CHAPTER II

BACKGROUND KNOWLEDGE

This chapter provides background on the biochemical techniques used for rice classification, chemical expression in the experiment, and data preparation. The statistical were used to analyse the chemical data were described in previous literature reviews.

2.1 Biochemical Background

Rice is the seed of a monocot plant of the grass family [1], which is a major cereal crop in developing countries and an important staple food for over half of the world's population. White rice is the name given to milled rice which has had the husk, bran, and germ removed. *Oryza sativa* and *Oryza glaberrima* are two main rice species which are consumed by humans. *Oryza sativa* can be found in many areas including tropical Latin America, the West Indies, East, South and Southeast Asia, while *Oryza glaberrima* is only found in Africa. The family *Oryza sativa* has been classified in three subspecies; *Indica, Japonica and Javanica*, which are basically distinguished by habitat, example *Indica* occurs in tropical areas (Thailand, India), while *Japonica* mainly occurs in temperate areas (Japan, Korea).

Over the last decade, fragrant rice has been in great demand in Asian rice trading. This fragrant rice has also been widely accepted in several areas such as European countries, North America, and Hong Kong. It's price is more expensive than others varieties because it has volatile compounds which have are good smell especially after cooking. One of the main volatile compounds is 2-acetyl-1-pyrroline

(2AP) and has also been found in various other kinds of food including cooked shrimp, bread power (baguette crusts), toast wheat bread, canned sweet corn [15-18], and can be recognized with an aroma threshold value as low as 0.1 ppb in water [17]. This compound has also been reported to be present in the volatile extract of various processed foods [15-18]. Since food processing methods usually involve heating, the occurrence of 2AP in foods has been suggested to take place during cooking at elevated temperatures via a reaction between amino acids and carbohydrates, called the Maillard reaction [19, 20]. As a result, early quantitative studies concerning the 2AP analysis in foods frequently utilized a heat extraction, such as stream-distillation and solvent extraction (SDS), followed by the extract analysis using gas chromatography-mass spectrometry (GC-MS). Although there have been a number of studies on rice flavour chemistry, for many uniquely flavoured specialty rice types. Over 300 volatile compounds have been identified from various cultivars of fragrant and non-fragrant rice [21]. The quantity of the chemical compounds identified vary with several factors such as the milling processes [14, 22], cooking methods [17], storage duration and temperature of storage [14, 23, 24] and chemical extraction condition [17, 25, 26].

During storage, a number of physicochemical and physiological changes occur which are usually termed “ageing”. These changes include pasting properties; and changes to colour, flavour, and composition which can affect rice quality [24, 27-30]. Aged rice tends would be fluffier and harder after being cooked [27-29]. Pushpamma and Reddy [31] reported that an optimum cooking time for milled rice was 4-6 min for longer than 6 months storage duration after harvest. The evaluation techniques

have been used by several researchers to evaluate the effect of storage on the end-use quality of rice [32-34]. The storage conditions are important in the ageing process. Nitrogen was superior to air in preserving palatability of cooked rice during brown rice storage at 10 °C for 2 years [35]. No great difference in quality was found between the brown rice stored in nitrogen versus carbon dioxide [35]. The nitrogen storage conditions had little effect on the rice-texture changes while cooking related to air storage [28]. Hermetic storage of milled rice at 30 °C for 3 months under vacuum or in nitrogen, carbon dioxide, and air atmospheres had little effect on the quantity of the reducing sugar, fat acidity, texturometer hardness and adhesiveness of cooked rice at 14.7% storage moisture [36]. At 15.7% moisture storage, vacuum package showed the least changes in reducing sugar, acidity hardness and adhesiveness, followed by gas package and air package [35]. Perez and Juliano [28] found that fragrant rice preservation at 15 °C for the first 3 – 4 months effected the reduction of 2AP as statistically significant different.

As described in other studies, changes to the chemical compounds follow three basic patterns as follows.

1. The chemical compounds decrease when the sample is kept in the longer-storage durations for example 2-acetyl-1-pyrroline (2AP) [14].
2. Some chemical compounds decrease for some storage durations and then increase for other storage durations. The chemical compounds with this property could be due to oxidation and reduction reactions among the chemical compounds [23, 24, 31].
3. Some chemical compounds did not change or were only slightly changed in their chemical-compound quantities. There are fatty acids groups [23, 24].

2.2 Rice Variety Classification

There are a number of rice characteristics including colour, flavour and composition which affect the quality of different varieties [24, 27-30]. However, the characteristics can be changed by environment, postharvest methods and storage duration [24, 27-30]. Several publications [5, 7-9, 21-24, 28, 29, 32-36] compared and classified rice characteristics of different varieties. The examples of rice characteristic measured were grain colour, grain size, amylose content, aroma components, chemical profiles, and DNA sequences. Those techniques could be applied to compare the rice characteristics; for example, rice type as determined by grain size, amylose content percentage in order to classify plain or sticky rice [27, 37], and 2AP measurements to discriminate aromatic versus non-aromatic rice [9, 21]. In terms of the aroma component measurement, the chemical profile data could be measured by HS-GC, GC-MS, and the other methods of chemical fingerprint. Rice samples are practically suitable to only some techniques. Gas measurement is one of the most popular methods because it can precisely determine low molecular weights without any waste in the measurement process [13, 24]. DNA fingerprint technique is another modern method to analyse the nucleotide-base sequence [5, 8, 15]. However, the fingerprint cannot discriminate some rice conditions (rice ageing, storage condition) so it is not an appropriate method to apply for this research [13, 14, 22, 24, 26]. The chemical profile data were applied with several chemical extractions [6, 10-13, 18, 22-26]. These methods can detect chemical-quantity changes compared between different conditions (storage conditions, milling method, and storage temperature) in the same rice variety [14, 22-24, 27, 30, 36].

A gas extraction technique, called “headspace sampling”, has been reviewed as a rapid and efficient technique usually used with capillary GC for the volatile fractions analysis in many food samples [13, 38]. In regards to rice, this technique has successfully been applied to the volatile analysis in rice foliage [39] and in rice cake [6, 40]. The introduction of the handy headspace solid-phase microextraction (SPME) process allowed more costly traditional techniques to be replaced. Despite its widespread application, SPME has not been reported as a successfully analytical tool for the presence of 2AP in grain of fragrant rice. Its main limitation comes from the difficulties in obtaining valid samples due to poor extraction reproducibility [25]. Another approach related to headspace sampling called static headspace (SH), has also shared popularity in food and flavour research [41-43]. This technique has successfully been applied to both qualitative and quantitative approaches. With static headspace sampling, sample headspace volatiles are automatically brought directly to the GC, thus offering good validation as well as the possibility for a high number of samples to be processed.

2.3 Chemical Measurement

In this experiment, the samples of fragrant rice were randomly measured using the headspace gas chromatographic (HS-GC) technique to extract the chemical components. The HS-GC technique measures the volatile compounds in a closed vial containing various types of samples including extract solvent, or samples that needed to be directly analysed. Only volatile compounds–evaporated above the sample in the close vial can be analysed in the GC.

2.3.1 Headspace Analysis

The headspace contains the evaporated volatile components under each measurement condition or objective detail. The HS technique can measure gas, liquid and solid samples. Then the GC is used to test the resulting samples..

1. **Static Headspace** is the vapor, which is distributed in the space above the sample placed in the closed vial in each condition, and is then brought to be analysed using the GC technique or another suitable technique in the next steps.

2. **Dynamic Headspace** is the vapor, which is distributed in the space above the sample placed in the closed vial for each condition, and is absorbed by the adsorbent; Silica. The adsorbent emits volatile compounds, which were there taken to analyse with the appropriate technique in next step.

In this study, an Agilent G1888 headspace sampler was equipped with the Agilent 6890 GC as shown in Figure 2.1.

Figure 2.1 Headspace gas chromatography (HS-GC)

2.3.2 Headspace Sampler Instrumentation

1. The Oven (Figure 2.2) is the equipment machine used for heating and controlling the temperatures of the samples, as well as spinning and shaking the sample vials. There are 12 channels that load in the same timing duration. The oven can control the temperature between 40-200 °C.

Figure 2.2 Headspace oven

2. The Vial Tray (Figure 2.3) is used to hold the sample vials; there are 72 channels for sample loading. The samples are moved from the vial tray to the oven. This method can be automatically setup to follow the measurement of volatile compounds.

Figure 2.3 Vial tray

3. Sample Injector System (Figure 2.4) is the equipment part that transfers the gas sample from the vial to the GC system. There is a vial-sampling 0.5 mm needle, which is used to push nitrogen gas into the vial to increase the pressure. After that, the vapor is transferred to sample loops consisting of 2 sizes (1 and 3 mm) which can be selected for appropriate samples. The six-port valve is a vapor-sample controller in the vapor-leading system. There are injections with syringe, sample loop and six ports valve while the system is kept temperature controlled between 40-200 °C.

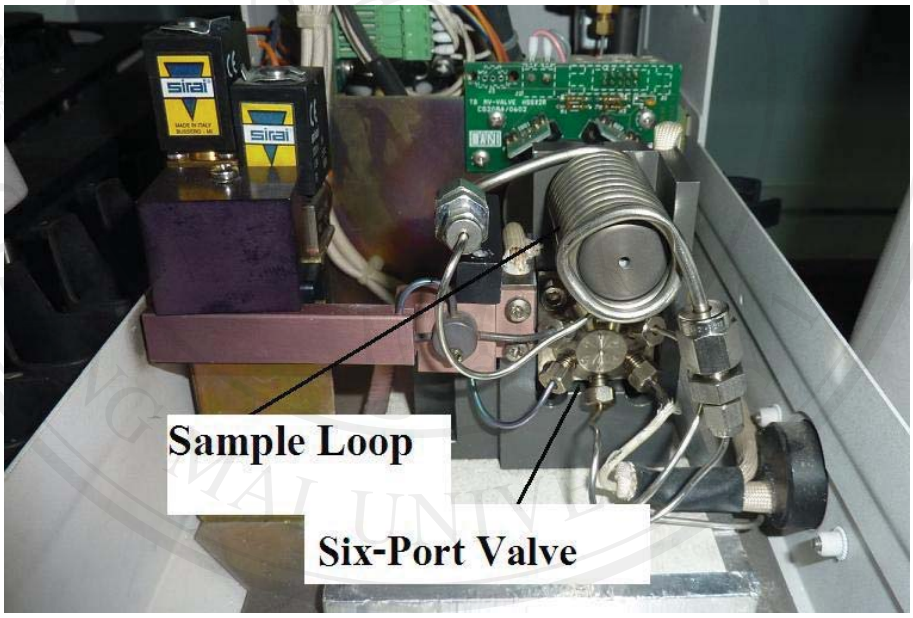

Figure 2.4 Sample injector equipments

4. The Transfer Line (Figure 2.5) is the transports the gas from the headspace sampler to the GC. This transfer line consists of a small Nickel tube in the heat coil which is covered with an insulator. This allows the temperature of volatile sample to be maintained between 40-220 °C.

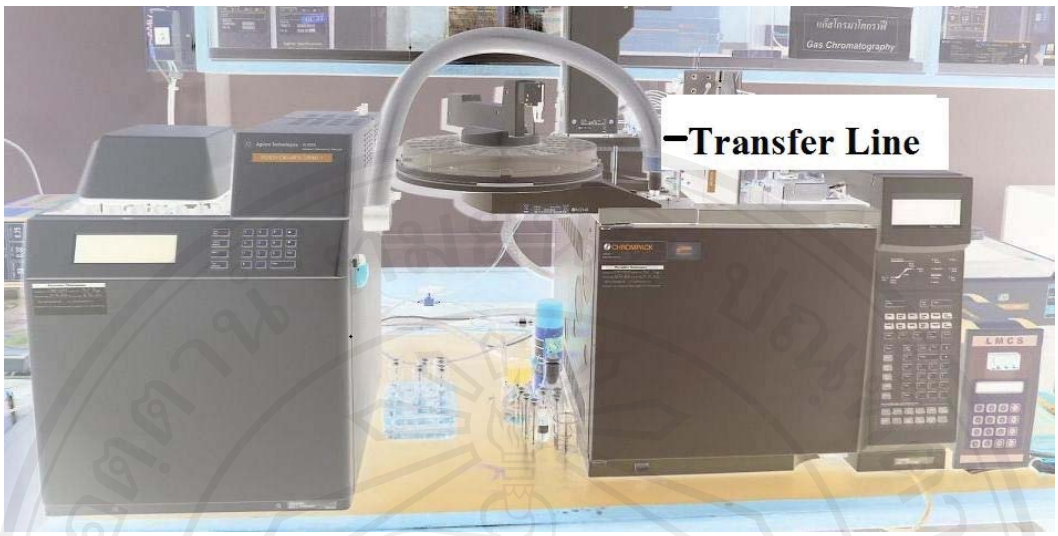

Figure 2.5 Transfer line

2.3.3 Gas Chromatography (GC)

The GC can analyse the volatile components in the mixtures which change to vapor at the proper temperature. The GC components (Figure 2.6) include 3 parts 1) carrier gas (bring the vapor samples from injector to GC column), 2) oven (including the injection port, column and detector inside), 3) data processing and storage unit.

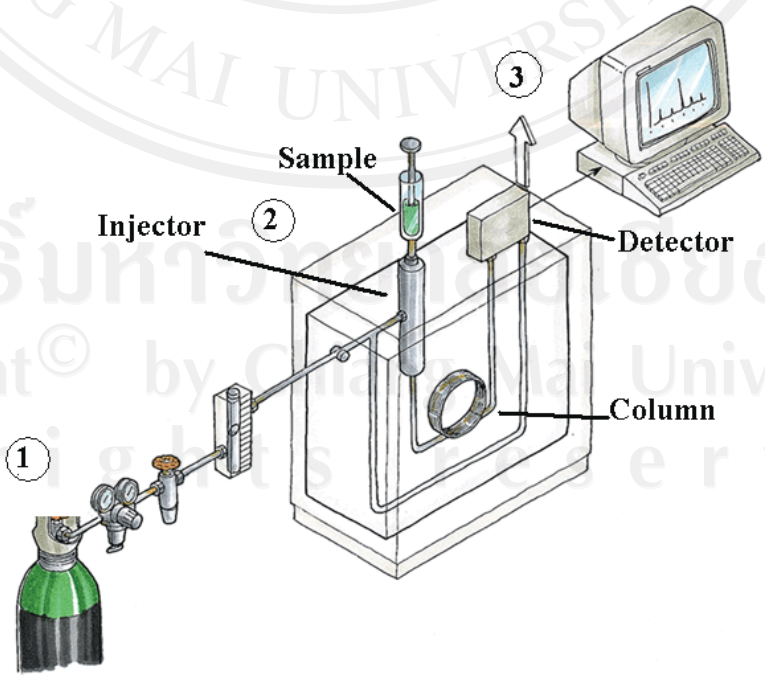

Figure 2.6 Gas chromatography components [44]

2.3.4 Separation Mechanism in GC

Every GC pattern is related to the distribution and partition of chemical molecules in the two different phases; stationary and mobile. The molecules are differently distributed or partitioned based on the relative solubility in each phase. The particular chemical compounds are used to extract the components (Figure 2.7) when the molecules have past the stationary phase, and are arrested with the different ratios and velocities. The components are spilt from the other components, which have a high potential to dissolve in the stationary phase; these components are slower than the low potential components.

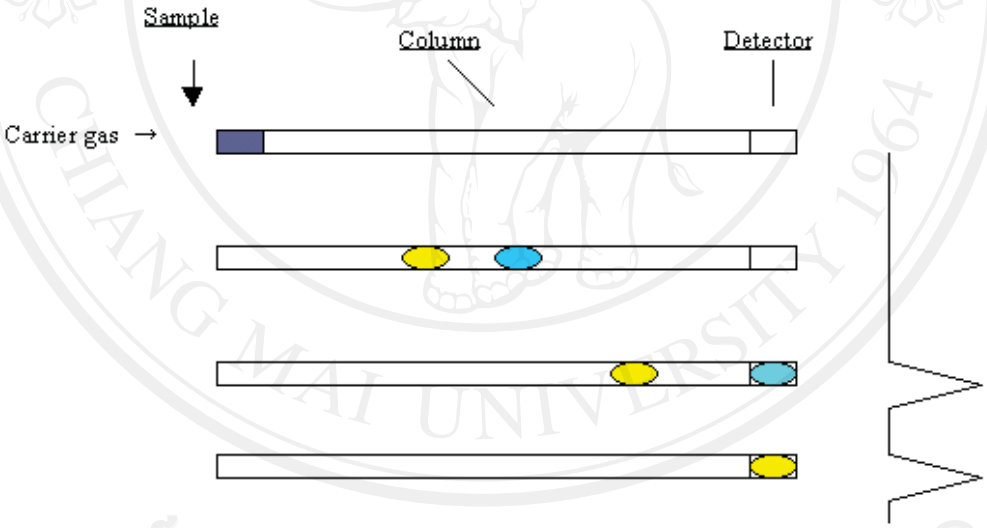

Figure 2.7 Chromatography [45]

The chemical-components were first extracted using the GC technique. First, the carrier gas from the compressed gas cylinder is introduced in the mobile phase controlled its pressure by a pressure regulator. Gas flows through the filter when the valve is opened. The moisture and other contaminants are eliminated before it reaches the temperature-controlled injector.

The samples are injected to the injection port by the syringe. The gas components in the injection port are subjected to higher temperature than their boiling points. After that, the gas sample, solvent and the vapor are flown to column by carrier gas. The chemical molecules in the gas phase are flown through the column and are separated from each others. Next, the molecules of each component are sent to the FID detector. The chemical signal is measured and sent to be processed and recorded in a computer. The results are plotted in a form of graph having the quantity signal (Y axis) with retention time (X axis) which called "chromatogram" (Figure 2.8).

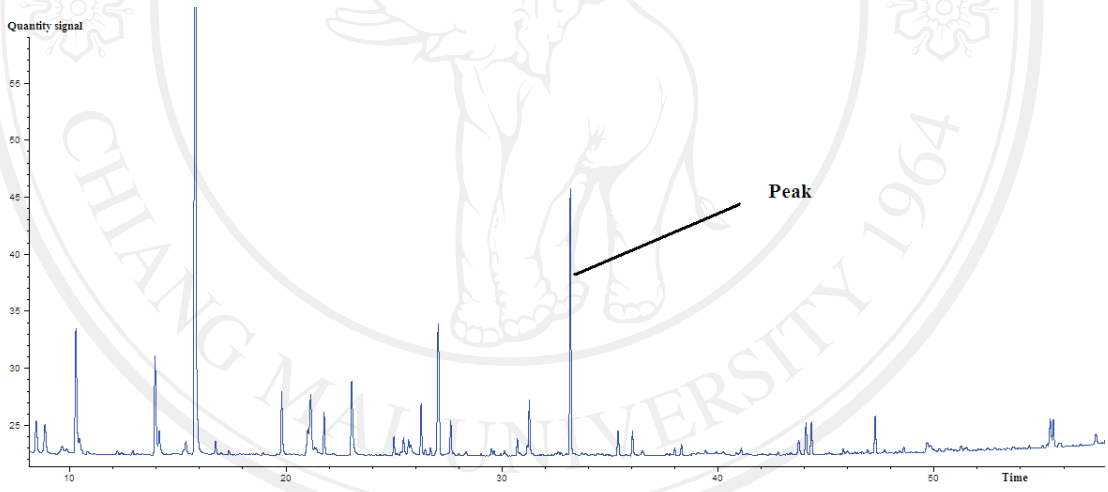

Figure 2.8 GC chromatogram

2.3.5 Chromatogram

The gas samples are flown from the beginning to the end of stationary phase with different durations based on the structures or chemical characteristics and there is the specific time for each peak-area variable. The specific time is called the retention time. Each retention time is characteristic of each peak-area variable (Figure 2.9).

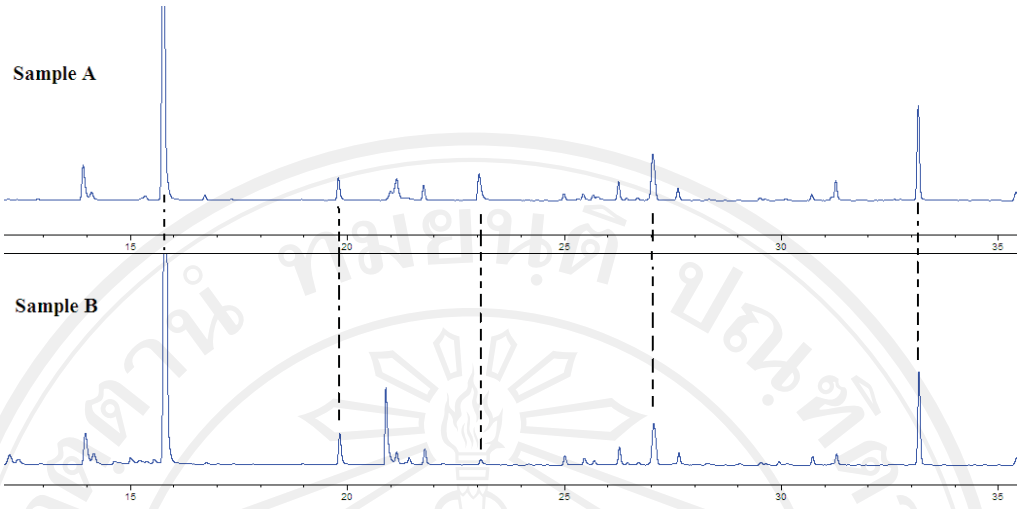

Figure 2.9 Retention time comparison between samples A and B

The result from the GC technique is called a "GC Chromatogram", which shows the relationship between time (X axis) and the quantity signal (Y axis). The quantity of a peak-area measured is related to the area under its chromatogram peak.

2.4 Statistical Methods

The objective of several chemical studies is to compare the chemical quantity in each component, and to classify or identify the group of subjects based on these chemical properties [10-12, 14, 21-24 and 26-29]. There are many statistics to compare the mean, median and percentiles which were suitable to the different situations. ANOVA is a standard statistical method to compare the means between two or more groups. The PCA is used to form the principal components (PCs) from all chemical components and brought to the main PCs to group the chemical properties.

One of several classification techniques, the discriminant analysis (DA) is the best method for the work proposed here, because the corrected classification and calibration that is resulted from DA produced models that are optimized over other

techniques [53]. The classification model is evaluated by cross validation technique [46].

2.4.1 Central Tendency Measurement

The central-tendency value is the data representative of each group or subject. For different solutions, different statistics are appropriate in each case including the mean which is suitable with normal distributed data, and the median which is better than mean when dealing with non-normal data or small-size samples [10, 11, 30, 42, 47].

2.4.2 Analysis of Variance (ANOVA)

ANOVA is used to determine whether samples from two or more groups come from populations with equal means. ANOVA test is fairly straightforward. As the name "analysis of variance" implies, two independent estimates of the variance for the dependent variable are compared, one that reflects the general variability of respondents within the groups (*MSW*) and another that represents the different between groups attributable to the treatment effects (*MSB*):

1. Within-groups estimate of variance (*MSW*: mean square within groups): This is an estimate of the average random respondent variability on the dependent variable within a treatment group and is based on deviations of individual scores from their respective group means. *MSW* is comparable to the standard error between two means calculated in the t test as it represents variability within groups. The value *MSW* is sometimes referred to as the error variance.

2. Between-groups estimate of variance (*MSB*: mean square between groups): The second estimate of variance is the variability of the treatment group means on the dependent variable. It is based on deviations of group mean from the overall mean of all scores. Under the null hypothesis of no treatment effects (i.e.,$\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$), this variance estimate, unlike *MSW*, reflects any treatment effects that exist; that is, different in treatment means increase the expected value of *MSB*.

Given that the null hypothesis of no group different is true, *MSW* and *MSB* represent independent estimates of population variance. Therefore, the ratio of *MSB* to *MSW* is a measure of how much variance is attributable to the different treatment versus the variance expected from random sampling. The ratio of *MSB* to *MSW* gives us a value for the *F* statistic, and can be shown as

$$F\ statistic = \frac{MSB}{MSW}$$

Since group different tend to inflate *MSB*, large value of the *F* statistic lead to rejection of the null hypothesis of no different in mean across groups. If the analysis has several different treatments (independent variables), then estimates of *MSB* are calculated for each treatment and *F* statistics are calculated for each treatment. This allows for separate assessment of each treatment.

Determine the critical value for the *F* statistic ($F_{critcal\ value}$) by referring to the *F* distribution with (*k-1*) and (*N-k*) degrees of freedom for a specified level of α (where $N = N_1 + \ldots + N_k$ and *k* = number of groups). If the value of the calculated F statistic exceeds $F_{crit,}$ conclude that the means across all groups are not equal [47].

2.4.3 Principal Component Analysis (PCA)

The PCA method is widely used in multivariate statistics to address several problems [10-12]. The most common PCA is used to extract factors and to form uncorrelated combinations of the observed variables. The first component accounts for the maximum variance. Successive components progressively explain smaller portions of the variance. The components are all uncorrelated with each other. The PCA method is used to provide the initial factor solution that can even be applied when a correlation matrix is singular.

The data was set as $X' = (X_1, X_2, \ldots, X_p)$ where X' is a random variable vector, Σ is a covariance matrix, and there are $(\lambda_1, e_1), (\lambda_2, e_2), \ldots (\lambda_p, e_p)$ as the eigenvalue and eigenvectors with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. The random variables $(X_1, X_2, \ldots, X_p)$ were transformed to the linear function of principal components;

$$Y_1 = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p$$

$$Y_2 = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \qquad 2.1$$

$$Y_p = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p$$

where

$$Var(Y_i) = e_i' \sum e_i = \lambda_i \; ; \; i = 1,2, \ldots, p \qquad 2.2$$

$$Cov(Y_i, Y_k) = e_i' \sum e_k = 0 \; ; \; i, k = 1,2, \ldots, p \; ; \; i \neq k \qquad 2.3$$

The eigenvectors of the covariance matrix are used to create the linear function in equation 2.1, which define the eigenvalue and eigenvector from values of covariance and correlation matrix.

2.4.3.1 Covariance Matrix and Correlation Matrix.

The covariance matrix (Σ) is defined with respect to the population matrix,

$$\Sigma = Cov(X) = E(X-\mu)(X-\mu)' = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \quad 2.4$$

where

$\sigma_{ii} = \sigma_i^2$ is the population variance value of variable *i* and σ_{ik} as population covariance value of variable *i* and *k* ($i \neq k$).

The sample covariance matrix (*S*) are,

$$S = \begin{bmatrix} \frac{1}{n}\sum_{j=1}^{n}(x_{1j}-\bar{x}_1)^2 & \cdots & \frac{1}{n}\sum_{j=1}^{n}(x_{1j}-\bar{x}_1)(x_{pj}-\bar{x}_p) \\ \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{j=1}^{n}(x_{1j}-\bar{x}_1)(x_{pj}-\bar{x}_p) & \cdots & \frac{1}{n}\sum_{j=1}^{n}(x_{pj}-\bar{x}_p)^2 \end{bmatrix}$$

$$= \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \quad 2.5$$

where x_{ij} = the i^{th} variable of j^{th} sample

$\bar{x}_i$ = the mean value of i^{th} variable.

The correlation matrix (ρ) are defined with respect to the population matrix,

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \quad 2.6$$

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}} \quad ; \quad i,k = 1,2,\dots,p \qquad 2.7$$

where ρ_{ik} = population coefficient correlation between variable i^{th} and k^{th}

The sample correlation matrix (R) are,

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix} \qquad 2.8$$

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)(x_{kj}-\bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{kj}-\bar{x}_k)^2}}; i,k = 1,2,\dots,p \qquad 2.9$$

where r_{ik} = sample coefficient correlation between variable i^{th} and k^{th}

x_{ij} = the i^{th} variable for j^{th} sample

2.4.3.2 Eigenvalue and Eigenvector

An eigenvalue is a value that characterizes the variances of a single mode of the full matrix solution. Each eigenvalue corresponds to a distince state where the eigenvalue (λ) can be calculated from equations 2.10 and 2.11.

$|S - \lambda I| = \underline{0}$ for covariance matrix 2.10

or $|R - \lambda I| = \underline{0}$ for correlation matrix 2.11

and the eigenvector (e) can be compute from equations 2.12 and 2.13.

$(S - \lambda I)e = \underline{0}$ for covariance matrix 2.12

or $(R - \lambda I)e = \underline{0}$ for correlation matrix 2.13

From $Y_1 = e_1'X, Y_2 = e_2'X, \dots, Y_p = e_p'X$ as value of p PCs

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i) \quad 2.14$$

Equation 2.14 shows the sum of all variables in the *X* data matrix and the sum of all PCs variances are equal to the value of the sum of the eigenvalues, which is proportion to the PCs variance as described in equation 2.15.

$$(Proportion\ of\ the\ k\ rank\ PC) = \frac{\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p} \ ; \ k = 1,2,\ldots,p \qquad 2.15$$

Equation 2.15, if the $k^{th}\ PC$ has a high variance proportion, the $k^{th}\ PC$ can be described by the high variation of the original variable, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ the first *k* PCs were high variance. This principal was used to describe the number of PCs to decrease the dimension of the original data, where the number of PCs is less than the number of variables. *Y* as matrix of independent variable can then be computed the eigenvector from equation 2.16.

$$Y = e'X \qquad 2.16$$

where e' = eigenvector from the high eigenvector

2.4.4 Linear Discriminant Analysis (LDA) and Stepwise Linear Discriminant Analysis (SLDA)

The LDA is the most common statistical method used for classification by determining the discriminate varieties or classes [10, 47-49]. It builds a predictive model for membership groups.

Discriminant Analysis Assumption

1. Independent variables $(X_1, X_2, \ldots, X_p)$ in each group or class are distributed as a Multivariate Normal Distribution.

$X_i \sim N(\mu_i, \Sigma_i);\ \ i = 1, 2, \ldots, k\ \ k \geq 2$

2. Covariance matrices in every group are equal.

$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$ for k groups.

where Σ_i = Covariance Matrices of group i^{th}

which $i = 1, 2, \ldots, k\ ; k \geq 2$

The model is composed of a discriminant function for two groups (or m-1 functions for m groups) based on linear combinations of the predictor variables that provide the best discrimination among the groups. The functions are generated from samples in each membership group and its functions can be applied to new cases that could predict other unknown-group classes. The LDA model constructs a set of linear functions over the predictors, known as discriminant functions such as;

$$Y_g = \beta_{0g} + \beta_{1g} X_{1g} + \beta_{2g} X_{2g} + \ldots + \beta_{pg} X_{pg} + \varepsilon \qquad 2.17$$

where

X_{ig} = value of the i^{th} variable and gth function

β_{ig} = discriminant weight of i^{th} variable in g^{th} function

Y_g = predicted groups or classes from g^{th} function

ε = error value

Equation 2.17, can rewrite to matrix format in 2.18.

$$Y_g = \beta_g^T X \qquad 2.18$$

Equation 2.19 is discriminant function from sample data

$$\hat{y}_g = b_{0g} + b_{1g} x_{1g} + b_{2g} x_{2g} + \ldots + b_{pg} x_{pg} \qquad 2.19$$

Equation 2.19, can rewrite to matrix format in 2.20.

$$\hat{Y}_g = b_g^T X \tag{2.20}$$

which $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$ 2.21

where $X =$ matrix data (p variables and n samples)

$\hat{Y}_g$ = vector of the prediction group

The β coefficients were estimated by b, which b gave maximum λ

$$\lambda = \frac{Between\ group\ sum\ of\ square}{Within\ group\ sum\ of\ square} = \frac{b^T B b}{b^T W b} \tag{2.22}$$

where

B = sample between groups matrix

W = sample within groups matrix

which can be computed B and W from equations 2.23 and 2.24.

$$B = \sum_{i=1}^{k} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \tag{2.23}$$

$$W = \sum_{i=1}^{k} \sum_{l=1}^{n_i} (\bar{x}_{il} - \bar{x}_i)(\bar{x}_{il} - \bar{x}_i)^T \tag{2.24}$$

where

$$\bar{x}_i = \frac{1}{n_i} \sum_{l=1}^{n_i} x_{il} \tag{2.25}$$

and $$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} \sum_{l=1}^{n_i} x_{il}}{\sum_{i=1}^{k} n_i} \tag{2.26}$$

where $\bar{x}$ = the mean value of k groups

$\bar{x}_i$ = the mean value of i^{th} group

From the principal condition, the best discriminant function should be the maximum λ value.

$$\lambda = \frac{b^T Bb}{b^T Wb} \qquad 2.27$$

where λ is the discriminant criterion that can be computed from equation 2.28.

$$|W^{-1}B - \lambda I| = 0 \qquad 2.28$$

where W^{-1} = invert matrix of W

I = identity matrix

λ = discriminant criterion

The estimation b should be the multiply b to equation 2.28.

$$(W^{-1}B - \lambda I)b = \underline{0} \qquad 2.29$$

Find the b value from equation 2.29. This b was the eigenvector of $W^{-1}B$ matrix which was according with λ at $b \neq 0$ (Nontrivial Solution).

So λ = eigenvalue of $W^{-1}B$ matrix

and b = maximum eigenvector

From the discriminant function in equation 2.19, the discriminant weight was brought to give in the discriminant function, and the total linear combination of all samples was given a vector discriminant score, where $\hat{Y}_g$ was the best discriminant function. In addition, the discriminant score was used to adjust or predict the group.

Wilks' Lambda

Wilks' lambda [47] is a statistic used in particular by Discriminant factor analysis as a measure of the class center separation. When the classes are multinomial with identical means and covariance matrix, the distribution of Wilks'

lambda is known, and therefore it can be used for testing the identity of the population means.

The LDA is a method that minimizes the variance within group and maximizes the variance between groups.

Stepwise linear discriminant analysis (SLDA) is the variable selection method in order to choose variables for the equation based on the lower optimized Wilks' lambda [10, 47-49].

Wilks' lambda [47] may also be used for variable selection in the discriminant analysis. It is possible to build a statistic that is approximately *F* distributed, and which is a function of the Wilks' lambdas pertaining to: a given subset of variables, and that same subset to which a new variable has been added.

An *F* statistic is then used for identifying which new variable will mostly increase the group separation. This variable is then added to the model.

Prediction Criterion

The discriminant function was first constructed from the training data, which were used to predict the unknown group. The criterion was then used with square Euclidean distances which brought the data from independent variables (x_0) to give in equation 2.30.

$$D_i^2 = \sum_{g=1}^{k-1}\left(\hat{y}_g - \bar{y}_{ig}\right)^2 = \sum_{g=1}^{k-1}\left[b_g'(x_0 - \bar{x}_i)\right]^2 \quad ; i = 1,2,3,..,k \qquad 2.30$$

when $\bar{y}_{ig} = b_g'\bar{x}_i$

where $\hat{y}_g$ = discriminant score from g^{th} function

b_g' = discriminant coefficient from g^{th} function

x_0 = new independent data

$\bar{x}_i$ = independent variable mean in i^{th} group

x_0 was in i^{th} group which $D_i^2 = min\{D_1^2,\ D_2^2, \ldots, D_k^2\}$ for separation the x_0 distance from each group with these group centroids, finding the minimum distance and give x_0 in that group.

2.4.5 Cross Validation (CV)

CV is a technique that potentially examined the created models by dividing data into two segments: one used to learn or train a model and the other used to validate the model [46, 50]. Leave-one out cross validation (LOOCV) technique is the one form of CV that is more suitable to use for small-size samples. The LOOCV technique can be automatically preformed through the computer software. Firstly, one of the training samples is taken off. Secondly, the calibration is set up from the remaining training-samples. Next, the established calibration is used to predict the samples that were picked out from the previous step. After that, the procedure will be repeated with the next training samples until the last samples. Finally, the obtained predication from the procedure will be compared with the expected value in order to measure predictive abilities of the calibration [50].

2.5 Related Work

A number of researches have reported studies on classifying characteristics from volatile compounds such as the flavor characterization of cheeses [11], Nebbiolo-based wines from Piedmont [10], the senescence of climacteric or non-climacteric melon fruit [12] and these results are related to fragrant rice volatile compounds in many aspects. However, several reports have focused on the statistical methods to classify characteristics or phenotype. These reports are listed as follows:

P. Dirinck and A. De Winne [11] studied the flavor characterization and classification of cheeses by gas chromatographic-mass spectrometric profiling. PCA was found to be able to classify Gouda and Emmental cheese. However in the Emmental cheese group (Austrian, French and Swiss Emmental cheese), only the Austrian Emmental was clearly differentiated from French and Swiss products.

E. Marengo, et al [10] studied the classification of Nebbiolo-based wines from Piedmont (Italy) by mean of a solid-phase microextraction-gas chromatography-mass spectrometry (SPME-GCMS) of volatile compounds. They used principal component analysis (PCA), hierarchical cluster analysis, Kohonen self organizing map, stepwise linear discriminant analysis (SLDA) and soft independent modeling of class analogy to reveal a good separation between five methods and found that SLDA is the best classification method and the interpretability of models improves the result of the statistical analysis. A main factor, connected to wine vintage, was identified and related to some analysts.

M. Kusano, et al [51] studied the application of a metabolomic method combining one-dimensional (1D) and two-dimensional (GCxGC) gas chromatography-time-of-flight/mass spectrometry to metabolic phenotyping of nature variants in rice. They developed a comprehensive method combining analytical techniques. This method was applied to metabolic phenotyping of natural variants in rice for the 68 world rice core collection (WRC) and two other varieties. Ten metabolites, were selected as metabolite representatives, and the selected ion current of each metabolite peak obtain from both techniques were statistically compared. The method of combining 1D- and GCxGC-TOF/MS is useful for the metabolic phenotyping of natural variants in rice for further studies in breeding programs.

N. Jaisieng, et al [52] studied rice varieties classification based on gas chromatographic profiles in rice grain using artificial neural network (ANN) and discriminant analysis that were calculated. Using ANN, the obtained results indicated good classification and prediction capabilities. Furthermore, a similar success rate could be achieved by using LDA. However, in LDA the assumption of a multivariate normal distribution in the data was not conserved and this method seemed too inappropriate to be used for rice variety classification.

F. Markowetz and R. Spang [53] studied and compared the classification model from microarray data which were expressed the various gene variables. They applied several methods to that research; LDA, QDA, T-statistics test and CV technique. The selected gene by T-statistics data with LDA model (DLDA) is the best classification model which has the lowest gap between the test error and the training error. The DLDA was of less complexity than the all gene data with LDA and QDA. The QDA model provided a better classification result from the training data than the LDA. However, the classification result in test set QDA provided less accuracy than in the LDA.