225-3142-5.ch030

ANALYTICAL CHEMISTRY AND CHEMOMETRICS

Over the years, various Chemometric tools have emerged and have been utilized as data evaluation instruments generated by various hyphenated analytical techniques including their application since its advent today (Kumar, Bansal, Sarma & Rawal, 2014). Although its primary applications are geared toward Multicomponent Analysis, its applications have even been extended to the area of genetic epidemiology and Bioinformatics in the recent years (Dumancas, 2012; Dumancas et. al., 2014; Dumancas et. al., 2015).

The advances that are now visible in Process Analytical Technology (PAT) in Chemometrics can be attributed to the rapid development of both analytical instrumentation and mathematical methods involved in multivariate data analysis (Bogomolov, 2011; Dubrovkin, 2014; Kessler, 2013; Pomerantsev & Rodionova, 2012). Specifically, the rapid growth of a wide multitude of novel analytical methods and the continuous expansion in the area of their applications are the two driving forces that led to the success of PAT (Dubrovkin, 2014).

With the vast array of information emanating from various analytical instruments comes the challenge of processing these data in a rapid fashion. Thus, the process of Data Fusion, a subclass of Chemometrics is now considered an important topic (Esteban et. al., 2005; Ovalles & Rechsteiner, Jr., 2015). Data Fusion simply refers to the integration of data and knowledge from several sources (e.g. analytical instruments) (Castanedo, 2013). Many other definitions for data fusion exist in the literature. It is defined by the Joint Directors of Laboratories (JDL) as a "multi-level, multifaceted process handling the automatic detection, association, correlation, estimation, and combination of data and information from several sources" (Steinberg et. al., 1999). The corresponding informational models from data fusion should simulate extremely complex problems by fitting to the massive amount of empirical semi-structured and unstructured data (Isaeva et. al., 2012). Consequently, the algorithmic support and the interface of a computerized analytical system (often with limited computer resources) should be adjustable to systems with features of new types. Such challenge arising from analytical information management led to several perspective solutions such as the concept of Cloud Computing all of which is part of the development of "Big Data Approach" (BDA) (Dubrovkin, 2014).

In this chapter, the major aspects of Big Data utilization and processing in Analytical Chemistry (Chemometrics) will be discussed. Specifically, some commonly used algorithmic and instrumental techniques and aspects of computerized analytical systems will be discussed.

APPLICATIONS OF CHEMOMETRICS

Chemometrics is a fast spreading area which has many avenues of applications in both descriptive and predictive problems in experimental life sciences especially in Chemistry. It is considered to be a highly interfacial discipline employing Multivariate Statistics, Computer Science and Applied Mathematics using methods employed in core data analytics with the ultimate goal of addressing problems in Biochemistry, Medicine, Chemistry, Chemical Engineering and Biology (Khanmohammadi, 2014).

The biological and medical applications of Chemometrics encompass a wide area of expertise. Support Vector Machines (SVMs), Partial Least Squares Discriminant Analysis (PLS-DA) are widely used

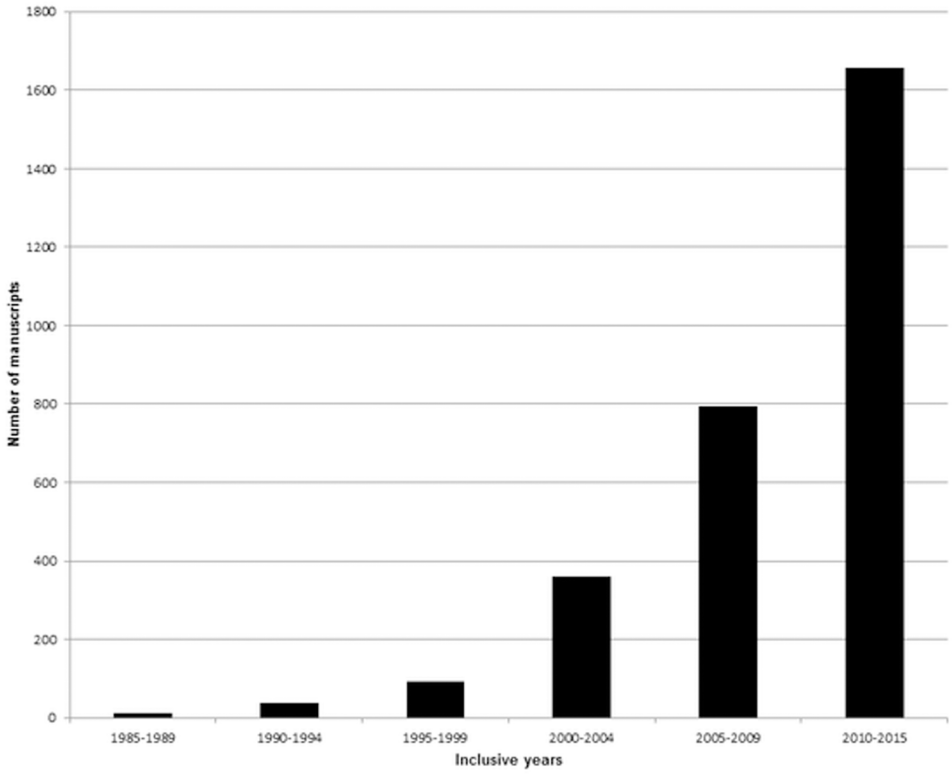
techniques for classification purposes involving microorganisms, medical diagnosis using spectroscopy and metabolomics using Coupled Chromatography and Nuclear Magnetic Resonance Spectrometry (Brereton, 2007).

Other widely used applications of Chemometrics is in food science. Specifically near Infrared Spectroscopy (Near IR) is used for calibration, classification and exploratory purposes. The ultimate goal is for sensory analysis which links composition to products using sensory panels and PCA. (Brereton, 2007)

Over the years, industries have also been employing Chemometric statistical designs for improving the performance of synthetic reactions. Specifically, factors are screened that are known to influence the performance of a reaction as well as implementing optimization of variables (Brereton, 2007).

Chemometrics and the methods involving in it are versatile and there is a high level of abstraction involved in this field. This is due to the fact that the field is characterized by the use of statistical and mathematical methods—the multivariate methods. The algorithms and the techniques used in the processing and evaluation of data can be implemented to various fields including Pharmacy, Food Control, Medicine and environmental monitoring among others (Matero, 2010; Mocák, 2012; Singh et. al., 2013). The number of manuscripts published involving Chemometrics has ever since increased significantly over the past four decades (Refer to Figure 1).

Figure 1. Number of manuscripts published over the years using the keyword "chemometric" in PubMed (Designed by authors, 2016)



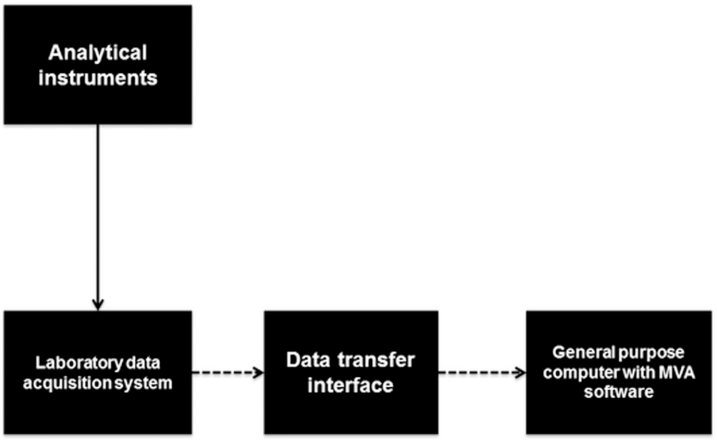
BIG DATA ANALYSIS IN ANALYTICAL CHEMISTRY

Definition of Big Data in the Context of Analytical Chemistry

In order to be aware of the features of Big Data within the context of Analytical Chemistry, the major focal points or definition of Big Data should first be discussed. Modern analytical laboratories can routinely generate large volumes of detailed data for complex samples. To acquire the most information from the data, systematic organization and reduction is needed. These data handling procedures consist of data collection, organization and creation of databases. Automatic instrumentation and data handling procedures manage these tasks of data handling. These automated procedures offer the advantages of speed and accuracy and therefore greatly reduce the difficulty for routine applications of Chemometrics. Data collection including sample extraction and analysis typically require considerable amount of time. Manual organization and creation of these databases greatly increase the amount of time and effort required to carry out the preparation for data analysis (Burgard & Kuznicki, 1990).

There are two different approaches to data handling in Chemometrics. Both of these assume the use of a computerized data acquisition as part of the entire instrumental system. The first one is a typical data acquisition and analysis configuration (Refer Figure 2). It consists of a stand-alone data instrumental acquisition system that requires a data transfer interface for serially uploading the data for each sample. This type of arrangement is relatively easy to implement but requires much user interaction to move data from one system to another and to organize the database for analysis. Many times, the Chemist or Laboratory Technician assumes the roles of the data transfer interface who is responsible for manually organizing, tabulating and entering the data into the computer where data analysis will occur. A better arrangement is when the data transfer interface is electronic and information is uploaded via a standard protocol to a host computer. Personal computers (PCs) and PC-based instrumental data systems are useful for this step since commercial database managers make it easy to manipulate large quantities of data. It is even possible to use the PC for the final data analysis as multivariate software packages for data analysis are now available for PCs (Burgard & Kuznicki, 1990).

Figure 2. Typical data acquisition and analysis configuration (Burgard & Kuznicki, 1990)

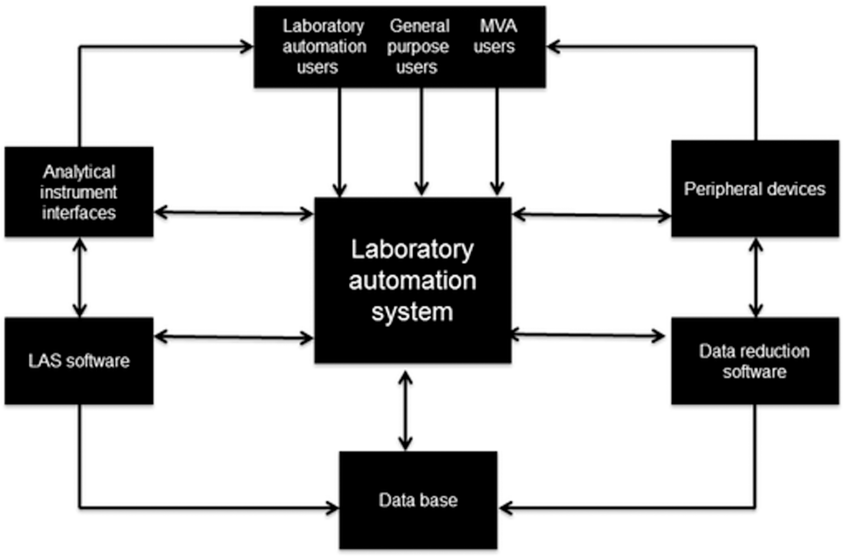


The second type of approach involves a completely integrated system where all functions are performed in the same computer (Refer Figure 3). Such an approach offers the advantages mentioned in the first approach to data handling in addition to supporting multiple instruments and users simultaneously. Multi-user Laboratory Automation Systems (LASs) can be configured for such operation. The vendor software provides support for the data acquisition and possibly the data management systems Laboratory Information Management Systems (LIMS). General purpose statistical software packages are available for various laboratory computer systems. Data analysis packages have been developed privately and integrated into the laboratory computer systems. However, the major disadvantage of this approach is the need for custom development and/or installation of the data reduction software (Burgard & Kuznicki, 1990).

Data Preprocessing

Chemometric methods are influenced by methods used for data processing. Data preprocessing is considered to be the second most important step in a Chemometric study after study definition and data collection (Burgard & Kuznicki, 1990). Pretreatments in Chemometrics are applied for various reasons, to overcome such problems as scaling differences between variables, background errors, noisy data, etc. It may also be necessary to reduce the total amount of data. Preprocessing simply means to put the data into a meaningful form for further comparisons; that is, the conversion of raw data to units or scales that allow direct comparison of measurements for different samples. The technique of preprocessing is often accomplished in three simple steps. The first step converts the data to units appropriate for the comparisons to be made. The second step involves the organization and creation of a database. The third and last step involves mathematically conditioning the data in preparation for the actual data analysis (Burgard & Kuznicki, 1990).

Figure 3. Ideal data acquisition and analysis environment (Burgard & Kuznicki, 1990)



In Pattern Recognition, the commonly used methods in data preprocessing include:

1. No pretreatment
2. Mean-Centering where the mean of each column is subtracted from each entry in the column (usually a minimum treatment for Principal Component Analysis (PCA))
3. Standardization: as well as Mean Centering wherein each entry in the column is divided by the column Standard Deviation. Thus, the mean of each column is zero and Standard Deviation (SD) = 1 (also often applied prior to PCA and methods which are not scale invariant such as Soft Independent Modeling by Class Analogy (SIMCA). This method is also called Autoscaling.
4. Row scaling, so the rows sum to a constant total (usually 1 or 100). Row scaling is useful where the absolute concentration of a sample cannot be controlled.

Other methods to reduce scaling and other problems include log transforms and weighting the variables. If the amount of data is too large to be handled in the software, it can be reduced by doing PCA first and using the principal components in place of the original data set. All these methods affect the final result of pattern recognition, thus, it needs to be understood why a pretreatment should be necessarily applied prior to analysis.

In the aforementioned methods involved in data preprocessing, methods (a) and (c) are done automatically or are transparent to the user and are not considered to be a separate step in an analysis sequence. One good example is the conversion of chromatographic peak areas to parts per million which can be accomplished in most instrumental data systems. This might be considered as the first step in the preprocessing of chemical data. Centering and scaling procedures in the third choice (c) are often considered to be part of a standard data reduction procedure and can be performed automatically by most software packages. The second choice (b), on the other hand, is typically not available in commercial software packages and, thus must be performed manually or by user created software or sequences. This method involves aligning all the data for each sample to ensure that same data points represent the same variable for each sample and the entry of the data into the database should be in a form that can readily be accessed by using an analysis software (Burgard & Kuznicki, 1990).

DATA ANALYSIS

Regression-Based Methods

Partial Least Squares (PLS)

Partial Least Squares (PLS) regression was introduced in the early 1980s and since then has gained much popularity in Chemometrics (Helland, 2004). It is closely related to Principal Component Regression (PCR), another Chemometric multivariate technique. However PLS differs in that it uses the response information during the decomposition of the X data matrix. The main idea of PLS is to get as much response information as possible into the first few loading vectors. Unlike PCR, PLS is a one-stage process as PLS performs decomposition on both the X and Y matrices simultaneously. There is no separate regression step as in PCR. PLS, like PCR, can be performed when the predictors are highly correlated (collinear).

In literature, PLS is often introduced and explained as a numerical algorithm that maximizes an objective function under certain constraints. The objective function is the covariance between X and Y scores and the constraint is usually the orthogonality of the scores (Varmuza & Filzmoser, 2009).

There are two separate PLS algorithms – PLS1 and PLS2. PLS1, or one-block PLS, is performed when there is one y response. PLS with several responses is called two-block PLS or PLS2. Also the vectors generated by PLS more closely relate to the constituents of interest than those from PCA (Helland, 2004).

As with PCR, the key to PLS is the decision on how many 'significant' components to include and the optimum number of components that can be decided from $R_{ev}{}^2$ where $R_{ev}{}^2$ is R^2 for the cross-validated model. However, because PLS is a nonlinear technique, the cross-validated residuals must be calculated by the leave-one-out technique repeating the model calculation many times.

The raw data matrix, Z is initialized by carrying out a pretreatment step (usually Mean Centering or Standardization) to give X. The same pretreatment is carried out on the response vector, y to give c. There are several different algorithms used. This version of the PLS algorithm is a non-iterative version (Brereton, 2003). Below, the steps involved in the PLS1 technique are given:

The scores are then given by:

$$\frac{X * h}{\sqrt{\left(\sum h^2\right)}} \text{ (Brereton, 2003)} \tag{1}$$

The x loadings are given by:

$$\frac{t' * X}{\sqrt{\left(\sum t^2\right)}} \text{ (Brereton, 2003)} \tag{2}$$

The c loadings are given by:

$$\frac{c' * t}{\sqrt{\left(\sum t^2\right)}} \text{ (Brereton, 2003)} \tag{3}$$

The x residuals are computed as:

$$X_{resid} = X - t * p \text{ (Brereton, 2003)} \tag{4}$$

And the new response estimate is given by:

$$C_{new} = C + t * q \text{ (Brereton, 2003)} \tag{5}$$

To get new calculated values of y, the pretreatment must be reversed (e.g. if mean-centered, add the mean). The c residuals can also be determined using the equation below:

$C_{resod} = C - C_{new}$ (Brereton, 2003) (6)

This is PLS with one component. Further components can be included stepwise by replacing X and C in equations (1) to (3) by the residuals X_{resid} and C_{resid} and recalculating.

PLS2 regression is a variant of PLS that is a generalization to several dependent variables (Helland, 2004). In other words, this type of regression predicts simultaneously several dependent variables (Banks, House, McMorris, Arabie & Gaul, 2011). The algorithm is an extension to the PLS1 algorithm except that a concentration matrix C is used. If a mixture is being analyzed, for example, PLS1 can be applied to the concentration vector for each component, or PLS2 can be applied in one process using a concentration matrix C where each column is a concentration vector for each component.

Principal Component Regression (PCR)

In Ordinary Least Squares (OLS), the number of variables must be less than that of the number of samples (objects) and these variables cannot be highly correlated. If the variables are strongly correlated, this leads to a singular data matrix which cannot be inverted in the least squares process. To avoid the problem of collinearity, and hence the need to decide which variables to use, PCR or PLS can be used. By definition, the variables in PCR and PLS are orthogonal (uncorrelated), so collinearity is no longer an issue.

The PCR model is as follows:

$y = T * a - e$ (Brereton, 2003) (7)

where T is a matrix of the first 'm' principal component scores, a is a vector of coefficients and e is the error vector. This can be compared to the general linear model:

$Y = X * b - e$ (Brereton, 2003) (8)

where X is the original data matrix (usually auto-scaled i.e. subtract the mean and divide by the Standard Deviation of each column or mean-centered i.e. subtract the mean of each column)

The matrix T can be calculated by determining L, the set of eigenvectors of the matrix $X^T * X$

$T = X * L$ (Brereton, 2003) (9)

$y = T * a + e = [X * L] * a + e$ (Brereton, 2003) (10)

Hence, the coefficients for the PCR model are given by $b = L * a$ and a is determined from the scores using OLS as given by:

$a = (T^T * T)^{-1} * T^T * y$ (Brereton, 2003) (11)

This is essentially the method used to determine *b* in OLS. However, the matrix *T* has orthogonal columns so the problem of collinearity is avoided.

To calculate the predicted responses from the original (unscaled) data, the coefficients can be determined by dividing the elements of b by the Standard Deviation of the corresponding column of *X*.

If the whole score matrix is used, the results will be the same as OLS. However, the secret to successful PCR is using enough 'significant' eigenvectors in *L* (and corresponding columns of *T*) to get a successful model but screen out noise. If the eigenvalues (and corresponding eigenvectors) of $X^T * X$ are arranged in order of decreasing eigenvalues, then only the eigenvectors corresponding to 'significant' contribution are used in the matrix L.

Ridge Regression

Ridge Regression is a general term encompassing different forms of regression (linear, logistic, survival, etc.) that incorporates the 'ridge penalty'. The ridge penalty is known by many names, e.g. Tikhonov Regularization, Constrained Linear Inversion, etc. It first gained wide recognition through the landmark 1970 paper by Hoerl and Kennard (Hoerl & Kennard, 1970). It is ideal for use in models with multicollinearity and other ill-posed problems. Like other regularization techniques, it involves imposing a constraint on the parameters in a model in order to mitigate variance inflation. Say we have a linear model with *p* parameters, which we represent as a *p*-dimensional vector $\boldsymbol{\beta}$, then the ridge penalty effectively controls the L_2 norm of $\boldsymbol{\beta}$. In other words, it coerces the sum of squares of the parameters to fall below a particular value which is usually a tunable parameter:

$$\left\|\beta\right\|_2^2 \leq c^2 \text{ (Hoerl \& Kennard, 1970)} \tag{12}$$

The ridge penalty exploits the bias-variance tradeoff by increasing the bias of the parameter estimates in exchange for a reduction in their variance. The latter is particularly important in ill-posed problems (e.g. where strong correlations exist among independent/explanatory variables) which are often plagued by variance inflation. The strength of the penalty (regularization) can be tuned to suit the particular problem it is applied to.

Validating Regression Based Methods

The validity of the regression model needs to be tested to have confidence in its use to predict properties of new samples. The type of validation used depends on the number of samples in our training set (samples where there is an independent assessment of the property that is being determined in the modelled experiment).

1. **Cross-Validation (CV):** In full Cross-Validation (sometimes called 'leave-one-out' CV), an object is left out of the training set and then its property is determined based on a model established from the rest of the training set. Thus, this object does not form part of the process of finding the model. This is repeated for each object in the data set. If it is a good method, then the predicted properties will be close to the values determined independently. A superior method, if there are

sufficient objects or samples available in the data set, is to divide the set into a training set and a test set. Roughly a 2:1 training: test ratio is said to be the best. The model is established based on the training set only and is used to predict the properties for the test set.

Even better is to use a 'bootstrap' method where a test set is selected at random and a model is constructed from the remaining samples. The procedure is applied, the selected samples are replaced and then select a new test set. This is repeated many times (many thousands of resamplings are quite possible with modern computing methods).

Data Reduction and Pattern Recognition in Chemometrics

The ultimate goal of data reduction is the replacement of a large amount of measurements by a few characteristic numbers in which all relevant information has been preserved. Depending on the type of data measured and the type of information needed, there are a number of methods involved in data reduction. In most cases, the data reduction method can be obtained by fitting a model through the data points. Consequently, the obtained model is then used to describe the data instead of the data themselves. The process of fitting models is one of the principal cores used by chemometricians. Within the context of calibration, the model is usually a straight line. In multicomponent analysis, on the other hand, the model consists of a system of linear equations. Lastly in optimization, the model is a polynomial consisting of several independent variables (Deming, Michotte, Massart, Kaufman & Vandeginste, 1988).

Principal Component Analysis (PCA)

The ultimate goal of Principal Component Analysis (PCA) is to obtain a set of K variables and identify a smaller number of components that can be determined from the data while representing a large proportion of the variance in the data. This is simply accomplished by identifying relationships between the K variables and producing a set of K uncorrelated components (new variables). Each component is a function of the original variable. Such a function is called an eigenvector (O'Donoghue, 2013).

It is often useful to reduce the number of variables prior to exploratory or supervised data analysis (Brereton, 2009). Principal components (PCs) can be used for many different purposes and in addition to data visualization can also be used for data reduction. Given a matrix X, instead of utilizing the original J raw variables in such original matrix, 'A' orthogonal variables or PCs are used as represented by a scores matrix T as input to the classifier (Brereton, 2009).

PCs are often ordered according to their size or eigenvalues with PC1 being the largest and PCA the smallest. By this, it simply means that PC1 consists of scores that have the largest sum of squares or largest eigenvalue. PCA is known to be an unsupervised method of data reduction. This means that the calculation of PCs does not take group membership of samples into account. This creates advantages within the context of data reduction prior to model testing. For example, there is no risk of over-fitting if PCA is performed on the overall dataset including that of the testing and training sets together prior to classification (Brereton, 2009).

PCA is a *data reduction* method. Its aim is to reduce a large data set into a much smaller set but one which retains the essential information of the original set. For example, consider the measurement of the IR spectra of 100 samples from 400-4000 cm^{-1}. If, as commonly done, every 1 cm^{-1} is performed, then the final data set consists of 100 x 3600 or 36,000 points. However, much of this data is superfluous. Two

Table 1. Example data used for Principal Component Analysis (PCA)

Compound	Wavelength 1 (cm^{-1})	Wavelength 2 (cm^{-1})	Wavelength 3 (cm^{-1})	Wavelength 4 (cm^{-1})
	300	350	400	450
A	16	62	67	27
B	15	60	69	31
C	14	59	68	31
D	15	61	71	31
E	14	60	70	30
F	14	59	69	30
G	17	63	68	29
H	16	62	69	28
I	15	60	72	30
J	17	63	69	27
K	18	62	68	28
L	18	64	67	29
Mean	15.75	61.25	68.92	29.25
Standard deviation	1.485	1.658	1.505	1.485

(Brereton, 2009)

wavenumbers that are only $1cm^{-1}$ apart will contain essentially the same information. Thus, we could leave out a lot of the data, but which wavelengths to be omitted remains the challenge.

In this example, there are 4 variables $X_1 - X_4$ (the 4 wavelengths). The idea behind PCA is to find new variables $Z_1 - Z_4$ which are linear combinations of the original variables:

$$Z_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + a_{14}X_4 .... \; Z_4 = a_{41}X_1 + a_{42}X_2 + a_{43}X_3 + a_{44}X_4, \tag{13}$$

(Brereton, 2009). Of course, there is no reduction if there are still 4 variables but in PCA, the Z variables (called principal components) are ranked so that Z_1 contains the largest amount of information (i.e. accounts for the largest amount of variation in the original data set) and decreasing to Z_4. In this case, the first two PCs account for 88% of the total variation so the total data set can be represented adequately with only 2 variables. This isn't a big reduction here but in the IR case described above, with 3600 variables, it is often possible to describe the whole data set with only a few variables.

Another feature of PCA is that the coefficients a_{ij} are chosen so that the Z's are orthogonal (i.e. uncorrelated). The following are the steps in calculating the coefficients:

Step 1: Pretreatment

Let X be our original data set. Commonly in PCA there are 3 forms of pretreatment applied to X:

1. No pretreatment
2. Mean-Centered: The mean of each column is subtracted from each entry in the column
3. Standardization: as well as Mean Centering wherein each entry in the column is divided by the column Standard Deviation. Thus, the mean of each column is zero and Standard Deviation (s.d.) =1

Which one to apply? In spectroscopy, data is commonly mean-centered. If the data variables are parameters which differ in scale (e.g. pH and temperature), then the data should be standardized.

Step 2: Determine the Eigenvalues and Eigenvectors of the Pretreated Data Matrix

The PCA of X (correlation option) gives eigenvalues and eigenvectors of the correlation matrix

= eigenvalues, eigenvectors of $[W^T * W] / (n-1)$ where W is standardized matrix

$(x_{ij} - x_{av,i}) / s_i$

The PCA of X (covariance) is eigenvalues, eigenvectors of $[V^T * V] / (n-1)$ where *V* is a centered matrix $(x_{ij} - x_{av,i})$ The aim of PCA is to decompose the original (treated) matrix *W* or *V* as follows:

$W = T * L^t$ (Brereton, 2009) (14)

In the aforementioned equation, L is the loadings matrix and has columns which are the eigenvectors of W. The loadings relate the new (latent) variables to the original variables. These are the coefficients a_{ij} listed above. T is the scores matrix and shows how the objects (rows of X) relate to the latent variables. It is important to remember that loadings relate to variables (columns) and scores to objects (rows).

T can be determined by: $T = W * L$. There are several algorithms used to obtain T and L. One is called Single Value Decomposition which decomposes W (or V) as follows

$W = U * diag(\lambda) * L^t$ (Brereton, 2009) (15)

Thus, $T = U * diag(\lambda)$ and $diag(\lambda)$ is a square matrix with the eigenvalues along the diagonal and zeroes everywhere else.

The PCA analysis of the above data set gave the following output (Table 2).

Note that the first PC describes 72% of the total variability and PC1+ PC2 describes 88% of the variability (Refer Table 2). The scores give useful information on groupings of objects (this is called a score plot) (Refer Figure 4).

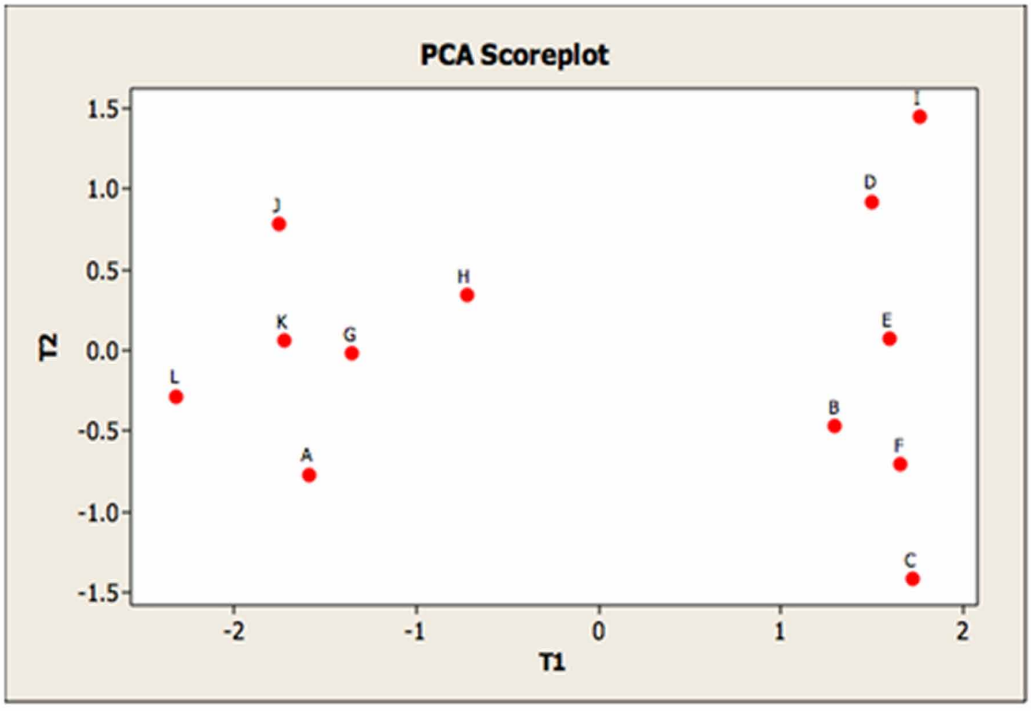
There are clearly two groups of objects. The separation is along the PC1 axis. Examining the loadings shows that the two higher wavelengths separate in a positive direction and the two lower ones in a negative direction. Closer examination of the original data should confirm that one group has slightly higher readings for the higher wavelengths and lower readings for the lower wavelengths.

Table 2. PCA analysis of data from Table 1

Eigen Analysis of the Correlation Matrix							
Eigenvalue	2.880	0.654	0.389	0.0844			
Proportion	0.720	0.161	0.097	0.021			
Cumulative %	72	88	98	100			
Loadings				Scores			
PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
-0.546	0.237	0.395	-0.699	-1.595	-0.766	-0.991	0.185
-0.546	0.298	0.324	0.712	1.291	-0.466	0.561	-0.128
0.400	0.912	-0.072	-0.043	1.722	-1.412	0.147	-0.058
0.493	-0.145	0.856	0.048	1.4929	0.926	0.660	0.243
				1.592	0.078	-0.330	0.280
				1.656	-0.708	-0.477	-0.120
				-1.362	-0.015	0.575	0.181
				-0.731	0.348	-0.511	0.160
				1.755	1.451	-0.160	-0.248

(Brereton, 2009)

Figure 4. PCA scoreplot of data from Table 1 (Brereton, 2009)



Clustering techniques have been employed in a wide range of disciplines. In Archaeology, clustering has been used to investigate the relationship between various types of artifacts. In psychiatry, the methods have been utilized to refine existing diagnostic categories. Further, in market research, clustering techniques have been employed to produce groups of consumers with different purchasing patterns (Everitt et. al., 2011).

Essential to the understanding of various clustering techniques is the correct identification of the number *k* of clusters that is somehow inherent in the data. The ultimate goal of cluster analysis is to find clusters where the objects within the clusters are as similar as possible and objects between different clusters are as dissimilar as possible. In order to assess the similarity and dissimilarity of objects within the clusters, a measure of 'homogeneity' and 'heterogeneity' between the clusters is defined. Homogeneity measures between clusters can be based on the maximum, minimum or average of the distances between all of a cluster or an average distance of the objects within a cluster to the cluster center (Varmuza & Filzmoser, 2009). One possible choice for a measure of homogeneity w_j within a cluster *j* is:

$$w_j = \sum_{i=1}^{n_1} \left\| x_i^{(j)} - c_j \right\|^2 \text{ (Varmuza \& Filzmoser, 2009)} \tag{16}$$

A measure of heterogeneity between two clusters, on the other hand, can be based on the maximum, minimum or average of all pairwise distances between the objects of the two clusters, or on the pairwise distances between the cluster centers (Varmuza & Filzmoser, 2009). Thus, the measure of heterogeneity B_{jl} between cluster j and l can be described as:

$$B_{jl} = \left\| c_j - c_l \right\|^2 \text{ (Varmuza \& Filzmoser, 2009)} \tag{17}$$

By combining the two aforementioned criteria, we come up with the validity measure as given by the equation below which depends on the chosen number of k clusters. In order to determine the number of clusters, a graph showing the number of clusters versus the validity measure is essential with a knee indicating the optimal number of clusters (Varmuza & Filzmoser, 2009). The results in a validity measure V(k) can be defined as:

$$V(k) = \frac{\sum_{i=1}^{k} w_j}{\sum_{i=1}^{k} B_{jl}} \text{ (Varmuza \& Filzmoser, 2009)} \tag{18}$$

A number of methods exist for Clustering. The most commonly used and simplest method is the k-Means Clustering. In this technique, the original dataset is split into *k* clusters where *k* is known. Consequently, each sample x_i should be attributed to one of the clusters S_k, *k*=1, …..,*K*. It should be noted that each cluster is characterized by having a centroid m_k which is defined as the center of masses of all samples in the cluster (Pomerantsev, 2014).

1. **Pattern Recognition:** Pattern Recognition is the area of Chemometrics where the patterns or structure in our data can be discovered. Basically, the methods can be divided into unsupervised and supervised methods. In unsupervised methods, no prior assumptions about structure or groupings in our data is made. Infrared spectra of a large range of polymers may be recorded, for example. A technique like Principal Component Analysis (PCA) is used that produces a plot of the first two principal components. Then examine this plot to see if there are any groupings of the samples and if so do these samples have anything in common structurally. PCA requires no assumptions to be made about possible groupings before carrying out the analysis. PCA is an exploratory method of data analysis used to first look at any structure in our data. Unsupervised Pattern Recognition is also referred to as Cluster Analysis. The chief method used in cluster analysis is Hierarchical Cluster Analysis which attempts to group objects which are 'similar' (there are a range of methods used to measure this similarity) and this is displayed on a dendrogram. In supervised methods, a training set of objects is used where the groups are known. The aim is then to form a rule, based on the measurements, which will assign each object to its correct group. The object is assigned to a group according to which group is the 'closest'. It is the choice of a distance measure that varies between the methods.
2. **Unsupervised Cluster Analysis**
    a. **Principal Component Analysis (PCA):** PCA is a data reduction method which reduces the initial data set to a set of new variables (called principal components) which are much smaller in number than the original number of variables but retains most of the information in the original data set. Visually, PCA can be displayed as the score plot of the first two principal components. This topic is covered further in the section above 'Principal Component Analysis.'
    b. **Hierarchical Cluster Analysis (HCA):** The initial step in this analysis is to determine similarity between objects. The key is the measure of similarity used. Options are:
        i. **Correlation Coefficient Between Samples:** This is a statistical measure of the strength of a linear relationship between paired data. In a sample, it is denoted by *r* and is by design constrained as $-1 \leq r \leq 1$ where positive values denote positive linear correlation, negative values denote negative linear correlation, a value of zero denotes no linear correlation and the closer the value is to 1 or -1, the stronger the linear correlation.
        ii. **Euclidean Distance:** The distance between 2 samples *k* and *l* is defined as:

$$d_{kl} = \sqrt{\sum_{j=1}^{J} (x_{kj} - x_{lj})^2} \text{ (Brereton, 2009)} \tag{19}$$

where there are J measurements and xkj is the jth measurement on sample k. Each measurement might be an absorbance in a spectrum at one wavelength, for example (Brereton, 2009).

        iii. **Manhattan Distance:**

$$d_{kl} = \sum_{j=1}^{J} \left| x_{kj} - x_{lj} \right| \text{(Brereton, 2009)} \tag{20}$$

iv. **Mahalanobis Distance:**

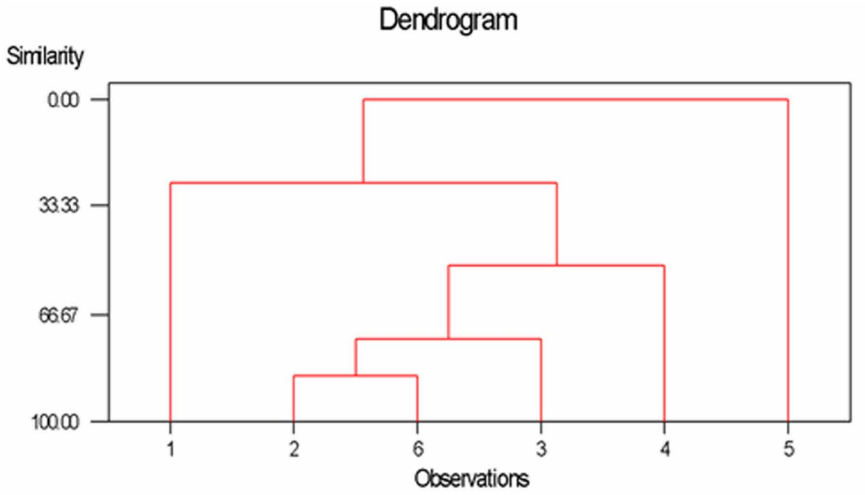
$$d_{kl} = \sqrt{(x_k - x_l) * C^{-1} * (x_k - x_l)^T} \text{ (Brereton, 2009)} \quad (21)$$

Here the x's are column vectors of measurements on a single object and C is the variance-covariance matrix whose elements represent the covariance between any two variables.

Once a measurement of similarity is decided, the next step is to link the objects. This can be done in several ways. The most common approach is to link the objects one at a time using the chosen similarity measurement. This can be depicted in a dendrogram (Refer Figure 5). For example, suppose that we have 6 objects or samples and the correlation coefficient is used as the similarity measurement. The two most similar objects are linked as a branch at the bottom of the tree. This group is then linked to the object which is most similar to this group to form a new group. This is continued till all objects are linked in the tree. An example of such a tree is shown below.

3. **Supervised Pattern Recognition:** With these types of methods, look at whether measurement of a property or set of properties can be used to assign an object to a group. For example, could the infrared spectrum of a polymer be measured and determine whether it is a polyolefin or a condensation polymer? First establish a model using a training set of objects. This set must contain sufficient members of each of the groups. The question is then how to assign membership of the group.
   a. **Discriminant Analysis:** This is a collection of parametric classification methods that models each class by its centroid and its covariance matrix and assigns objects to the 'closest' class. The methods differ in the way the object-class 'distance' is calculated.
      i. **Nearest Mean Classifier (NMC):** This is the simplest method. The simple Euclidian distance is used as the distance measure to the class centroid. The class centroid is the average for each measure for the objects in the group. For example, if the measure was

Figure 5. Example dendrogram for classification of observations (Brereton, 2009)



an IR spectrum, the centroid would be a J-vector where each j element was the average of absorbances at the j^{th} wavenumber. This method is a poor performer because it ignores scale differences. In the IR example, this would not be so important but if the measures have very different scales (e.g. if we had environmental data and one measure was pH and another EC in uS/cm then the EC measure would dominate).

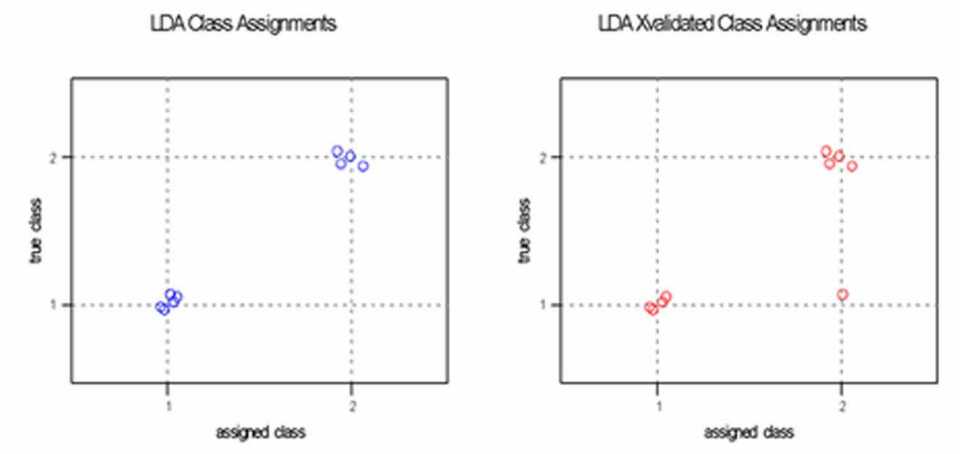
ii. **Linear Discriminant Analysis (LDA):** The distance method used is the Mahalanobis Distance. The method assumes the same covariance structure for each group i.e. each group is equally 'scattered'. Boundaries between the classes are straight lines or planes (Refer Figure 6).

iii. **Quadratic Discriminant Analysis (QDA):** This method also uses the Mahalanobis distance measure but a covariance matrix is calculated for each group. Thus, the measure varies depending on what group the object is in. QDA works well when group differences depend on scale and not location. Boundaries between groups can be curved.

iv. **Regularized Discriminant Analysis (RDA):** This method spans all the discriminant methods above. Often the optimum method is found between the above methods. RDA is a biased method as it has two adjustable parameters λ and γ. λ biases the method towards a single class covariance matrix. γ shrinks the class covariance matrix towards a multiple (the average of the eigenvalues) of the identity matrix. Both parameters have values between 0 and 1. $\lambda = 0$ and $\gamma = 0$ regularized discriminant analysis is same as quadratic discriminant analysis. $\lambda = 1$ and $\gamma = 0$ is same as linear discriminant analysis. $\lambda = 1$ and $\gamma = 1$ is same as the nearest mean classifier.

v. **Soft Independent Modeling of Class Analogy (SIMCA):** SIMCA, like RDA, is a biased version of discriminant analysis. Instead of calculating unbiased class covariance matrices, each group is represented by a principal components model. An object is classified according to its distance from this model. The method is 'soft' in that classes can overlap and objects can belong to more than one class.

Figure 6. Example Linear Discriminant Analysis (LDA) plots (Brereton, 2009)



vi. **Partial Least Squares Discriminant Analysis (PLSDA):** PLS can be used to carry out class modeling. A PLS model is set up the usual way with the y variable, a number indicating group membership. For example, consider the investigation of a 2-way classification such as determining gender from lifestyle preferences. A training set with columns (the X block) indicating preferences for various lifestyle choices is set up $y = 1$ for male and $y = -1$ for female is then assigned. The model is then applied to the test set to predict y and if y is positive, assign male and female for negative y.

Wavelet Transforms

Most analytical instruments often develop noises and fluctuations at the recording stage of the spectrum. This causes a reduction in the original signal of the analyte leading to decreased signal to noise ratio also called noise effect. Noise effect is often eliminated by various means so as to yield quality information from the acquired experimental data. Wavelet Transforms represents one of the most powerful methods to improve Signal to Noise ratio. A wavelet is defined as a family of functions derived from a basic function called the wavelet basis function by dilation and translation. Wavelet basis functions are those functions with some special properties such as orthogonality, compact support, symmetry and smoothness. Wavelet Transform is a projection operation of a signal onto the wavelet (Chau et. al., 2004).

In general, Wavelet Transform is a wavelike function that upon scaled and translated, can be used to decompose a signal into its basic constituents at different scales. Each scale component can be converted into a frequency range. Thus, the resulting Wavelet Transform measures the time-frequency variations of frequency components in a non stationary sign (Liang, 2014). The Wavelet Transform method is often considered to be advantageous over the traditional Fourier Transform method when the signal contains discontinuities and sharp spikes. The method also offers good localization properties in both, the time domain and the frequency domain (Bos & Vrielink, 1994).

Several applications of Wavelet Transforms have been documented such as pre-processing of infrared spectra deionizing or compression of signals through thresholding (Alsberg et. al., 1997; Mittermayr et. al., 1996; Walczak & Massart, 1997), Pattern Recognition and compression of data (Walczak et. al., 1996), and qualitative analysis based on linear models (Depczynski et. al., 1999; Jouan-Rimbaud et. al., 1997).

Continuous and Discrete Wavelet Transforms

Wavelet Transform comprises two distinct parts called the Discrete Wavelet Transforms (DWT) and Continuous Wavelet Transforms (CWT) which were developed independently in several fields (Daubechies, 1992; Ma & Shao, 2004; Walczak, 2000). While CWT is popular among physicists, the DWT is more common in numerical analysis, signal and image processing.

1. Continuous Wavelet Transform

The CWT is an operator that displays and analyzes the characteristics of a signal depending on two variables: time and scale. Hence, as a two-variable function, CWT can be considered as a surface or image. CWT is typically defined with respect to a specific function, called a mother wavelet that satis-

fies some particular properties. The Continuous Wavelet Transform $W_{(a\,b)}$ for a time signal $x(t)$ using a wavelet function $\psi(t)$ is given by the equation

$$W_{(a,b)}^{\psi} = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi^{*}\left(\frac{t-b}{a}\right)dt \quad \forall\, a,b \in R \text{ (Dinç \& Baleanu, 2004; Maldague,1994)} \tag{22}$$

where a, b and R represents scale variable, shift variable and set of real numbers respectively. The wavelet function ψ(t) is a continuous function in both the time domain and the frequency domain called the mother wavelet and the superscripted asterisk (*) symbol represents operation of complex conjugate. For real-valued wavelets, $\psi^{*}(t) = \psi(t)$, the mother wavelet acts as a source function to generate daughter wavelets by dilation and shift operations from the mother wavelet and is represented by the equation

$$\psi_{(a,b)}(t) = \frac{1}{\sqrt{a}}\,\psi\left(\frac{t-b}{a}\right) \text{(Dinç \& Baleanu, 2004; Maldague, 1994)} \tag{23}$$

where a and b are dilation and translation parameters respectively.

It should also be taken into consideration that not every function can qualify to be a mother wavelet (Sadowsky, 1996). In order for a function to be a mother wavelet, it should satisfy an essential property called the "admissibility condition" as described by the following equation

$$C_{\psi} = \int_{0}^{\infty} \frac{\left|\widehat{\psi}(\omega)\right|^{2}}{\omega}\, d\omega < +\infty \text{ (Dinç \& Baleanu, 2004; Maldague, 1994)} \tag{24}$$

where C_{ψ} is a constant corresponding to a particular wavelet $\psi(t)$, ω is the frequency and $\psi(\omega)$ is the Fourier Transform of the wavelet $\psi(t)$.

2. Discrete Wavelet Transform

The main difference between the Discrete Wavelet Transform (DWT) and CWT is that it decomposes the signal into mutually orthogonal set of wavelets. It is therefore an implementation of the Wavelet Transforms using a discrete set of the wavelet scales and translations obeying some defined rules (Dinç & Baleanu, 2004; Liang, 2014; Maldague, 1994). The basic functions for DWT are the scaled and dilated versions of the wavefunctions $\psi(t)$ and scaling function $(\phi(t))$ and can be conveniently expressed by the equation below:

$$\phi(x) = \sum_{k=-\infty}^{\infty} a_{k}\phi(S_{k} - k) \text{ (Dinç \& Baleanu, 2004; Liang, 2014; Maldague, 1994),} \tag{25}$$

where S is the scaling (normally chosen as 2).

One of the distinct features of DWT is that it can be regarded as a mathematically formalized subband coder and are normally implemented as a bank of bandpass filters (Dinç & Baleanu, 2004; Maldague, 1994). Digital Filter banks are a set of four filters consisting of both low pass and high pass filters and are used in the analysis and reconstruction of the signals. The filters are collectively called Quadrature –Mirror Filters (QMFs). The QMFs ensures perfect reconstruction of the signal with no loss of information.

Wavelet Translation and Dilation

Translations and dilations are two characteristic operators applied to single real valued functions of the general form $\psi \in L^2(R)$ Here, ψ is typically referred to as the analyzing wavelet (otherwise known as the 'mother' wavelet), and $L^2(R)$ refers to the space of infinite-energy (also known as square-integrable) functions. A translation is defined as a shift of the argument along the real axis. For example, for a given function $\psi(t)$ and a real value τ, the translation of ψ is given by $\psi(t-\tau)$. Dilation, on the other hand, simply refers to a scaling of the argument, e.g. for a given function $\psi(t)$ and a positive parameter *s*, a dilation of ψ is given by $s^{-1/2}\psi(t\ /\ s)$. Parameter *s* here refers to a continuous, positive real parameter indicative of scale. Therefore, a dilation of a function corresponds to either an expansion or contraction of the function ψ. The extra multiplicative term $s^{-1/2}$ in the dilation expression is introduced simply as a Normalization factor to guarantee an orthonormal wavelet basis.

Other Data Reduction Techniques

1. Linear Calibration Correlation

In order to clearly understand the concept of Least Squares Line, it is important to examine the equations used in the process. Given two sets of data, *X* and *Y*, which are related to each other by the following equation below:

$Y = mx + b$ (Robinson, Frame & Frame II, 2014) (26)

The least squares slope of a line fitting this data is given by the equation below:

$$m = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2} \text{ (Robinson, Frame \& Frame II, 2014)} \quad (27)$$

And the least squares Y-intercept is given by:

$b = \vec{Y} - m\overline{X}$ (Robinson, Frame & Frame II, 2014) (28)

In the above equations, *i* is the data point index, *m* represents the slope and *b* is the *Y*-intercept of the line (Robinson, Frame & Frame II, 2014).

The calculations involved in determining the calibration line constitute subtracting the average *X* from all *X* values, the average *Y* from all *Y* values, then consequently performing the appropriate summing, multiplication, squaring and division (Robinson, Frame & Frame II, 2014).

2. Errors and Confidence Limits

Standard Error is a factor that tells us how accurate our estimate of the mean is likely to be. Confidence Limits, on the other hand, are indicators to determine how accurate the mean is likely to be (Robinson, Frame & Frame II, 2014).

3. Partly Straight, Partly Curved Calibration Plots

As previously stated, calibration plots are generated by expressing the relationship between sample concentration and a measurable variable as

$y = f(x)$ (Vandecasteele, 1997) (29)

where y is the measurable variable and x is the sample concentration. In linear calibration plots, this relationship takes the form of

$Y = mx + b$ (Vandecasteele, 1997) (30)

where *Y* is the measurable variable, 'm' is the slope of the linear curve, 'x' is the sample concentration and 'b' is the y-intercept of the linear curve. There are, however, numerous reasons why a theoretically linear plot can exhibit curving or deviance from the projected least-squares slope. An unavoidable cause of data variance is the indeterminate, or random errors which are the result of uncertain measurements or unknown human inaccuracies. Indeterminate errors cause scattering along either side of the least-squares slope.

Some calibration plots exhibit minimal scattering but are not entirely linear. Partly straight, partly curved calibration plots can occur when using spectrometric methods of obtaining a measurable variable (y). Curves which move away from a previously determined least-squares slope typically move in the negative direction on the high-concentration side of the calibration plot and may represent the maximum detection capabilities of the method or machinery in question. These non-linear deviations can appear quadratic in nature as the analyte approaches maximum detectability and peaks at the limit of detection. Calibration plots exhibiting this type of curving cannot be analyzed using standard addition unless the calibration plot of the machinery's internal standard displays the same type of nonlinear behavior at the same x-values as the analyte (Vandecasteele, 1997).

Regression Diagnostics

Comparison of regression models is best done by first calculating the residuals.

The residuals are calculated as follows:

$e_i = y_i - Y_i$ (Brereton, 2009) (31)

where Yi are the calculated values using the following model:

$$ESS = \sum_i e_i^2 \; TSS = \sum_i (y_i - mean(y))^2 \text{ and } R^2 = 1 - \frac{ESS}{TSS} \tag{32}$$

R2 is called the 'Coefficient of Regression' and can also be calculated as the square of the correlation coefficient of y and Y. This parameter gives a measure of the 'goodness of fit' of the model and gives the percentage of variation in the data which can be explained by the regression model.

The predictive power of the model can be determined from the Cross-Validated residuals

$e_{CV(i)} = y_i - \hat{y}_{(-i)}$ (Brereton, 2009) (33)

where $\hat{y}_{(-i)}$ denotes the predicted value of the ith observation from a model calculated without that observation. Therefore, the Predicted Residual Error Sum of Squares (PRESS) can be calculated as:

$$PRESS = \sum_{i=1}^{n} e_{CV(i)}^2 \text{ and } R_{CV}^2 = 1 - \frac{\mathrm{P}\, RESS}{TSS} \text{ (Brereton,2009)} \tag{34}$$

The R_{CV}^2 values are a measure of the predictive ability of the model. The 'best' model, in terms of the number of principal components used in the model, is thus the one which gives the highest value of R_{CV}^2. Note that while R2 always improves with the addition of more components, this is not true of R_{CV}^2 The situation of too many components is an example of 'overfitting'.

Another diagnostic used is RMSEC (Root Mean Square Error of Calibration) as given by:

$$RMSEC = \sqrt{\frac{ESS}{df}} \text{ (Brereton, 2009)} \tag{35}$$

Here, *df* is the degrees of freedom. RMSEP (Root Mean Square Error of Prediction) is a similar diagnostic summed over the prediction samples.

Variable Selection

In regression modeling for data sets with a small number of variables, the common method of variable selection is stepwise variable selection. The common approaches are forward and backward selection. Forward Selection starts off with a single variable (the variable which is most strongly associated with the

response y). In subsequent steps, variables not present in the current model are considered for addition, and then add the variable which has the highest association with the residuals from the current model. In backward elimination, the model starts with all variables and eliminates at each step the variable whose exclusion results in the lowest increase in residual sum of squares. Termination rules are usually evoked when addition or elimination of variables achieves no significant improvement. Stepwise Regression is a combination of the two processes where variables may be added or removed according to certain criteria. A variable previously removed can come back into the model at a later stage, for example.

Stepwise Regression methods are only feasible when the number of variables is relatively small and independent. With data sets from Chromatography, with measurements at each wavelength or time interval there may be thousands of variables for each sample and this number will greatly exceed the number of objects or samples. While using PLS or PCR, variables can be combined into a small number of 'latent' variables, a process of variable selection can greatly improve the performance of the model.

Simple inspection of the data, as overlapping spectra or chromatograms, for example, can sometimes be sufficient. 'Baseline' areas which are predominantly noise can be removed and possible outliers identified. Simple functions of the variables can then be examined to help choose regions to be excluded. These include, for each column: (i) Mean (ii) Standard Deviation (s.d.) (iii) Correlation of the column with the response variable, y (iv) S.D./Mean.

The mean indicates large responses but this may not vary much from sample to sample (e.g a contaminant at a constant level in all samples). The Standard Deviation shows which variables have the most variation across all the samples. This variation, however may not be due to variation in y. A component, which is not the response variable being analyzed, could vary between samples. The variation between samples due to y can be examined by using the Correlation Coefficient. Using S.D./Mean may help to identify less intense peaks that may be interesting.

There is a very large number of other variable selection processes which have been advocated in the literature. Criterion-based procedures look at all possible models and evaluate them using some criterion. R^2 may appear to be the simplest but R^2 always increases as variables are added to a model so using this as a criterion results in 'overfitting' and useless models for prediction. Adjusted R^2, called R_a^2, can be used where

$R_a^2 = 1-(n-1)(n-p)(1-R^2)$ N is the number of samples and 'p' the number of variables. Another commonly used criterion is Predicted Residual Sum of Squares (PRESS) which is the sum of squares of the residuals (the i^{th} residual is calculated using a model with the i^{th} variable left out of the model). The model with the lowest PRESS is selected.

The problem with criterion methods is that the number of models grows exponentially as 'p' increases as there are $2^p - 1$ possible models. Even with as few as 100 variables, it would take about 1021 years to evaluate all models even at a rate of 10 evaluations per second. Algorithms have thus been developed to improve the selection process for models. The best known is probably the Genetic Algorithms methods which follow the concept of 'survival of the fittest' when competing with other models. Other procedures such as Particle Swarm Optimization and Ant Colony Optimization, which both mimic biological processes, have been suggested. Overall, though, the processes of simple inspection and plotting functions of the variables that were described previously will be sufficient to select regions for modeling.

Programming Tools Used in Analytical Chemistry Big Data Storage and Visualization Techniques

Various programming tools are available for Big Data Storage and Visualization Techniques in Analytical Chemistry. MapReduce is a programming model correlated with the implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster (Shankland, 2008). It was proposed by Google for distributed processing of large datasets on massively parallel systems (Dean & Ghemawat, 2004). The program of MapReduce consists of two parts: *Map()* procedure (a method) and *Reduce()* method. The *Map()* procedure is used for filtering and sorting of data (e.g. classify students by first name). The *Reduce()* method, on the other hand, functions for summary operation. (e.g. determining the total number of students in each queue). The MapReduce System is also called an "infrastructure" or "framework" system. It coordinates the processing by marshalling the distributed servers as well as running various tasks in parallel. It also supervises all communications and data transfers between various parts of the system, and provides warnings for redundancy and Fault Tolerance (Dean & Ghemawat, 2004).

Within the context of Chemometrics, the MapReduce model has been used in an algorithm for calculating Principal Component Regression (PCR). The algorithm consists of several steps: centering of input matrix of regressors, optional regressors scaling, Principal Components Decomposition of the preprocessed input matrix, PCR parameters calculation, Regression Quality evaluation and calculation of prediction for a given set of samples. All these steps except for one are implemented in terms of map-reduce functions and could, therefore, be parallelized and scheduled by Hadoop (discussed below). Principal Components Decomposition is the only computational step realized in non-parallel manner because sequential implicit QL-algorithm (Demmel, 1997) used in this step solves the eigenvalue problem for up to 1000 regressors faster than what the Hadoop starts and warms (Nuzhdin & Zhilin, 2012).

Apache Hadoop, an open- source software implementation of MapReduce, is used for distributed storage and also for processing of relatively large size of data sets (i.e. Tera- and even petabyte scale) (Nuzhdin & Zhilin, 2012). It is made up of computer clusters which is built from commodity hardware. It is believed that all the modules in Hadoop are outlined with a foundational supposition such that hardware failures are common phenomena and should be automatically handled by the framework (The Apache Software Foundation, 2014). Hadoop detects itself and manages failures at the application layer and convey a highly-available service on top of a cluster of commodity machines rather than rely on hardware to deliver high-availability (BigFoot Team, 2013). As previously mentioned, Hadoop was used in PCR algorithm. Specifically, the algorithm was tested on experimental 2-node Hadoop cluster for synthetic datasets of the dimension 1,000,000 x 500 and demonstrated speedup factor of 1.8 (Nuzhdin & Zhilin, 2012).

Apache Hive is considered to be the *defacto* standard for interactive Structured Query Language (SQL) queries over petabytes of data in Hadoop since its developement in 2008.The Apache community has considerably improved Hive's speed, scale, and SQL semantics with the completion of the Stinger Initiative, and the next phase of Stinger. Hive easily integrates with other critical data center technologies using a familiar Java Database Connectivity (JDBC) interface. According to data analysts, the usage of Hive is to query, summarize, explore and analyze data, and then turn these into actionable business insight. The advantages of using Hive for Enterprise SQL in Hadoop include its familiarity by many users as well as its compatibility with many devices. It is also considered fast, scalable and extensible (Apache Software Foundation, 2017). While there may be none or limited studies showcasing the use

of Apache Hive in analyzing analytical data, it may potentially be used for facilitating, querying and managing massive datasets generated by various sophisticated analytical instrumentation.

Spark is an in-memory data analysis with a *Mapreduce* programming model written in Scala. Resilient Distributed Datasets (RDDs), fault-tolerant data structures for Cluster Computing is where Sto store and mpark is based. The RDDs are established, subdivided assemblage of objects that support a wide range of transformations and this allow the apps to keep working sets in memory for efficient reuse (caching). By efficacy of working in memory, Spark is extraordinarily efficient for iterative algorithms and interactive mining (BigFoot Team, 2013; Zaharia et. al, 2010). A crop breeding data analysis platform on Spark has been proposed. The platform consists of Hadoop Distributed File System (HDFS) and cluster based on memory iterative components. With this cluster, crop breeding large data analysis tasks in parallel through API provided by Spark was achieved (Chen et. al., 2016).

Apache Pig is a platform for analyzing huge data sets consisting of a high-level language for expressing data analysis programs coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization which in turns enables them to handle very large data sets. Pig is a high level scripting language that is used with Apache Hadoop (Apache Software Foundation, 2016; Apache Software Foundation, 2017). Within the context of applied Spectroscopy in Analytical Chemistry, Pig provides execution framework for parallel computation in a study involving a novel quantitative spectral analysis method based on parallel BP neural network for dissolved gas in transformer oil. The parallel BP Neural Network model is performed on the Hadoop Cluster Computing platform for component prediction. The experimental results verify that the proposed model can predict the component concentrations of the dissolved gas in transformer oil correctly and has high effectiveness (Zhong et. al., 2016).

Apache Cassandra is a tool which is developed by Facebook and this is a distributed data storage system comparable to BigTable. Apache Cassandra is designed for superintending sizable amounts of structured data dispersed across many commodity servers, thus, delivering a key-value store with concordant consistency. The Cassandra API consists of three very simple methods. These are the insert, get and delete. This allows the user to operate data with the use of multi-dimensional map indexed by the key. Highly available service with no single point of failure is the main goal of Cassandra (BigFoot Team, 2013; Lakshman & Malik, 2010). Apache Cassandra has possible applications in analyzing large genomic datasets or data output from analytical instrumentation.

Navigating large datasets in Chemometrics and Analytical Chemistry might be facilitated using various distributed storage systems such as Amazon Dynamo and Google Bigtable. Such storage systems will be discussed below. Further, the advent of powerful

Chemometric and analytical high-throughput methodologies have paved a way to generate massive datasets. These datasets will be stored in databases using modern data compression and data management as mentioned in this section. Basic visualizations such as bar charts and scatter plots are now realized as JavaScript-based interactive views and might have potential applications in Chemometrics and Analytical Chemistry.

Amazon Dynamo has been developed and used by Amazon which is a key-value distributed storage system. Dynamo is a structured overlay which is based on unchanging assortment with utmost one-hop request routing. It uses a vector clock scheme and a write operation. Clock scheme is used to perceive conflicts while a write operation requires a read of the timestamps (BigFoot Team, 2013; DeCandia et. al., 2007).

Google Bigtable was designed by Google which is a distributed storage system. Google Bigtable is used to store and manage petabytes of structured data across thousands of commodity servers. At the start, Google outlined Bigtable as distributed data storage solution for several applications (like Google Earth and Google Finance), which aims in providing adjustable, high performance solution for different application requirements (BigFoot Team, 2013; Chang et. al., 2007).

jQuery is a cross-platform JavaScript library which is designed to make more comprehensible the client-side scripting of HyperText Markup Language (HTML) (The jQuery Foundation, 2017). In addition, jQuery is the most popular JavaScript library in use at the present time. It has an installation on the 65% of the top million highest-trafficked sites on the web. The jQuery's syntax is designed to make it uncomplicated to steer a document. It also provides potential for developers to generate plug-ins on top of the Javasript library. This empowers developers to create abstractions for low-level interaction and animation, advanced effects and high-level, themeable widgets (jQuery, n.d.).

Wildfly is a free and open-source software. This is an application server authorized by JBoss and is currently reinforced by Red Hat. Wildfly is written in Java and executes the Java Platform, Enterprise Edition (Java EE) specification and also runs on numerous platforms (Wildfly, n.d.).

JavaScript library is a library of pre-written JavaScript which allows for easier development of JavaScript-based applications, especially for AJAX and other web-centric technologies (JavaScript library, n.d.).

Various programming languages such as Perl, Java, Scala, C, C++, C#, Python, PHP and Ruby on Rails are considered commonly used languages that may have potential applications in both Chemometrics and Analytical Chemistry. For example, a set of Perl scripts was written to extract structural parameters from the x-rays in one study (Worley, 2015). In another study, Haystack, which is a web-based server uses the scripting languages Perl and R and a website interface powered by PHP (Grace et al., 2014). Some background information about these programming languages is provided below.

Perl is a popular open-source programming language. It is a scripting language consisting of a sequence of commands that the computer must execute and perform (Berman, 2009).

Java is an all-purpose computer language that is concurrent, class-based, object-oriented, and specifically designed to have as few implementation dependencies as possible (Gosling et. al., 2014). An RCDK package, a Java framework for Chemoinformatics was developed in R that provides the user with access to the CDK. The library allows the user to load molecules, evaluate fingerprints and calculate molecular descriptors (Guha, 2007).

Another all-purpose programming language is the Scala. Scala is an acronym for "Scalable Language". This means that Scala grows with you. You can play with it by typing one-line expressions and observing the results (Odersky, 2017). It has been applied in the areas of Chemometrics specifically in its implementation of the PLS algorithm (Dayal & MacGregor, 1997).

C, C++, C# are other programming languages that have been used in scriptwriting in Chemometrics (Einax, 1995). C is a powerful system programming language, and C++ is an excellent general purpose programming language with modern bells and whistles. C# is a programming language designed by Microsoft. It is based on C/C++, and bears a striking similarity with Java in numerous ways. C# aims to combine the high productivity of Microsoft's Visual Basic and the raw power of C++ (Rajaram, 2007).

Python is a general-purpose, high-level programming language whose design philosophy emphasizes code readability (Miller & Ranum, 2014). Hypertext preprocessor (PHP), on the other hand, is a server-side scripting language designed primarily for web development but also used as a general-purpose

programming language (Gosselin et. al., 2010). Lastly, Ruby on Rails is an open source framework developed to increase programmer productivity and reduce entry barriers to programming Web applications (Bachle & Kirchberg, 2007).

In general, Chemometrics and Analytical Chemists have a wide array of programming tools to choose from. The choice must depend on many factors such as the technical ability of the programming, hardware and Operating System availability, User Interface options, time scale of the project and necessity to interface to other people's programs (Einax, 1995).

Signal Processing in Chemometrics

Two broad questions in Chemometrics relate to detection and estimation. While detection seeks to answer the question of "Is a compound present?", estimation seeks to answer "How much of the compound is present?" The qualitative nature of detection, and the quantitative nature of estimation have been immensely aided with tools such as Spectrometry. An excellent reference to Signal Processing in Analytical Chemistry is in (Wentzell & Brown, 2000).

A major component of spectrometry data analysis is Signal Processing, which is accomplished through techniques such as Fourier Analysis and Wavelet Transforms. While Fourier Transforms are useful in detecting the frequencies present in a time-/frequency – series analysis, Wavelet Transforms possess the added advantage of determining the frequency and location of the event on the time scale. Fourier Transforms have been used in applications such as discrimination of cyanobacterial strains (Kansiz et. al., 1999) in conjunction with Fourier Transform Infrared Spectroscopy (FTIR), detection of specific varieties of the coffee bean in coffee samples (Briandet et. al., 1996) and many other applications as diverse as food spoilage (Ammor et. al., 2009), wine composition (Coimbra et. al., 2002), ivory analysis (Brody et. al., 2001), and bacterial sample analysis (Goodacre et. al., 2002; Kim et. al., 2005). Spectral signal estimation using Wavelet Transforms has been studied (Jetter et. al., 2000) and used for applications such as detecting moisture content in wheat samples and determining the quality of pulp in processing paper using Acoustic Chemometrics (Bjork, 2007). The presence of noise in these signals can be attributed to signal transfer methods, conversion from analog to digital representation, statistical errors or a priori baseline parameters. The effects of noise are mitigated by using techniques such as thresholding, filtering and temporal or spatial processing to suppress the effects of noise at specific locations in the signal. Improving the Signal-to-Noise Ratio (SNR) has tremendously benefited from the multi-variate signals obtained from chemometric tools.

Techniques such as PCA and PLS presented earlier in this chapter have for years been the primary methods to separate the signal from the noise. Coupled with advances in detection and estimation theory, statistical modeling and prediction theory using neural networks, noise and other disturbances can be better modeled using Fourier Transforms and Wavelet Transforms. Wavelet Transform applications in Chemometrics have been studied in (Tan & Brown, 2002) to obtain smoother multivariate signals with less background noise and to better isolate variables for process analysis in a multivariate analysis using Partial Least Squares and multiresolution analysis. Specific applications of Wavelet Transforms in Chemometrics have been presented in areas such as Chromatography (Daszykowski & Walczak, 2006), prediction of the sugar content in apples (Nicolai, Theron & Lammertyn, 2007) and detection of sweeteners in honey (Zhu, et. al., 2010).

Data Storage in Chemometrics and Analytical Chemistry

The very large-scale data sets generated by statistical and mathematical tools used for analyzing chemometric data is driven from the current trends in data generation. A computing revolution has facilitated large-scale data generation in applications as diverse as genomics, pharmaceutical applications (Singh et. al., 2013), Spectrometry (Marhuenda-Egea et. al., 2013), Food Quality (Cruz et. al., 2013), biological and environmental analysis (Szefer, 2003), atmospheric precipitates (Ofner et. al., 2015), Medicinal Chemistry (Lusher et. al., 2014) and Toxicity (Azmi et. al., 2005). The data generated in these real-world applications are multi-dimensional leading to an n-dimensional data space where the data representation and semantics where the memory and computational power requirements might lead to the inadequacy of main memory to store all the data. Two different approaches to data storage have been outlined (Kantardzic, 2011) to address this problem:

1. **Divide and Conquer:** Data can be stored in secondary memory and clustering is performed on these subsets independently. These clusters are then merged to yield a clustering of the entire data set.
2. **Incremental Clustering:** Data is stored in secondary memory and is transferred to main memory for clustering.

Incremental Clustering

Consider five 2-dimensional points $x_1, x_2, ..., x_5$ with the following coordinates: (1,1), (0,2), (1,3), (4,1) and (5,0).

We apply the incremental clustering algorithm (Kantardzic, 2011) with threshold of $\sigma = 2$ to test for similarity within an existing cluster. A distance between points exceeding the threshold of $\sigma = 2$ signals the creation of a new cluster with the latest data point.

Since $x_1(1,1)$ is the first sample, it is assigned to the first cluster C_1 . The centroid of this cluster is calculated as $P_{C_1} = (1,1)$.

The distance from the next sample $x_2(0,2)$ calculated as the Mahalanobis distance (Brereton, 2009):

$d(x_2, P_{C_1}) = \sqrt{(1-0)^2 + (1-2)^2} = 1.414$ (Brereton, 2009)

Since $d(x_{2,}, P_{C_1}) < \sigma$ the point x_2 belongs to the cluster C_1.

The new value of the centroid of this cluster is then given by:

$$P_{C_1} = \left(\frac{1+0}{2}, \frac{1+2}{2}\right) = (0.5,\ 0.5)$$ (Kantardzic, 2011)

For the sample x3(1,3) the distance $d(x_3, P_D) = \sqrt{(1-0.5)^2 + (3-1.5)^2} = 1.58$. Since $d(x_{3,}, P_{C_1}) < \sigma$ the point x3 belongs to the cluster C_1. The new value of the centroid of this cluster is given by:

$$P_{C_1} = \left(\frac{1+0+1}{3}, \frac{1+2+3}{3}\right) = (0.66,\ 2.0)\text{.}$$ (Kantardzic, 2011)

The subsequent distance calculation is $d(x_4, P_{C_1}) = \sqrt{(4-0.66)^2 + (1-2)^2} = 3.486$ (Kantardzic, 2011)

Since $d(x_{4,}, P_{C_1}) < \sigma$, point x_4 is assigned to a new cluster C_2. The new value of the centroid of this cluster is given by $P_{C_2} = (4,1)$.

The last sample is compared with centroids of both clusters P_{C_1} and P_{C_2}:

$d(x_5, P_{C_1}) = \sqrt{(5-0.66)^2 + (1-2.0)^2} = 4.77$ (Kantardzic, 2011)

$d(x_5, P_{C_2}) = \sqrt{(5-4)^2 + (0-1)^2} = 1.414$ (Kantardzic, 2011)

Since the distance between point x_5 and the centroid of cluster C_2 is less than the distance to the centroid of cluster C_1, point x_5 is assigned to cluster C_2. The new centroid of cluster C_2 is given by:

$$P_{C_2} = \left(\frac{4+5}{2}, \frac{1+0}{2}\right) = (4.5,\ 0.5)$$ (Kantardzic, 2011)

The incremental clustering approach will result in different calculations if the order of points considered is changed. Although not applied iteratively in this example, the incremental clustering approach may be used in such a manner. Also, while the above example demonstrates incremental clustering with 2-D data and Euclidean distance between points, it can also be applied to n-dimensional data with different distance metrics such as the simple matching coefficient, Jaccard Coefficient and Rao's Coefficient. An analysis of Clustering Coefficients with an application in entomology is described in (Dalirsefat et. al., 2009).

Divide and Conquer

The Divide-and-Conquer clustering approach has been studied in (Andrews & Fox, 2007; Khalilian et. al., 2016) and (Cui et. al., 2014). In Khalilian, Mustapha & Sulaiman (2016), the authors present two approaches to data analysis in large-scale data sets. The first approach deals with algorithmic analysis where algorithms are used to generate clusters and test incoming samples for similarity with existing clusters. However, the large-scale data sets in applications such as Chemometrics cause challenges in insufficient data storage capacity. A second approach is that of clustering on streaming data. Thus, the data is not stored but is analyzed on the run. This approach called ''data-stream clustering' alleviates data storage problems but results in unique issues related to the data such as change detection in the data stream, detection of gradual or abrupt changes, empty clusters resulting from monotonous data and expenditure of computational and time resources for generating empty clusters. The authors in (Khalilian

et. al., 2016) propose a modified K-Means Divide and Conquer Algorithm for data stream clustering, which also detects outliers, outdated micro-clusters and change in the data stream.

The general concept behind a Divide and Conquer Algorithm, which falls into the general category of divisible algorithms is that the data set is divided into two categories and then each category is divided into two categories. This repetitive division continues until enough partitions have been created to result in data sets of manageable size, both in terms of storage space and computational capacity of the algorithms. Thus, the Divide and Conquer Algorithm works opposite to the incremental clustering approach described above where each data point is considered individually at the outset and is merged with existing clusters to form incrementally large clusters based on the threshold distance for similarity with a cluster. In this sense, the Divide and Conquer Algorithm may be viewed as a top-down approach for data storage while the Incremental Clustering Algorithm may be viewed as a bottom-up approach to clustering. The Incremental Clustering Algorithm is an example of agglomerative algorithm that merge clusters and is more frequently used in the real-world applications.

The problem of data storage and analysis of large-scale data sets requires computational tools in data reduction where the existing data is parsed for the most relevant information for an application. For example, an electronic assistant such as Siri or a search engine will focus on the most relevant terms in a query to generate relevant search results. Consequently, search terms such as "how can I find the recipe to bake a cake?" and "cake recipe" are structured to achieve similar search results, yet data reduction to reduce the dimensionality of data can reduce the storage requirements for the first query by 80% (number of words in the queries). In (Andrews & Fox, 2007), the authors use a data reduction technique based on the K Means Divide and Conquer Algorithm. The k-Means Divide and Conquer Algorithm along with other clustering algorithms is described in Section 3.3.2 of this chapter by Andrews & Fox (2007), and the interested reader is encouraged to refer to this section for detailed analyses.

Visualization Plots, Softwares and Toolboxes Used in Chemometric Techniques

Visualization techniques exist for different areas of Chemometrics. One particular area of Chemometrics that relies heavily on Visualization is multivariate analysis which attempts to make sense of high-dimensional data. Principal Components Analysis, multidimensional scaling and factor analysis all rely on visualization and, to some extent or the other, aim to project high-dimensional data to lower (usually 2-dimensional) subspaces.

Cluster analysis is another area that relies on visualization and in this section the visualization techniques and software toolboxes for creating cluster plots are described.

PCA is the most commonly used technique in cluster analysis and for visualizing clusters and has been discussed above. Various software packages primarily available in R can perform PCA analysis such as prcomp, FactoMineR, cmdscale, hclust, mclust and lm. MATLAB and SAS programs can also perform PCA analysis using the Chemometrics Toolbox and PRINCOMP procedures respectively. R, however, offers a major advantage in that it is freely available. Other commercially available software tools that can perform PCA analysis include Eigenvector PLS Toolbox, Camo Unscrambler, Infometrix and Symbion QT (Refer to Table 3). The primary visualization tools for PCA include screen plots for displaying eigenvalue magnitudes associated with a component and biplots for the simultaneous visualization of *data points* projected to a lower (often 2-dimensional) subspace and variables displayed as vectors.

Factor Analysis (FA) can be performed under the R Program using the FactoMineR Package. MATLAB can also perform FA using its Factor Analysis Toolbox. SAS also offers a FACTOR procedure. The

Eigenvector PLS Toolbox is a comprehensive software system that can perform FA, PCA and Clustering. Other commercially available software tools that can also perform FA include Camo Unscrambler, Infometrix and Symbion QT (Refer to Table 3). The main visualization techniques for FA results include screen plots for displaying eigenvalue magnitudes associated with each component. These are plotted in descending order of size to identify the most important ones. FA can also generate factor-loading plots which provide a visualization for the degree of loading on each factor allowing easy comparison of relative loading magnitudes among a large number of factors. Vector plot of loadings can also be generated using FA.

In Multi-Dimensional Scaling (MDS), a scatterplot showing the projection of the data points into 2- or 3-dimensional space, is the chief visualization tool. Any generic 2-D or 3-D plotting technique usually works for this purpose.

In Regression Analysis including both simple and multivariable regression methods, residual plots are an invaluable and standard tool for showing the residuals of the model, i.e. the difference between the actual values of the dependent variable and the predicted values based on the model. These types of plots are simple plots where the x-axis represents an independent variable and the y-axis represents the residual. This plot shows whether the Linear Regression model is an appropriate model for the data. For example, if strong non- linear relationships exist between the independent and dependent variables, a linear model will usually be inadequate. If the residual plot shows a random pattern around 0, this is usually a fairly reliable indicator that the data supports a linear model. If a strong, systematic, non-random pattern is seen in the residuals, then this is usually seen as evidence that the relationship between the independent and dependent variables is non-linear and therefore a linear model is inappropriate for this type of data. Regression Analysis is a commonly used method for predicting the outcome of a specific independent variable(s) and can be performed using various software packages under the R program, MATLAB and SAS statistical software (Refer Table 3).

A wide range of visualization tools exist for cluster analysis. The main tools for the 2 major types of Clustering (discussed below) can be performed using various software packages such as R, MATLAB, SAS, Eigenvector Toolbox, Camo Unscrambler, Infometrix Pirouette, Symbion QT. The MODECLUS procedure clusters observations in a SAS dataset using any of several algorithms based on nonparametric density estimates (Refer Table 3). There are two types of Clustering Techniques: Hierarchical and Non-Hierarchical Clustering. Hierarchical Clustering utilizes dendrograms which have been briefly discussed and illustrated above. They are top-down tree-like diagrams used to visualize the hierarchy of clusters produced by clustering algorithms. They provide an easy way to examine groupings of variables that are deemed similar and to visualize the degrees of dissimilarity (inter-cluster distance) among variables. Non-Hierarchical Clustering method, on the other hand, is a category of clustering method distinguished from Hierarchical Clustering methods by the fact that they do not have tree like structures and often they work by grouping individuals rather than variables. The most popular algorithm in this class of clustering methods is the k-means algorithm which has been described previously. The main visualization tool used here is a simple scatterplot which allows the visualization of the resulting clusters.

CONCLUSION

Data mining and associated data storage challenges all result from a fundamental problem that probes for insight and hidden patterns in large data sets. Informally, the question presents itself as: What can

Table 3. Summary of Softwares for visualization tools used in Chemometrics (PCA=Principal Component Analysis, FA=Factor Analysis, MDS=Multi Dimensional Scaling)

Software	Functions/ Packages/ Libraries	PCA	FA	MDS	Regression	Clustering
R	prcomp	X				
	FactoMineR	X	X			
	cmdscale			X		
	hclust					X
	mclust					X
	lm, plot				X	
MATLAB	Chemometrics Toolbox	X			X	X
	Factor Analysis Toolbox		X			
SAS	FACTOR procedure		X			
	PRINCOMP procedure	X				
	MDS procedure			X		
	CLUSTER, DISTANCE, MODECLUS procedures					X
	GLM procedure				X	
Eigenvector PLS Toolbox		X	X			X
Camo Unscrambler		X	X	X	X	X
Infometrix Pirouette		X	X		X	X
Symbion QT		X			X	

(R Core Team, 2016; The MathWorks Inc., 2012; SAS Institute, 2011; Eigenvector Research, Inc., 2016; CAMO Software, 2016; Infometrix, Inc., 2016; Symbion Systems, Inc., 2016)

one learn from this data? The challenges of converting data to information has spurred a novel discipline called Data Science, or Big Data, that seeks to combine tools from Computer Science, Statistics and Machine Learning. In this chapter, the applications of Big Data to Analytical Chemistry and Chemometrics are analyzed and focused on five distinct chemometric aspects of large data sets: data acquisition, data preprocessing, data analysis, data storage and chemometric software and toolboxes. The techniques and algorithms presented in the chapter reinforce the idea of using tools from Computer Science, Statistics and Machine Learning for information retrieval and analysis challenges inherent in large-scale data sets. Data analyses involved in Chemometrics include an array of Regression based methods such as PLS, PCR and Ridge Regression methods, as well as data reduction and Pattern Recognition techniques such as PCA and Wavelet Transformation methods.

In order to be effective with data mining in any discipline, one would need to examine the storage of Big Data. Storing Big Data is no easy task as the storage needs to be able to handle large amounts of data and have the ability to scale upwards as more data are added to storage as time progresses. The storage device also needs the ability to handle receiving inputs and delivering outputs as necessary for storage practices or delivery of data to analytic tools respectively. Once the fundamental step of Big Data Storage is satisfied, the data in question can proceed to be preprocessed.

Data Storage in Analytical Chemistry and Chemometrics has benefited from technological advances in paradigms such as MapReduce, Cloud Computing and Parallel Computing. Additional data processing mechanisms for efficient storage and pattern mining have been presented such as Divide and Conquer and Incremental Clustering to optimize storage space and computational efficiency. The diversity of tools for data storage and analysis presented in this chapter lend themselves readily to similar research areas in Big Data involving complex data types (Bioinformatics, Geographical Information Systems), Graph-based and Network Mining (Social Networks, Chemical Structures, Biological Pathways) and Engineering and Science (Software and System Engineering, Recommender Systems and Data Warehousing).

Once the data is preprocessed, it is then modeled to find any possible trends/correlations that could be of use. To understand the findings, the information is often visualized in the form of a graph (e.g., Linear Regression and Multivariable Regression) or as a Scatter plot (e.g., Clustering and Classification). This helps illustrate and further evaluate the possible predictions and patterns of the data that has been analyzed. Softwares that can visualize data range from GUI friendly software SPSS and RapidMiner to programs in software packages such as MATLAB and Microsoft Excel. If the software packages are not an option, it is possible to program in other languages such as C++, Python or Java.

FUTURE RESEARCH DIRECTIONS

As the field of data mining matures and continues to grow, we expect to see a wider range of applications to analytical chemistry. With analytical instruments and sensors beginning to play a more central role in various areas of chemistry, there is a greater awareness of the pivotal nature of data science in this field. For example, the Journal of Analytical Methods in Chemistry recently introduced a special issue called “Big Data and Data Science in Analytical Chemistry and Chemical Industry”. This issue invites submissions covering latest breakthroughs in data analysis methodology in analytical chemistry. Efforts such as this will broaden the reach of data analytic tools useful for dealing with challenges in analytical chemistry. Another important goal moving forward is the development of machine learning tools specifically for problems in analytical chemistry. This is a potential avenue for future research. Most statistical and analytical techniques used in chemometrics were originally designed for application to other fields, but there are certain aspects of the data generated by analytical instruments that are unique to this field. This requires a new breed of analytical chemists who are well-versed and comfortable in both areas, and can develop more powerful analytical approaches specifically tailored to analytical chemistry applications. Now, more than ever, training and educational programs in analytical chemistry need to begin incorporating elements of data science like statistical analysis, machine learning, and programming.

REFERENCES

Alsberg, B. K., Woodward, A. M., Winson, M. K., Rowland, J., & Kell, D. B. (1997). Wavelet Denoising of Infrared Spectra. *Analyst (London)*, *122*(7), 645–652. doi:10.1039/a608255f

Ammor, M. S., Argyri, A., & Nychas, G. J. E. (2009). Rapid monitoring of the spoilage of minced beef stored under conventionally and active packaging conditions using Fourier transform infrared spectroscopy in tandem with chemometrics. *Meat Science*, *81*(3), 507–514. doi:10.1016/j.meatsci.2008.10.015 PMID:20416598

Andrews, N. O., & Fox, E. A. (2007). *Clustering for Data Reduction: A Divide and Conquer Approach*. Department of Computer Science, TR-07-36. Retrieved on September 7, 2016 from http://vtechworks.lib.vt.edu/handle/10919/19848

Apache Software Foundation. (2017a). *Apache Hive*. Retrieved on January 24, 2017 from http://hortonworks.com/apache/hive/

Apache Software Foundation. (2016). *Apache Pig*. Retrieved on January 24, 2017 from https://pig.apache.org/

Apache Software Foundation. (2017b). *Beginners Guide to Apache Pig*. Retrieved on January 24, 2017 from http://hortonworks.com/hadoop-tutorial/how-to-use-basic- pig-commands/

Azmi, J., Griffin, J. L., Shore, R. F., Holmes, E. & Nicholson, J. K. (2005). Chemometric analysis of biofluids following toxicant induced hepatotoxicity: A metabonomic approach to distinguish the effects of 1-naphthylisothiocyanate from its products. *Xenobiotica: The Fate of Foreign Compounds in Biological Systems, 35*(8), 839–852. 10.1080/00498250500297940

Bachle, M., & Kirchberg, P. (2007). Ruby on rails. *IEEE Software*, *24*(6), 105–108. doi:10.1109/MS.2007.176

Banks, D., House, L., McMorris, F. R., Arabie, P., & Gaul, W. (2011). *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS)*. Springer Science & Business Media. Retrieved on October 22, 2016 from https://www.amazon.com/Classification-Clustering-Data-Mining-Applications/dp/3540220143

Barford, L. A., Fazzio, R. S., & Smith, D. R. (1992). *An introduction to wavelets*. Hewlett Packard, HPL:92-124. Retrieved on November 28, 2016 from http://www.hpl.hp.com/techreports/92/HPL 92-124.pdf

Berman, J. J. (2009). *Perl: The Programming Language*. Jones and Bartlett Publishers. Retrieved on January 25, 2017 from https://books.google.com/books?id=maDGMelhpg8C&printsec=frontcover&dq=Perl&hl=en&sa =X&ved=0ahUKEwjKzI2BodzRAhVX9GMKHfQ2DusQ6AEIMzAC#v=onepage&q=Perl&f=fa lse

BigFoot Team. (2013). *Current practices of big data analytics.* BigFoot FP7-ICT-ICT-2011.1.2 Call 8 Project No. 317858. Retrieved on January 24, 2017 from http://bigfootproject.eu/downloads/D.2.1.pdf

Björk, A. (2007). *Chemometric and signal processing methods for real time monitoring and modeling: Applications in the pulp and paper industry*. Retrieved on December 8, 2016 from http://swepub.kb.se/bib/swepub:oai:DiVA.org:kth-4383?vw=full

Bogomolov, A. (2011). Multivariate process trajectories: Capture, resolution and analysis. *Chemometrics and Intelligent Laboratory Systems*, *108*(1), 49–63. doi:10.1016/j.chemolab.2011.02.005

Bos, M., & Vrielink, J. A. M. (1994). The wavelet transform for preprocessing IR spectra in the identification of mono- and di-substituted benzenes. *Chemometrics and Intelligent Laboratory Systems*, *23*(1), 115–122. doi:10.1016/0169-7439(93)E0066-D

Brereton, R. G. (2003). *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. John Wiley & Sons. Retrieved on October 22, 2016 from https://www.amazon.com/Chemometrics-Analysis-Laboratory-Chemical-Plant/dp/0471489786

Brereton, R. G. (2007). *Applied Chemometrics for Scientists*. John Wiley & Sons. Retrieved on October 22, 2016 from https://www.amazon.com/Applied- Chemometrics-Scientists-Richard-Brereton/dp/0470016868

Brereton, R. G. (2009). *Chemometrics for Pattern Recognition*. John Wiley & Sons. Retrieved on October 22, 2016 from https://books.google.com/books?id= Lp5ImWw7bCAC&pg=PA12&lpg=PA12&dq=Chemometrics+for+Pattern+Recognition+bre reton&source=bl&ots=zxWN_Yvrq3&sig=CGmkeWObAXeRiOgEgZ0vA5- KBUY&hl=en&sa=X&ved=0ahUKEwiql9mtkfDPAhXLRSYKHVClCpIQ6AEIVTAH#v=o nepage&q&f=false

Briandet, R., Kemsley, E. K., & Wilson, R. H. (1996). Discrimination of Arabica and Robusta in instant coffee by Fourier transform infrared spectroscopy and chemometrics. *Journal of Agricultural and Food Chemistry*, *44*(1), 170–174. doi:10.1021/jf950305a

Brody, R. H., Edwards, H. G., & Pollard, A. M. (2001). Chemometric methods applied to the differentiation of Fourier-transform Raman spectra of ivories. *Analytica Chimica Acta*, *427*(2), 223–232. doi:10.1016/S0003-2670(00)01206-X

Burgard, D. R., & Kuznicki, J. T. (1990). *Chemometrics*. CRC Press. Retrieved on October 22, 2016 from https://www.amazon.com/Chemometrics-David-R- Burgard/dp/0849348641

CAMO Software. (2016). *The Unscrambler X: Version 10.4*. Retrieved on November 18, 2016 from http://www.camo.com/

Castanedo, F. (2013). A Review of Data Fusion Techniques. *The Scientific World Journal*, *2013*, 1–19. doi:10.1155/2013/704504 PMID:24288502

Chang, F. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems, 26*(2), 1-14. Retrieved on January 24, 2017 from https://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf

Chau, F.-T., Liang, Y.-Z., Gao, J., & Shao, X.-G. (2004). *Chemometrics: From Basics to Wavelet Transform*. John Wiley & Sons. Retrieved on October 22, 2016 from https://www.amazon.com/Chemometrics-Transform-Monographs-Analytical-Applications/dp/0471202428

Chen, S., Wu, C., & Yu, Y. (2016). Analysis of Plant Breeding on Hadoop and Spark. *Advances in Agriculture*, 1-6. Retrieved on January 24, 2017 from https://www.hindawi.com/journals/aag/2016/7081491/abs/

Coimbra, M. A., Gonçalves, F., Barros, A. S., & Delgadillo, I. (2002). Fourier transform infrared spectroscopy and chemometric analysis of white wine polysaccharide extracts. *Journal of Agricultural and Food Chemistry*, *50*(12), 3405–3411. doi:10.1021/jf020074p PMID:12033803

Cruz, A. G., Cadena, R. S., Alvaro, M. B. V. B., Sant'Ana, A. S., Oliveira, C. A. F., Faria, J. A. F., ... Ferreira, M. M. C. (2013). Assessing the use of different chemometric techniques to discriminate low-fat and full-fat yogurts. *Lebensmittel-Wissenschaft + Technologie*, *50*(1), 210–214. doi:10.1016/j.lwt.2012.05.023

Cui, H., Ruan, G., Xue, J., Xie, R., Wang, L., & Feng, X. (2014). A collaborative divide-and- conquer K-means clustering algorithm for processing large data. *Proceedings of the 11th ACM Conference on Computing Frontiers*. Retrieved on September 10, 2016 from 10.1145/2597917.2597918

Dalirsefat, S. B., da Silva Meyer, A., & Mirhoseini, S. Z. (2009). Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, Bombyx mori. *Journal of Insect Science (Online)*, *9*(71), 1–8. doi:10.1673/031.009.7101 PMID:20050782

Daszykowski, M., & Walczak, B. (2006). Use and abuse of chemometrics in chromatography. *TrAC Trends in Analytical Chemistry*, *25*(11), 1081–1096. doi:10.1016/j.trac.2006.09.001

Dayal, B. S., & MacGregor, J. F. (1997). Improved PLS algorithms. *Journal of Chemometrics*, *11*(1), 73–85. doi:

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics (SIAM). Retrieved on October 22, 2016 from https://www.amazon.com/Lectures-Wavelets-CBMS-NSF-Conference-Mathematics/dp/0898712742

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107–113. doi:10.1145/1327452.1327492

Decandia, G. (2007). Dynamo: Amazon's highly available key- value store. *SOSP 2007 Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles*, 205-220. Retrieved on January 24, 2017 from http://s3.amazonaws.com/AllThingsDistributed/sosp/amazon-dynamo- sosp2007.pdf

Deming, S. N., Michotte, Y., Massart, D. L., Kaufman, L., & Vandeginste, B. G. M. (1988). *Chemometrics: A Textbook*. Elsevier. Retrieved on October 22, 2016 from https://books.google.com/books/about/Chemometrics.html?id=G8JMac7OCtAC

Demmel, J. W. (1997). *Applied Numerical Linear Algebra*. Society of Industrial and Applied Mathematics. Retrieved on January 24, 2017 from http://epubs.siam.org/doi/book/10.1137/1.9781611971446

Depczynski, U., Jetter, K., Molt, K., & Niemöller, A. (1999). Quantitative analysis of near infrared spectra by wavelet coefficient regression using a genetic algorithm. *Chemometrics and Intelligent Laboratory Systems*, *47*(2), 179–187. doi:10.1016/S0169-7439(98)00208-1

Dinç, E., & Baleanu, D. (2004). Application of the Wavelet Method for the Simultaneous Quantitative Determination of Benazepril and Hydrochlorothiazide in Their Mixtures. *Journal of AOAC International, 87*(4), 834–841. Retrieved on September 10, 2016 from http://www.ingentaconnect.com/content/aoac/jaoac/2004/00000087/00000004/art00006

Dubrovkin, J. (2014). Big Data Approach to Analytical Chemistry. *International Journal of Emerging Technologies in Computational and Applied Sciences, 14*(142), 205–210. Retrieved on October 22, 2016 from http://www.academia.edu/6315299/Big_Data_Approach_to_Analytical_Chemistry

Dumancas, G. G. (2012). *Simultaneous Spectrophotometric and Chemometric Determination of Cholesterol and Mono-/Polyunsaturated Fatty Acids* (Doctoral dissertation). Retrieved on August 7, 2015 from https://shareok.org/handle/11244/6451

Dumancas, G. G., Bello, G., Hughes, J., & Diss, M. (2014). Comparison of Chemometric Algorithms for Multicomponent Analyses and Signal Processing: An Example from 4-(2- Pyridylazo) Resorcinol-Metal Colored Complexes. *Recent Patents on Signal Processing, 4*(2), 106–115. Retrieved on September 7, 2016 from http://www.ingentaconnect.com/content/ ben/rptsp/2014/00000004/00000002/art00006

Dumancas, G. G., Ramasahayam, S., Bello, G., Hughes, J., & Kramer, R. (2015). Chemometric regression techniques as emerging, powerful tools in genetic association studies. *TrAC Trends in Analytical Chemistry*, *74*, 79–88. doi:10.1016/j.trac.2015.05.007

Ekanayake, J., Pallickara, S., & Fox, G. (2008). *MapReduce for Data Intensive Scientific Analysis*. Presented at the 4th IEEE International Conference on e-Science, Indianapolis, IN. Retrieved on September 10, 2016 from http://escience2008.iu.edu/sessions/mapReduce.shtml

Eigenvector Research, Inc. (2016). *PLS Toolbox 8.2.1*. Retrieved on November 18, 2016 from http://www.eigenvector.com/

Einax, J. (1995). *Chemometrics in environmental chemistry – statistical methods*. SpringerVerlag. Retrieved on January 25, 2017 from https://books.google.com/books?id=aUnmCAAAQBAJ&pg=PA68&dq=C%2B%2B+chemo metrics&hl=en&sa=X&ved=0ahUKEwjH6oLHrNzRAhVE3WMKHf5DBKYQ6AEIIzAA#v=onepage&q=C%2B%2B%20chemometrics&f=false

Esteban, J., Starr, A., Willetts, R., Hannah, P., & Bryanston-Cross, P. (2005). A Review of data fusion models and architectures: Towards engineering guidelines. *Neural Computing & Applications*, *14*(4), 273–281. doi:10.100700521-004-0463-7

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. John Wiley & Sons. Retrieved on November 9, 2016 from https://www.amazon.com/Cluster-Analysis-Brian-S-Everitt/ dp/0470749911

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google File System. In *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles* (pp. 29–43). New York, NY: ACM. 10.1145/945445.945450

Goodacre, R., Shann, B., Gilbert, R. J., Timmins, É. M., McGovern, A. C., Alsberg, B. K., ... Logan, N. A. (2000). Detection of the dipicolinic acid biomarker in Bacillus spores using Curie-point pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. *Analytical Chemistry*, *72*(1), 119–127. doi:10.1021/ac990661i PMID:10655643

Gosling, J., Joy, B., Steele, G. L., Bracha, G., & Buckley, A. (2014). *The Java Language Specification, Java SE 8th Edition.* Addison-Wesley Professional. Retrieved on January 25, 2017 from http://dl.acm.org/citation.cfm?id=2636997

Gosselin, D., Kokoska, D., & Easterbrooks, R. (2010). *PHP programming with MY SQL 2nd edition*. Cengage Learning. Retrieved on January 25, 2017 from https://books.google.com/books?id=yr02MM9UMA8C&pg=PT22&dq=Hypertext+Preprocessor+PHP&hl=en&sa=X&ved=0ahUKEwiRxrfh6t3RAhXEWCYKHYqKAsw4ChDoAQg4MAE#v=onepage&q=Hypertext%20Preprocessor%20PHP&f=false

Grace, S. C., Embry, S., & Luo, H. (2014). Haystack, a web-based tool for metabolomics research. *BMC Bioinformatics*, *15*(Suppl 11), S12. doi:10.1186/1471-2105-15-S11-S12 PMID:25350247

Guha, R. (2007). Chemical informatics functionality in R. *Journal of Statistical Software*, *18*(5), 1–16. doi:10.18637/jss.v018.i05

Helland, I. (2004). Partial Least Squares Regression. In *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc. Retrieved on September 7, 2016 from http://onlinelibrary. wiley.com/doi/10.1002/0471667196.ess6004/abstract

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. doi:10.1080/00401706.1970.10488634

Infometrix, Inc. (2016). *Pirouette*. Retrieved on November 18, 2016 from https://infometrix.com

Isaeva, J., Sæbo, S., Wyller, J. A., Nhek, S., & Martens, H. (2012). Fast and comprehensive fitting of complex mathematical models to massive amounts of empirical data. *Chemometrics and Intelligent Laboratory Systems*, *117*, 13–21. doi:10.1016/j.chemolab.2011.04.009

JavaScript library. (n.d.). In *Wikipedia*. Retrieved on January 24, 2017 from https://en.wikipedia.org/wiki/JavaScript_library

Jetter, K., Depczynski, U., Molt, K., & Niemöller, A. (2000). Principles and applications of wavelet transformation to chemometrics. *Analytica Chimica Acta*, *420*(2), 169–180. doi:10.1016/S0003-2670(00)00889-8

Jouan-Rimbaud, D., Walczak, B., Poppi, R. J., de Noord, O. E., & Massart, D. L. (1997). Application of Wavelet Transform To Extract the Relevant Component from Spectral Data for Multivariate Calibration. *Analytical Chemistry*, *69*(21), 4317–4323. doi:10.1021/ac970293n PMID:21639165

jQuery. (n.d.). In *Wikipedia.* Retrieved on January 24, 2017 from https://en.wikipedia.org/wiki/JQuery

Kansiz, M., Heraud, P., Wood, B., Burden, F., Beardall, J., & McNaughton, D. (1999). Fourier transform infrared microspectroscopy and chemometrics as a tool for the discrimination of cyanobacterial strains. *Phytochemistry*, *52*(3), 407–417. doi:10.1016/S0031-9422(99)00212-5 PMID:9933953

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons. Retrieved on October 22, 2016 from http://onlinelibrary.wiley.com /book/10.1002/9781118029145

Khalilian, M., Mustapha, N., & Sulaiman, N. (2016). Data stream clustering by divide and conquer approach based on vector model. *Journal of Big Data*, *3*(1), 1. doi:10.118640537-015-0036-x

Kessler, R. W. (2013). Perspectives in process analysis. *Journal of Chemometrics*, *27*(11), 369–378. doi:10.1002/cem.2549

Khanmohammadi, M. (Ed.). (2014). *Current Applications of Chemometrics*. New York: Nova Science Pub Incorporated. Retrieved on October 22, 2016 from http://www.worldcat.org/title/current-applications-of-chemometrics/oclc/902847099

Kim, S., Reuhs, B. L., & Mauer, L. J. (2005). Use of Fourier transform infrared spectra of crude bacterial lipopolysaccharides and chemometrics for differentiation of Salmonella enterica serotypes. *Journal of Applied Microbiology*, *99*(2), 411–417. doi:10.1111/j.1365-2672.2005.02621.x PMID:16033474

Kumar, N., Bansal, A., Sarma, G. S., & Rawal, R. K. (2014). Chemometrics tools used in analytical chemistry: An overview. *Talanta*, *123*, 186–199. doi:10.1016/j.talanta.2014.02.003 PMID:24725882

Lakshman, A., & Malik, P. (2010). Cassandra – A Decentralized Structured Storage System. *Operating Systems Review*, *44*(2), 35–40. doi:10.1145/1773912.1773922

Liang, H. (2014). Wavelet Analysis. In D. Jaeger & R. Jung (Eds.), *Encyclopedia of Computational Neuroscience* (pp. 1–3). Springer New York. Retrieved on November 20, 2015 from http://link.springer.com/referenceworkentry/10.1007/978-1-4614-7320-6_422-1

Lusher, S. J., McGuire, R., van Schaik, R. C., Nicholson, C. D., & de Vlieg, J. (2014). Data- driven medicinal chemistry in the era of big data. *Drug Discovery Today*, *19*(7), 859–868. doi:10.1016/j.drudis.2013.12.004 PMID:24361338

Ma & Shao. (2004). Continuous Wavelet Transform Applied to Removing the Fluctuating Background in Near-Infrared Spectra. *Journal of Chemical Information and Computer Sciences, 44*(3), 907–911. 10.1021/ ci034211+

Maldague, X. (1994). *Advances in Signal Processing for Nondestructive Evaluation of Materials*. Springer Science & Business Media. Retrieved on November 9, 2016 from https://www.amazon.com/Advances-Processing-Nondestructive-Evaluation-Materials/dp/9401044597

Marhuenda-Egea, F. C., Gonsálvez-Álvarez, R. D., Lledó-Bosch, B., Ten, J., & Bernabeu, R. (2013). New approach for chemometric analysis of mass spectrometry data. *Analytical Chemistry*, *85*(6), 3053–3058. doi:10.1021/ac303255h PMID:23394213

Matero, S. (2010). *Chemometric methods in pharmaceutical tablet development and manufacturing unit operations* (Doctoral dissertation). Retrieved on March 16, 2016 from http://epublications.uef.fi/pub/urn_isbn_978-952-61-0143-9/index_en.html

McDonald, J. H. (2014a). Confidence Limits. In *Handbook of Biological Statistics* (3rd ed.; pp. 115–120). Baltimore, MD: Sparky House Publishing. Retrieved on December 4, 2015 from http://www.biostathandbook.com/confidence.html

McDonald, J. H. (2014b). Standard Error of the Mean. In *Handbook of Biological Statistics* (3rd ed.; pp. 111–114). Baltimore, MD: Sparky House Publishing. Retrieved on December 4, 2015 from http://www.biostathandbook.com/confidence.html

Miller, B. N., & Ranum, D. L. (2014). *Python programming in context.* Jones and Bartlett Learning. Retrieved on January 25, 2017 from https://books.google.com/books?hl=en&lr=&id=2NzuKZznqUIC&oi=fnd&pg=PR1&dq=what+i s+Python+programming&ots=f-dXUy8OVp&sig=loP3nls- r_KyuRNwEFsQaoeiISk#v=onepage&q=what%20is%20Python%20programming&f=false

Mittermayr, C. R., Nikolov, S. G., Hutter, H., & Grasserbauer, M. (1996). Wavelet denoising of Gaussian peaks: A comparative study. *Chemometrics and Intelligent Laboratory Systems*, *34*(2), 187–202. doi:10.1016/0169-7439(96)00026-3

Mocák, J. (2012). Chemometrics in Medicine and Pharmacy. *Nova Biotechnologica et Chimica*, *11*(1). doi:10.2478/v10296-012-0002-3

Nicolai, B. M., Theron, K. I., & Lammertyn, J. (2007). Kernel PLS regression on wavelet transformed NIR spectra for prediction of sugar content of apple. *Chemometrics and Intelligent Laboratory Systems*, *85*(2), 243–252. doi:10.1016/j.chemolab.2006.07.001

Nuzhdin, P., & Zhilin, S. (2012). Calculating principal components regression in MapReduce arhictecture. *8th Winter Symposium on Chemometrics, Drakino*. Retrieved on January 24, 2017 from http://wsc.chemometrics.ru/media/files/conferences/wsc8/abstracts/WSC8-Nuzhdin-abstract.pdf

O'Donoghue, P. (2012). *Statistics for Sport and Exercise Studies: An Introduction*. Routledge. Retrieved on November 9, 2016 from https://www.amazon.com/Statistics-Sport-Exercise-Studies- Introduction/dp/0415595576

Odersky, M. (2017). *A scalable language*. Ecole Polytechnique Federale de Lausanne (EPFL). Retrieved on January 25, 2017 from https://www.scala-lang.org/what-is-scala.html

Ofner, J., Kamilli, K. A., Eitenberger, E., Friedbacher, G., Lendl, B., Held, A., & Lohninger, H. (2015). Chemometric Analysis of Multisensor Hyperspectral Images of Precipitated Atmospheric Particulate Matter. *Analytical Chemistry*, *87*(18), 9413–9420. doi:10.1021/acs.analchem.5b02272 PMID:26278430

Ovalles, C., & Rechsteiner, C. E., Jr., (Eds.). (2015). *Analytical Methods in Petroleum Upstream Applications*. CRC Press. Retrieved on October 22, 2016 from https://www.crcpress.com/Analytical-Methods-in-Petroleum-Upstream-Applications/Ovalles-Jr/p/book/9781482230864

Pomerantsev, A. L. (2014). *Chemometrics in Excel*. John Wiley & Sons. Retrieved on October 22, 2016 from https://www.amazon.com/Chemometrics- Excel-Alexey-L-Pomerantsev/dp/1118605357

Pomerantsev, A. L., & Rodionova, O. Y. (2012). Process analytical technology: A critical view of the chemometricians. *Journal of Chemometrics*, *26*(6), 299–310. doi:10.1002/cem.2445

Rajaram, J. (2007). *C# interview questions and answers*. Firewall Media. Retrieved on January 25, 2017 from https://books.google.com/books?id=75gN27NLQZAC&pg=PR9&dq=C,+C%2B%2B,+C%23&hl=en&sa=X&ved=0ahUKEwjO44_qq9zRAhWIKWMKHfPmBDcQ6AEILDAD#v=onepage&q=C%2C%20C%2B%2B%2C%20C%23&f=false

R Core Team. (2016). R*: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved on November 18, 2016 from https://www.R-project.org/

Robinson, J. W., Frame, E. M. S., & Frame, G. M., II. (2014). *Undergraduate Instrumental Analysis, 7th edition*. CRC Press. Retrieved on November 9, 2016 from https://books.google.com/books?id=KY7SBQAAQBAJ&pg=PA518&dq=linear+calibration+correlation+y+%3D+mx+%2B+b&hl=en&sa=X&ved=0ahUKEwjKm5LXj5zQAhVEyyYKHSSbAH8Q6AEINTAA#v=onepage&q=linear%20calibration%20correlation%20y%20%3D%20mx%20%2B%20b&f=false

Sadowsky, J. (1996). Investigation of Signal Characteristics Using the Continuous Wavelet Transform. *Johns Hopkins Applied Physics Laboratory Technical Digest, 17*(3). Retrieved on September 10, 2016 from http://www.jhuapl.edu/techdigest/td/td1703/sadowsky.pdf

SAS Institute. (2011). *The SAS system for Windows*. Release 9.2. SAS Inst. Retrieved on November 18, 2016 from http://www.sas.com/

Shankland, S. (2008). Google spotlights data center inner workings. *CNET.* Retrieved on January 19, 2017 from https://www.cnet.com/uk/news/google-spotlights-data-center-inner-workings/

Singh, I., Juneja, P., Kaur, B. & Kumar, P. (2013). *Pharmaceutical Applications of Chemometric Techniques*. International Scholarly Research Notices. 10.1155/2013/795178

Steinberg, A. N., Bowman, C. L., & White, F. E. (1999). Revisions to the JDL data fusion model. In B. V. Dasarathy (Ed.), Proc. SPIE 3719, Sensor Fusion: Architectures, Algorithms, and Applications III (p. 430). doi:10.21236/ADA389851

Symbion Systems, Inc. (2016). *Symbion QT*. Retrieved on November 18, 2016 from http://www.gosymbion.com/

Szefer, P. (2003). Application of Chemometric Techniques in Analytical Evaluation of Biological and Environmental Samples. In *New Horizons and Challenges in Environmental Analysis and Monitoring*. EU Projects. Retrieved on September 10, 2016 from http://www.chem.pg.gda.pl/CEEAM/Dokumenty/CEEAM_ksiazka/Chapters/chapter18.pdf

Tan, H. W., & Brown, S. D. (2002). Wavelet analysis applied to removing non-constant, varying spectroscopic background in multivariate calibration. *Journal of Chemometrics*, *16*(5), 228–240. doi:10.1002/cem.717

Apache Software Foundation. (2014). *Welcome to Apache Hadoop*. Retrieved on January 24, 2017 from http://www.apache.org/foundation/contact.html

jQuery Foundation. (2017). *jQuery*. Retrieved on January 24, 2017 from https://api.jquery.com/

MathWorks Inc. (2012). *MATLAB and Statistics Toolbox Release*. Retrieved on November 18, 2016 from https://www.mathworks.com/

Vandecasteele, C., & Block, C. B. (1997). *Modern Methods for Trace Element Determination*. John Wiley & Sons. Retrieved on November 5, 2016 from https://books. google.com/books?id=E1i3HQNNGncC&printsec=frontcover#v=onepage&q&f=false

Varmuza, K., & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press. Retrieved on October 22, 2016 from https://www.amazon.com/Introduction-Multivariate-Statistical-Analysis- Chemometrics/dp/1420059475

Walczak, B. (2000). *Wavelets in Chemistry*. Elsevier. Retrieved on October 22, 2016 from https://books.google.com/books?id=LGXgAMHjE3oC&pg=PA176&lpg=PA176&dq=Walczak,+B.+(2000).+Wavelets+in+Chemistry.&source=bl&ots=BfcTz9jhHd&sig=4oNd7kwbJGEWYBUqeC5QfqT29oE&hl=en&sa=X&ved=0ahUKEwiKga7m-u_PAhWEOCYKHaRyC_UQ6AEIMDAE#v=onepage&q&f=false

Walczak, B., van den Bogaert, B., & Massart, D. L. (1996). Application of Wavelet Packet Transform in Pattern Recognition of Near-IR Data. *Analytical Chemistry*, *68*(10), 1742–1747. doi:10.1021/ac951091z

Walczak, B., & Massart, D. L. (1997). Wavelet packet transform applied to a set of signals: A new approach to the best-basis selection. *Chemometrics and Intelligent Laboratory Systems*, *38*(1), 39–50. doi:10.1016/S0169-7439(97)00050-6

Wentzell, P. D. & Brown, C. D. (2000). Signal processing in analytical chemistry. *Encyclopedia of Analytical Chemistry.* 10.1002/9780470027318.a5207

WildFly. (n.d.). In *Wikipedia.* Retrieved on January 24, 2017 from https://en.wikipedia.org/wiki/WildFly

Worley, B. (2015). *Chemometric and Bioinformatic Analyses of Cellular Biochemistry* (Dissertation). University of Nebraska-Lincoln. Retrieved on January 25, 2017 from http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1064&context=chemistrydiss

Zaharia, M., Chowdhury, M., Franlin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster Computing with Working Sets. *HotCloud 2010 Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*. Retrieved on January 24, 2017 from https://people.csail.mit.edu/matei/papers/2010/hotcloud_spark.pdf

Zhong, Z., Tang, S., Peng, G., & Zhang, Y. (2016). A novel quantitative spectral analysis method based on parallel BP neural network for dissolved gas in transformer oil. *Power and Energy Engineering Conference (APPEEC), 2016 IEEE PES Asia-Pacific.* Retrieved on January 25, 2017 from http://ieeexplore.ieee.org/abstract/document/7779840/?reload=true

Zhu, X., Li, S., Shan, Y., Zhang, Z., Li, G., Su, D., & Liu, F. (2010). Detection of adulterants such as sweeteners materials in honey using near-infrared spectroscopy and chemometrics. *Journal of Food Engineering*, *101*(1), 92–97. doi:10.1016/j.jfoodeng.2010.06.014

ADDITIONAL READING

Akerkar, R. (2012). Improving data quality on big and high-dimensional data. *Journal of Bioinformatics and Intelligent Control*, *1*(2), 155–162. doi:10.1166/jbic.2013.1017

Bylesjö, M., Eriksson, D., Kusano, M., Moritz, T., & Trygg, J. (2007). Data integration in plant biology: The O2PLS method for combined modeling of transcript and metabolite data. *The Plant Journal*, *52*(6), 1181–1191. doi:10.1111/j.1365-313X.2007.03293.x PMID:17931352

Camacho, J. (2014). Visualizing big data with compressed score plots: Approach and research challenges. *Chemometrics and Intelligent Laboratory Systems*, *135*, 110–125. doi:10.1016/j.chemolab.2014.04.011

Camacho, J., Pérez-Villegas, A., Rodríguez-Gómez, R. A., & Jiménez-Mañas, E. (2015). Multivariate Exploratory Data Analysis (MEDA) Toolbox for Matlab. *Chemometrics and Intelligent Laboratory Systems*, *143*, 49–57. doi:10.1016/j.chemolab.2015.02.016

Cichocki, A. (2014). Era of big data processing: A new approach via tensor networks and tensor decompositions. Cornell University Library. Ithaca, NY, USA. Retrieved on October 22, 2016 from http://arxiv.org/abs/1403.2048

De Juan, A., & Tauler, R. (2003). Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. *Analytica Chimica Acta*, *500*(1–2), 195–210. doi:10.1016/S0003-2670(03)00724-4

Eriksson, L., Antti, H., Gottfries, J., Holmes, E., Johansson, E., Lindgren, F., ... Wold, S. (2004). Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Analytical and Bioanalytical Chemistry*, *380*(3), 419–429. doi:10.100700216-004-2783-y PMID:15448969

Eriksson, L., Trygg, J., & Wold, S. (2014). A chemometrics toolbox based on projections and latent variables. *Journal of Chemometrics*, *28*(5), 332–346. doi:10.1002/cem.2581

Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D., & Varnek, A. (2015). Chemical data visualization and analysis with incremental generative topographic mapping: Big data challenge. *Journal of Chemical Information and Modeling*, *55*(1), 84–94. doi:10.1021/ci500575y PMID:25423612

Geladi, P., Manley, M., & Lestander, T. (2003). Scatter plotting in multivariate data analysis. *Journal of Chemometrics*, *17*(8-9), 503–511. doi:10.1002/cem.814

Huma, L. (2010). *Chemoinformatics and advanced machine learning perspectives: complex computational methods and collaborative techniques: complex computational methods and collaborative techniques*. Idea Group Inc (IGI). Hershey, PA, USA. Retrieved on October 23, 2016 from https://books.google.com/books?hl=en&lr=&id=vWW-AQAAQBAJ&oi=fnd&pg =PA145&dq=big+data+chemometrics&ots=gDhetQw2mp&sig=0Ozc8nbQR85CgVQYpMPC51FQnEI#v=onepage&q&f=false

Kalinin, S. V., Sumpter, B. G., & Archibald, R. K. (2015). Big-deep-smart data in imaging for guiding materials design. *Nature Materials*, *14*(10), 973–980. doi:10.1038/nmat4395 PMID:26395941

Lohninger, H. (1994). INSPECT: A program system to visualize and interpret chemical data. *Chemometrics and Intelligent Laboratory Systems*, *22*(1), 147–153. doi:10.1016/0169-7439(93)E0054-8

McIntyre, R. S., Cha, D. S., Jerrell, J. M., Swardfager, W., Kim, R. D., Costa, L. G., ... Alsuwaidan, M. (2014). Advancing biomarker research: Utilizing "Big Data" approaches for the characterization and prevention of bipolar disorder. *Bipolar Disorders*, *16*(5), 531–547. doi:10.1111/bdi.12162 PMID:24330342

Megahed, F. M., & Jones-Farmer, L. A. (2015). Statistical perspectives on "big data." In S. Knoth & W. Schmid (Eds.), Frontiers in Statistical Quality Control 11 (pp. 29–47). Springer International Publishing. Retrieved on October 23, 2016 from http://link.springer.com/ chapter/10.1007/978-3-319-12355-4_3

Nikolov, M., Simeonova, P., & Simeonov, V. (2009). Chemometrics as an option to assess clinical data from diabetes mellitus type 2 patients. *Open Medicine: a Peer-Reviewed, Independent, Open-Access Journal*, *4*(4), 433–443. doi:10.247811536-009-0059-9

Offroy, M., & Duponchel, L. (2016). Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Analytica Chimica Acta*, *910*, 1–11. doi:10.1016/j.aca.2015.12.037 PMID:26873463

Pierce, K. M., Hoggard, J. C., Mohler, R. E., & Synovec, R. E. (2008). Recent advancements in comprehensive two-dimensional separations with chemometrics. *Journal of Chromatography. A*, *1184*(1–2), 341–352. doi:10.1016/j.chroma.2007.07.059 PMID:17697686

Qin, S. J. (2014). Process data analytics in the era of big data. *AIChE Journal. American Institute of Chemical Engineers*, *60*(9), 3092–3100. doi:10.1002/aic.14523

Tsai, C.-W., Yang, Y.-L., Chiang, M.-C., & Yang, C.-S. (2014). intelligent big data analysis: A review. *International Journal of Big Data Intelligence*, *1*(4), 181–191. doi:10.1504/IJBDI.2014.066957

Viceconti, M., Hunter, P., & Hose, R. (2015). Big Data, Big Knowledge: Big Data for Personalized Healthcare. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1209–1215. doi:10.1109/JBHI.2015.2406883 PMID:26218867

Wehrens, R. (2011). *Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Springer Science & Business Media. New York City, NY, USA. Retrieved on October 23, 2016 from https://books.google.com/books?hl=en&lr=&id=0jdNQ_Xrhu8C&oi=fnd&pg=PR3&dq=big+data+chemometrics&ots=UVu6IqjZ3k&sig=ibHW01vKkP6nvj2jC2iSE7R57Rk#v=onepage&q&f=false

Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E. J., Edlund, U., Shockcor, J. P., ... Trygg, J. (2008). Visualization of GC/TOF-MS-Based metabolomics data for identification of biochemically interesting compounds using opls class models. *Analytical Chemistry*, *80*(1), 115–122. doi:10.1021/ac0713510 PMID:18027910

Wold, S. (1984). Multivariate data analysis in chemistry. In B. R. Kowalski (Ed.), *Chemometrics* (pp. 17–95). Springer. The Netherlands. Retrieved on October 23, 2016 from http://link.springer.com/chapter/10.1007/978-94-017-1026-8_2

KEY TERMS AND DEFINITIONS

Big Data Approach: An approach that involves managing Big Data from different sources or databases.

Chemometrics: A branch of Analytical Chemistry that deals with the utilization of multivariate statistical techniques to come up with meaningful information about the data.

Continuous Wavelet Transform: Uses inner products to measure the similarity between a signal and an analyzing function.

Discrete Wavelet Transform: Is an implementation of the Wavelet Transform using a discrete set of the wavelet scales and translations obeying some defined rules.

Factor Analysis: A process in which the values of observed data are expressed as functions of a number of possible causes in order to find which are the most important.

Generalized Linear Model: Is a flexible generalization of ordinary Linear Regression that allows for response variables that have error distribution models other than a normal distribution.

Hierarchical Cluster Analysis: Is a method of cluster analysis which seeks to build a hierarchy of clusters.

Linear Discriminant Analysis: Is a generalization of Fisher's Linear Discriminant, a method used in Statistics, Pattern Recognition and Machine Learning to find a linear combination of features that characterizes or separates two or more classes of objects or events.

Process Analytical Technology: Has been defined by the United States Food and Drug Administration as a mechanism to design, analyze and control pharmaceutical manufacturing processes through the measurement of Critical Process Parameters which affect Critical Quality Attributes.

Principal Component Analysis: Is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

Principal Component Regression: Constructs new predictor variables, known as components, as linear combinations of the original predictor variables by creating components to explain the observed variability in the predictor variables, without considering the response variable at all.