# CHAPTER II
# THEORY

## 2.1 Analytical Part

### 2.1.1 Solid Phase Extraction (SPE) [23, 24]

Solid-phase extraction is a non-equilibrium sample preparation technique. The principle of this technique is an exhaustive removal of chemical constituents from a flowing liquid sample by retention on a solid sorbent and subsequent recovery of selected constituents by elution from the sorbent [23]. SPE is used in analytical laboratory to clean up samples and/or concentrate. Furthermore, SPE can be used to isolate analytes of interest from a wide variety of matrices; including urine, blood, water, beverages, soil, and animal tissue etc. There are four typical steps of SPE which consist of conditioning, loading, washing and elution (Figure 2). The conditioning step is for making the sorbent compatible with sample solution for close contact in small channels and sorbent should not be dry at any stage. Loading step is required gentle vacuum (or pump) at appropriate rate depend on column dimension and particle size. Washing step is for removal of interferences co-adsorbed. The wash solution must be finely chosen to avoid eluting the compounds of interest. The final step requires an eluting solvent which remove adsorbed analytes from the sorbent. If possible, the minimum volume has to be used to avoid diluting the analytes. Furthermore, eluting solvent should be compatible with analytical method (i.e. low boiling point), free from impurity and non-toxic. The advantages of SPE method over other methods (protein precipitation and liquid-liquid extraction) are small sample volume requirement, the use of small volumes of toxic organic solvents and this technique can be automated for routine analysis.

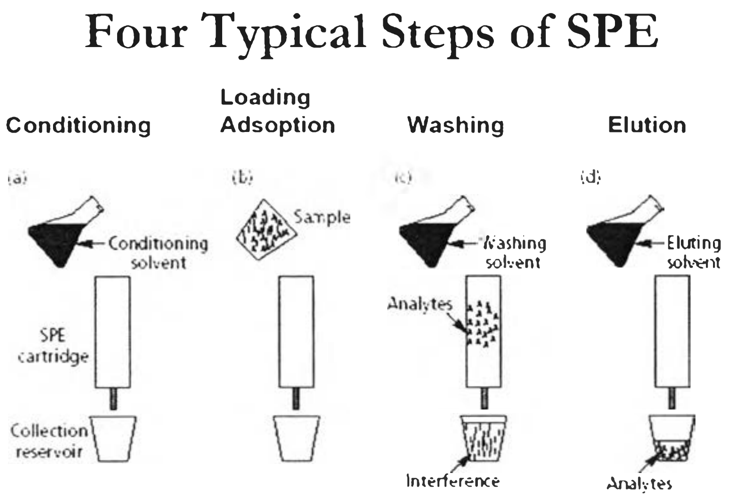

Figure 2 Four typical steps of SPE

2.1.2 Liquid Chromatography (LC) [25-27]

Liquid chromatography (LC) is a separation technique that can be used for the analysis of ions or molecules which are dissolved in a solvent. LC is based on mechanisms of adsorption, partition and ion exchange. The mechanisms depend on the type of stationary phase (SP) used. SP used in LC is solid which is normally packed inside a column. The mobile phase (MP) is the liquid pumped through the column. The sample to be separated is injected into the flowing MP by an injector. When the MP passes through the column, the molecules that SP adsorb most will transfer slowly through the column. When the MP has passed through the column it goes into the detector that detects the different molecules as they have pass through it. Figure 3 presents schematic diagram of LC instrument. The separation of the components result from the difference in the relative distribution ratios of each analytes between the two phases (SP and MP).

2.1.2.1 Instrumental Part of LC System

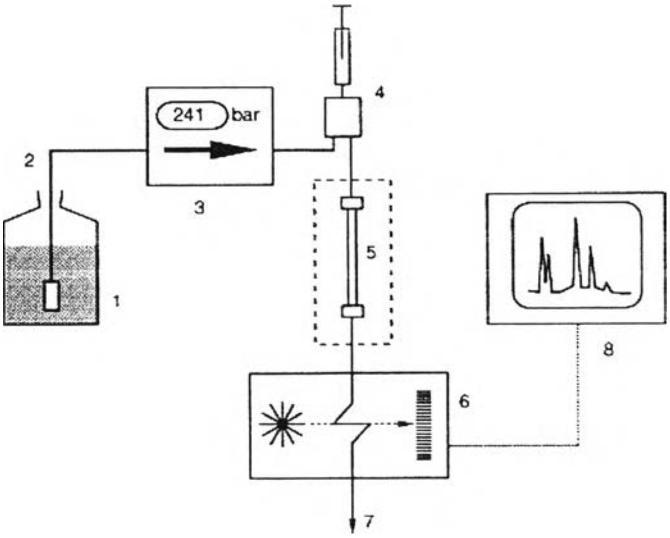

Figure 3 Schematic diagram of a LC instrument; 1=reservoir; 2=transfer line; 3=pump; 4=sample injection; 5=column (with thermostat); 6=detector; 7=waste; 8=data acquisition [25]

2.1.2.2 Type of Elution of MP

MP is usually composed of two or more solvents implemented or not by a modifier (i.e. acid or base). The elution efficiency depends on the MP composition and its interaction between analytes and column. There are two elution types for a bioanalytical separation:

Isocratic elution is an elution type which maintains the solvent compositions at a constant mixing ratio during the entire run.

Gradient elution is an elution type which changes the composition of the MP during the experiment. This type is appropriate for separation of complex mixtures and compounds with different solubility properties.

#### 2.1.2.3 LC Separation Mode

##### 2.1.2.3.1 Normal Phase Liquid Chromatography (NPLC)

NPLC is also called adsorption chromatography. In this mode, the SP is more polarity than the MP. The SP includes polar boned phase or bare silica such as cyano, amino and diol boned phases. MP used is an organic solvent (i.e. hexane and ethyl acetate), furthermore water does not use in this mode. In this mode, polar compounds will be eluted slowly and usually show long retention times. The separation mechanism of NPLC is based on polar adsorption or hydrogen boning or dipole-dipole interaction. NPLC is traditionally used for the separation of mixtures of isomers and polar compounds.

##### 2.1.2.3.2 Reversed Phase Liquid Chromatography (RPLC)

RPLC is commonly used for bioanalysis of less polar analytes solubilised in aqueous matrices. In this mode, the MP is more polarity than SP. The SP is octadecylsilyl (C18) or octyl silyl (C8) and it is hydrophobic and chemically bonded to surface of silica supported particle. The MP used is often water, buffers and organic solvent (i.e. methanol (MeOH) and acetonitrile (MeCN)). In this mode, non polar compounds will be eluted slowly and have longer retention time. The separation mechanism of RPLC is based on partition. RPLC is widely used to separate non polar to neutral polar compounds with low molecular weights (< 2000 Daltons). Furthermore, weak acids and bases and proteins/peptides can be separated in this mode.

##### 2.1.2.3.3 Ion Exchange Chromatography (IEC)

The SP used in this mode is a strong cation exchanger (SCX; sulfonates), a strong anion exchanger (SAX; quaternary ammonium), a weak cation exchanger (WCX; carboxymethyl) or a weak anion exchanger (WAX; diethylaminoethyl). MP used is an aqueous solution of a salt with buffer and a counterion. The separation mechanism of IEC is based on ion exchange. IEC is used for separation of ionic or polar analytes.

2.1.2.3.4 Hydrophilic Interaction Liquid Chromatography (HILIC)

HILIC is also called aqueous normal phase chromatography. This mode is a combination of NPLC, RPLC and IEC in term of adsorbent, eluent and analyte, respectively [28]. The SP is a polar bonded phase (i.e. silica, diol, amino, cation, anion and zwitterionic bonded phase) see Figure 4. The MP is composed by high ratio of organic solvent (80 to 96%) with a small amount of aqueous or polar solvent which is a strong solvent. HILIC was designed for eluting polar compounds. The advantages of HILIC over conventional RP are the retaining and separating ability for small polar analytes. Furthermore, using MP containing of high organic solvent content can enhance ESI-MS sensitivity compare with IEC which using MP with aqueous solution containing salt with buffer and/or a counterion. The separation mechanism of HILIC is based on a partition of analytes between MP and the water-enriched layer of SP and/or electrostatic (ionic) interaction with either positive or negative charges (Figure 5). Nowadays, HILIC is used for separation of small polar and hydrophilic compounds, especially those with a low molecular weight (MW < 1,000 Da).

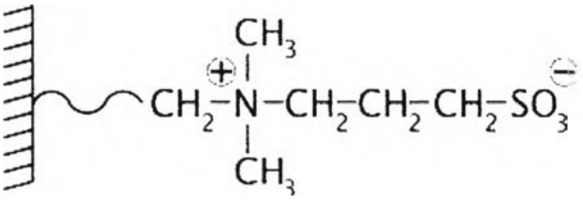

Figure 4 ZIC HILIC phase, SP used in this study which consists of porous silica particles covalently bonded with highly polar sulfobetaine type zwitterionic functional groups [29]

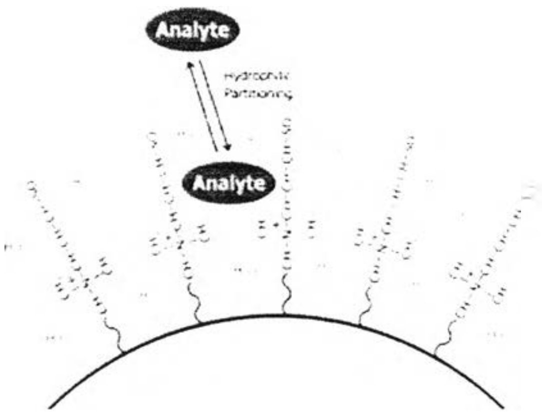

Figure 5 The separation mechanism of HILIC on the ZIC-HILIC column [30]

2.1.3 Mass Spectrometry (MS) [31-33]

MS is a powerful analytical technique which be used to quantify known compounds and/or to identify unknown structures. Furthermore, elucidation of the structure and chemical properties of different molecules could be performed by MS. MS process involves the conversion of a sample (i.e. solution) into gaseous ions with or without fragmentation and then characterization by their mass to charge ratios (m/z) and relative abundances. Ion source generates multiple ions from the samples and then mass analyzer separates them according to their specific m/z. The ion detector records the relative abundance of each ion type. Finally, mass spectrum of the molecule is produced and presented in the form of ion abundance (%) versus m/z plot. Figure 6 shows the example of mass spectrum of hexanal and Figure 7 presents a schematic diagram of MS instrument.

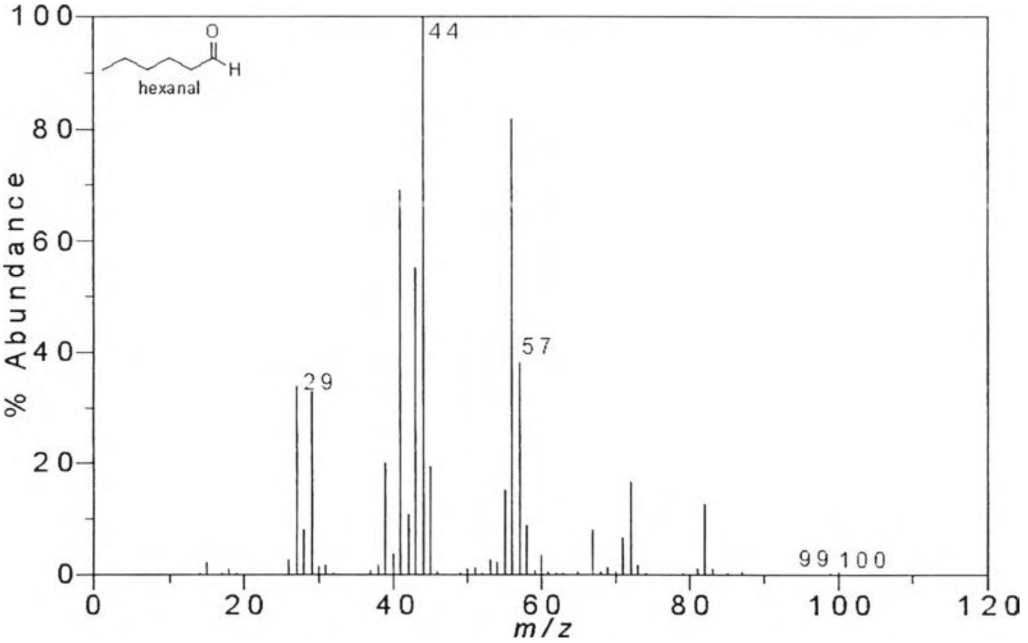

Figure 6 Mass spectrum of hexanal [34]

2.1.3.1 Instrumental Part of MS System

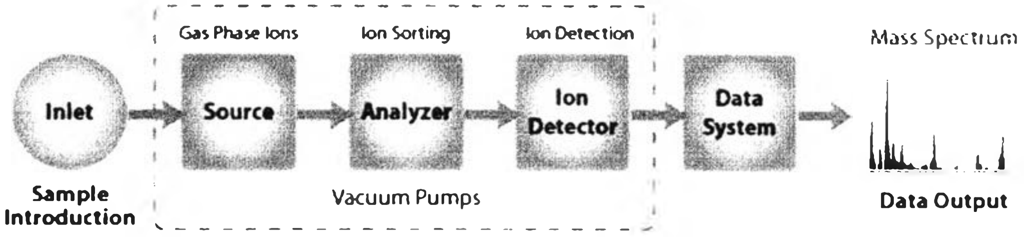

Figure 7 Schematic diagram of a MS instrument [32]

2.1.3.2 Electrospray Ionization (ESI) [35, 36]

ESI is an atmospheric pressure and soft ionization technique appropriate mostly for polar compounds and large proteins. ESI can be used in both positive and negative modes. ESI uses electrical energy to assist the transfer of ions from solution into gaseous phase without fragmentation before going to MS analysis.

The analyte is introduced in the source in solution either from a syringe pump or as the eluent flow from LC. The solution flow passes through the electrospray needle on which is applied a high potential difference (typically in the range from 2.5 to 4 kV). This forces the spraying of charged droplets from the needle with a surface charge of the same polarity to the charge on the needle. The droplets are repelled from the needle towards the source sampling cone on the counter electrode. Finally, the droplets pass through the space between the needle tip and the cone, and then solvent evaporation occurs (Figure 8).

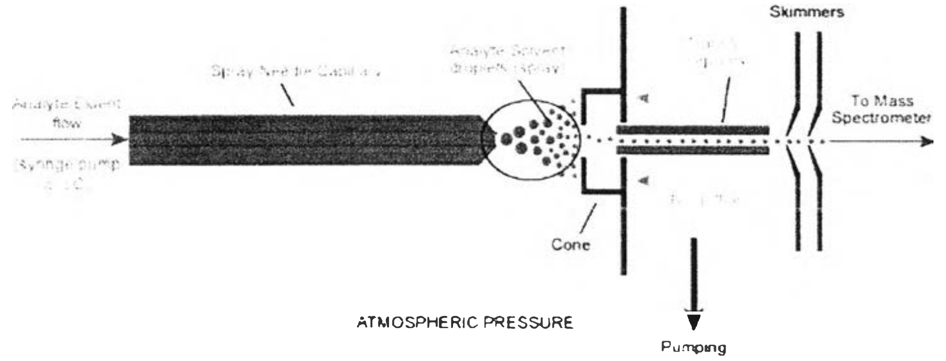

Figure 8 Schematic of the components of an ESI source [36]

2.1.3.3 Quadrupole Ion Trap (QIT) Mass Analyzer [33, 36]

QIT is an ion trapping device which uses a three dimension quadrupole field for trapping and analyzing ions. An evacuated cavity consists of two end cap electrodes spaced by ring electrodes which connect to direct current (DC) and radio frequency (RF) potentials for ion oscillation in QIT. Ions which are generated from external ion source (i.e. ESI) or inside the cell are transported to QIT. After that analyzed ions are stored in QIT by optimizing DC and RF component to accumulate analyzed ions in stable region. Ion are scanned (by ramping RF voltage) and ejected out to detector sequentially with one m/z at a time (Figure 9). Characteristics of QIT are high full mass scan sensitivity and the ability to perform tandem MS (MSn) by tandem MS in time.

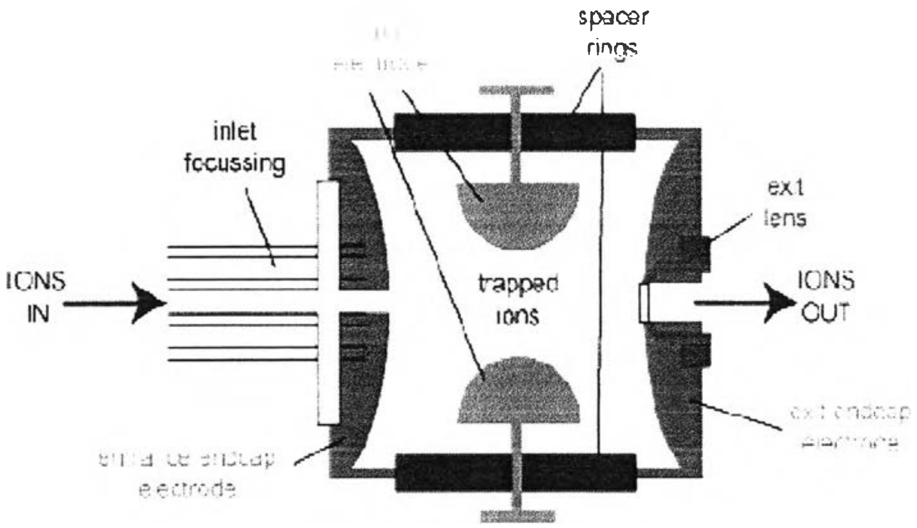

Figure 9 Schematic of the QIT mass analyzer [36]

### 2.1.3.4 Tandem Mass Spectrometry [33, 37]

Tandem MS or (MS/MS) is two or more MS systems coupled together for a simultaneous operation. Multiple stages of mass analysis separation can be achieved with individual mass spectrometer elements separated in space (tandem in space) or using a single mass spectrometer with the MS steps separated in time (tandem in time).

Tandem MS in space, molecular ions of all compounds in the mixture are generated and separated in the first MS (Q1). Individual molecular ion is then fragmented or dissociated in collision induced dissociation (CID) which is Q2. Finally, the fragment ions are detected and analyzed in the second MS (Q3) for identification of each component in the mixture (Figure 10). CID uses high energy and inert collision gas (i.e. Ar) to induce fragmentation of selected parent ion (precursor ion). Triple quadrupole (QQQ) is one of the MS instrumentation which separates ions in space (tandem MS in space) in Figure 10.

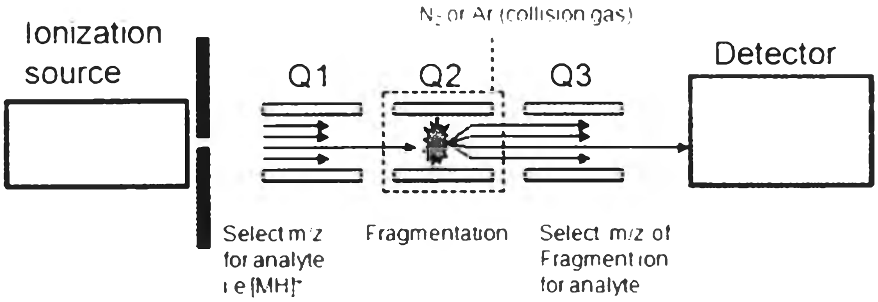
N₂ or Ar (collision gas)
Ionization source
Q1
Q2
Q3
Detector
Select m/z for analyte i.e [MH]⁺
Fragmentation
Select m/z of Fragment ion for analyte

Figure 10 Schematic of the QQQ mass spectrometer [37]

Tandem MS in time, QIT is one of MS instrumentation which has ability to do tandem MS ($MS^n$) without additional equipment (standalone tandem MS). There are five steps of QIT to perform tandem MS in time. The first step, parent ions are generated by soft ionization and then only selected ions are trapped by QIT and other ions are ejected in the second step. The third step, collision gas is added into QIT cell and RF is applied to end caps to fragment and induce motion of selected ions. In the fourth step, tandem MS scan fragmented ions of selected component for analysis. The final (5th) step, repeating 2nd to 4th step is performed for more tandem analysis (more selected daughter ions). Furthermore, QIT has a high full scan sensitivity since there is no ion lost (all ions are trapped during the collision process).

## 2.2 Pattern Recognition Part [13, 14]

### 2.2.1 Data Pre-Processing

Data pre-processing is the most important step to be considered before data analysis step. It is the step which transforming raw data achieved by analytical techniques into simple format to be processed by multivariate techniques (in this thesis used pattern recognition) for exploration, classification and quantification. The general input format for pattern recognition modelling is a data matrix (X) which consists of samples (row) and a set of variables (column) Figure 11.

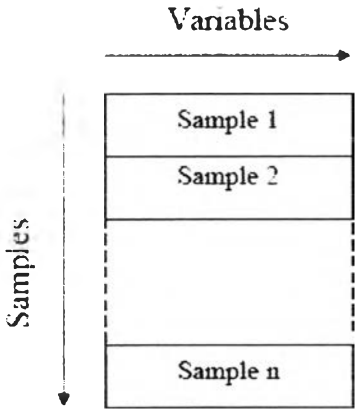

Figure 11 Data matrix used for pattern recognition modeling

Significant differences in the results can be achieved by using different pre-processing methods. The natural variations occur from several factors (i.e. different experimental date, instrumental condition, etc.). Therefore, the raw data should be transformed to reduce the variation before processing the pattern recognition by various pre-processing methods. There are three different steps (in order) to scale the data matrix (X). The first step is transforming the elements of the matrix, the second step is row scaling and the final step is column scaling. It is very important to perform the pre-processing steps in the correct order to obtain a correct result.

2.2.1.1 Transforming The Elements of The Matrix

The informative data (peaks) from biological samples (i.e. plasma and urine) are not naturally the most intense peaks and usually be small data (peaks) with high variation. Therefore, scaling the data is one of the important methods used to reduce the influence of large data (peaks). In this step, there are two methods used which are Logarithmic scaling and Power transformation.

1. Logarithmic scaling as a common equation; $^{l}x_{ij} = Log_{10}(x_{ij})$.

This section considers only data scaling using log to the base 10.

2. Power transformation as an equation; $^{n\_root}x_{ij} = (x_{ij})^{1/n}$

Where i = row (sample), j = column (variable) and n = order of square root. The limitations of Log scaling method are`the elements with zero values could not be defined and the linear relationships between variables could be eliminated by this method. In addition, power transformation method could define the elements with zero values and large nth roots could be used to reduce for further, however it cannot be used if the data matrix contains negative elements.

### 2.2.1.2 Row Scaling

It is one of an important step to prepare a data matrix before performing any data analysis methods, especially when the numbers of samples to be analysed is difficult to control. The main objective of row scaling is to make data of samples (i.e. amount, concentration, intensity) comparable. One of the most popular and simple methods is to scale each sample's element to the quantity of an internal standard (IS). Therefore, considering the accuracy of an IS should be taken. In addition, an alternative row scaling method is to scale the rows to a constant total. It is appropriate for the case that the amount of sample is not difficult to control. This is also called "Normalisation" (as an equation below) and it transforms the elements of the data matrix for proportions instead of absolute quantities.

$${}^{nr}x_{ij} = \frac{x_{ij}}{\sum_{j=1}^{J} x_{ij}}$$

The summation of intensities for all variables in each sample is equal to 1.

### 2.2.1.3 Column Scaling

This final step is very important if variables are on different scales and it is required to ensure that all variables have equal power in the analysis.

There are three methods in this step.

1. Mean centring is a common method for column scaling often included in chemometric software packages as an equation; ${}^{mean}x_{ij} = x_{ij} - \bar{x}_j$

Where $\bar{x}_j$ is the column mean of variable j.

This method involves subtracting the column mean for each variable and it can be useful if different variables have different means.

2. Weighted centring is very useful if the number of samples is very different in each class as an equation; $\bar{\bar{x}}_j = \sum_{g=1}^{G} \bar{x}_{g,j} / G$

Where G = total classes of data, $\bar{\bar{x}}_j$ = the global mean of variable j.

This method is not limited to binary class problems and it can be extended for datasets with several classes.

3. Standardisation is an extension of mean centring by subtracting the mean of each variable and then dividing by the population standard deviation as an equation;

$$^{\text{std}}x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^{I} (x_{ij} - \bar{x}_j)^2 / I}}$$

This step is very important if used for measuring different variables over different ranges. This method makes variation of all variables in the data matrix on an equal scale to prevent controlling outcome by only the most intense peaks. In case, the mean value is equal to zero and the standard deviation is one that means all variables are given equal importance for data analysis.

### 2.2.2 Unsupervised Pattern Recognition

Unsupervised pattern recognition methods are also called "exploratory techniques". The objective of this technique is to visualise the underlying information of the relationship between samples and variables in data matrix without prior knowledge (i.e. class information) required. This technique exposes the main patterns or groups of observations in the data and also identifies outliers. All unsupervised pattern recognition methods are based on the same concept which is the reduction of the number of variables into a small number of latent variables for simple analysis by visual inspection.

### 2.2.2.1 Principal Component Analysis (PCA)

PCA is the most commonly used method for data reduction and visualization. The objective of this method is to find the patterns, reduce the dimension of data for the most important parts and eliminate noise simultaneously.

This method is based on the hypothesis that several variables in a dataset are correlated and some variables present significant variance when compare with others. Therefore, the main trends of the data can be summarized by a few latent variables which is called "Principal Components (PC)". Normally, the original data matrix is projected by using PCA into PC space which ensures that the projection of the most significant variance of the data matrix is onto the first PC. PCA can identify useful information from the data by using only a few PCs. The significance of each PC can be determined by using the percentage variance represented by the PC. The PC which gives the largest percent variance will be assigned as the first PC and the PC with the second largest as the second PC, respectively. The application of PCA to datasets provides two quantitative vectors for each PC, which are the scores and loadings vectors.

In PCA, data matrix (X) is decomposed into PCs consisting of a scores matrix (T), a loadings matrix (P) and a residue matrix (E) as an equation; X = T P' + E, where data matrix contains I samples and J variables. Graphical example presents the decomposition of data matrix (X) into a scores matrix (T), a loadings matrix (P) with A principal components (PCs) in Figure 12.

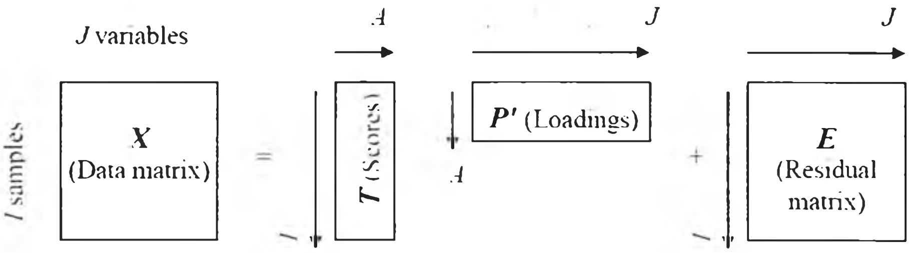

Figure 12 Graphical example presents the decomposition of data matrix (X) into a scores matrix (T), a loadings matrix (P') with A principal components [13]

The scores vector includes coordinates of each sample projected into the PC space and gives a visual image of the difference between samples.

Samples with similar properties will be aligned in the same region of the plot and away from samples with different properties. The loadings vector allows the contribution of each variable to be evaluated. The scores and loadings for each PC are orthogonal and independent to each other. Figure 13 presents graphical example of PC of data, the main variance is along PC1 and PC2 is orthogonal to PC1.

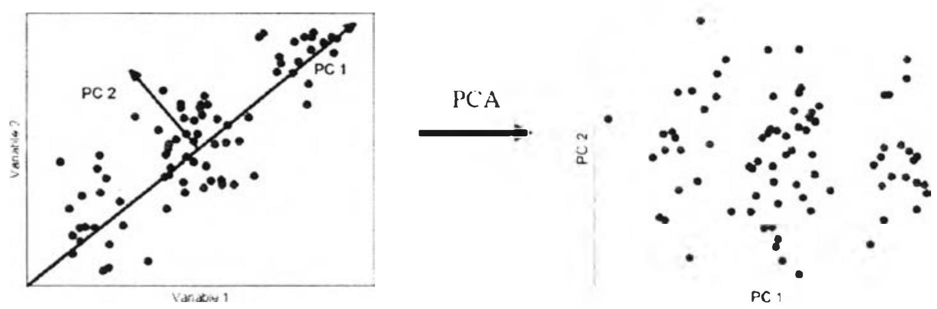

Figure 13 Graphical example illustrates of samples in the original space (left) and the position of samples in PC spaces (right) [13]

### 2.2.3 Supervised Pattern Recognition

The objective of supervised pattern recognition is to classify groups of samples for predicting the class membership of unknown samples. This technique uses the mathematical models (also called "classier") of known sample classes to create boundaries between each class and maximize the separation of them (i.e. patients vs. healthy individuals). In this thesis, Partial Least Squares Discriminant Analysis (PLSDA) and Linear Discriminant Analysis (LDA) techniques are used and explained for more detail below.

#### 2.2.3.1 Partial Least Squares Discriminant Analysis (PLSDA)

PLSDA is one of the most common supervised linear modelling techniques in the field of chemometrics. The advantage of PLSDA is flexibility for the case which the numbers of variables are very much more than the numbers of samples. The discriminatory power of PLSDA is similar to linear discriminant analysis

(LDA) when apply with appropriate data scaling (data pre-processing). PLSDA is a regression method which is processed by projecting the original data into latent variable space. The principal of PLSDA is similar to PCA method since it aims to find the best latent variables to represent the data. However, PLSDA is a method which constructs a regression model containing a set of regression coefficients that describes the relationship between data matrix and a class membership matrix while PCA considers only data matrix.

PLS is generally considered only for prediction of two class and thus it is used in this thesis as equations; X = T P' + E and c = T q + f
Where X is experimental data matrix, T is the scores, P' and q correspond to the loadings, E and f correspond to the residue, c is a classifier.

Furthermore,T and P' in PLSDA are different from T and P' in PCA. For two class classification, the value of elements in the c vector is set to +1 (Class A) or -1 (Class B) correspond to the class that a sample is in. The application of regression coefficients in PLSDA to unknown samples causes the prediction of the class membership of the unknown samples.

2.2.3.2 Linear Discriminant Analysis (LDA)

LDA is a method commonly used to calculate discriminant functions as linear combinations of selected variables for separating the groups. The basic principal of LDA is to calculate the centroids of each data class g ($\bar{x}_g$).
The centroids are calculated from the mean of all samples from each of variables in a group and it is assumed that the distribution of samples around the centroid is symmetrical. The calculation of distance is also considered the pooled variance-covariance matrix ($\boldsymbol{S}_p$). The distance between samples to the class centroid is weighted according to the overall variance of each variable. Therefore, the correlation between variables is now considering. This technique is used a measure called "Mahalanobis" to calculate distance to class centroid g which is based on a variance covariance matrix for the whole data, rather than for each class separately.

As equations; $d_{ig} = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)\boldsymbol{S}_p^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)'}$ and $\boldsymbol{S}_p = \dfrac{\sum_{g=1}^{G}(I_g - 1)\boldsymbol{S}_g}{\sum_{g=1}^{G} I_g - 1}$

Where $\mathbf{S}_p$ is the pooled covariance matrix, Ig provides sample numbers in class g and Sg provides the variance-covariance matrix for group g. The limitation of this technique is the pooled covariance matrix $\mathbf{S}_p$ is only applicable when there are similar variance-covariance matrices for all classes. Furthermore, LDA method has high potential to classify samples more than two groups.