CHAPTER 1

INTRODUCTION

1.1 Chemometrics

- *What is chemometrics?* [1]

Chemometrics is a multivariate data analysis that uses mathematical and statistical knowledge in order to extract information from large and complicated data, and to interpret that information.

- *How does chemometrics work?*

In chemical analysis, when the data are complicate, it is difficult to focus one's attention on one or two variables at the same time. However, the complicate data usually contain the great amount of the information. There are a wide range of variables or factors affecting the whole system, and these variables also correlate and interact with each other. Instead of using only one or two points of data to predict the analytical results, chemometric approach would put focus on as much as possible of the data concerning the useful information to result in the most reasonable results.

1.2 Chemometrics and a mathematical matrix [2, 3]

A mathematical matrix is a rectangular array of numbers. The numbers in the array can be any fields of any data, such as absorbance data in the analytical chemistry. The matrix can vary in size that can be described by the specifying number

of rows (horizontal lines) and columns (vertical lines). It appears that the matrix can be used as a good interface between experimental data and the mathematical calculation because a great numbers of the experimental data would contain in only a matrix and those of the data can be manipulated easily through the matrix operations.

The mathematical matrix system used throughout these studies is row-wise system. In the system, each row of the matrix is designated for each sample and each column of the matrix is designated for each variable studied.

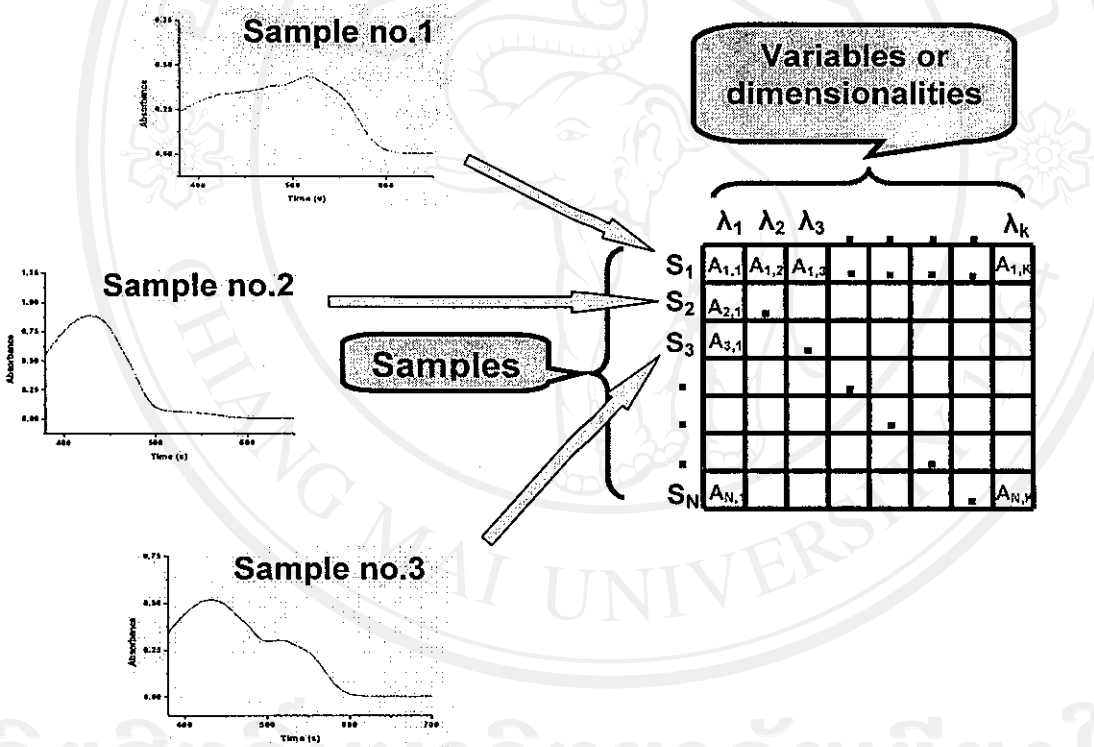

Figure 1.1 The schematic picture of how to manipulate some experimental data into a matrix using row-wise system.

An example of how to manipulate the experimental data in the row-wise system matrix is shown in Figure 1.1. In this example, supposing the data are

spectrum data: the studied variables are the wavelength and the collected data in the analytical process are the absorbance of each wavelength. Therefore, when the data no.1 are organized in the first row of the data matrix, each absorbance of each wavelength will be filled in each column of the first row of the matrix, similarly the data for samples no. 2 and no. 3 are organized in the second and the third row, respectively.

In this case, if the number of the samples were 50 samples (N = 50) and the number of the variables studied were 1000 variables or wavelengths (K = 1000), it could imply that in only one matrix, the data would contain as much as 50 samples and each sample has 1000 members of the variables. Moreover, there are 50 x 1000 or 50000 absorbance data containing in the matrix.

- *The data sets used in chemometrics* [2]

Three sets of data namely training set, validation set, and sample set involve in chemical analysis.

In the calibration step, the training set, a data set containing 2 sets of data often called independent and dependent variables which are a set of known samples and a set of the measurements on the known samples, respectively, is used. An attempt is made to find a relationship between the two groups of the variables in order to calibrate the model.

In the prediction step, the sample set which is the independent data that are obtained from the samples is used together with the model to predict the values or the predictive results for the dependent variables.

However, it is necessary that the model should be tested for the predictive abilities before it is used to predict the real samples in the prediction step. In this step called validation step, the validation set which is an additional data set containing independent measurements on the samples that are independent from the training samples is performed as it is a sample set for obtaining the predictive results. Then, those of the predictive results are compared with the expected results in order to calculate the predictive abilities of the model such as Predicted Residual Error Sum-of-Square or PRESS [1].

1.3 Principal Component Analysis, PCA [2, 3, 4]

Principal Component Analysis (PCA) is a multivariate statistical technique that reduces the dimensionality of a data set while the data were retained as much the variability of the original data as possible. By PCA approach, the variables of the data matrix will be summarized into a few latent variables called score vectors or Principal Components, PCs.

In theory, there are as many the numbers of the possible PCs as the numbers of the parent data variables. While PC number 1 or PC_1 always exhibits the greatest amount of the variation, PC_2 always exhibits the second greatest numbers of the variation, PC_3 always exhibits the third greatest numbers of the variation, and so on. When using PCA, it is hoped that the significant of the variation of most PCs in the rear of the data will be so low as to be negligible. In this case, the variation in the data set can be adequately described by means of a few first PCs in which most of the variations are contained.

- *PCA and matrices*

Figure 1.2 simplifies how PCA works though the data matrix. The data are supposed to be organized into the N x K data matrix and it can be seen that the data matrix X is composed of the three important data matrices that are $\overline{X}$ matrix, score matrix, and residual matrix.

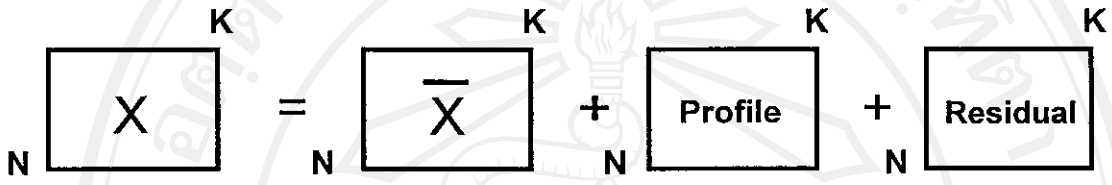

Figure 1.2 The matrix structure of a data matrix; $\overline{X}$ matrix, score matrix, and residual matrix.

➢ $\overline{X}$ **matrix** is the data matrix that keeps the mean value of each variable. The method used for calculating the mean value depends on the nature of the data. In this work, the mean values are obtained from the mean-centering approach.

➢ **Profile matrix** is the data matrix that contains the profile of the data. This profile will indicate the difference of each data from the mean value of each variable. Profile matrix is the most important part of the data because it is used to explain the variation of the data on the variables and when performing PCA, only this data part is usually in the interest.

➢ **Residual matrix** is the data matrix that contains the systematic error of the data. This feature provides that the theoretical model with the known parameters is not perfect. Occasionally, the residual matrix is decomposed into the PCs that contain noise.

In PCA, the profile matrix is decomposed into the score vectors $t = [t_i, i = 1, 2,\ldots, k]$ and loading vectors $p_i = [p_i', i = 1, 2,\ldots, n]$ as shown in the Figure 1.3. It is expected that as much as possible of the variation of the profile matrix can be explained by the combinations of only a few first components ($t_1p_1' + t_2p_2' + t_3p_3' + \ldots$). It should be as a little variation as possible remains in the rear combinations that can be classified as the residuals so that they can be negligible during the analysis procedure.

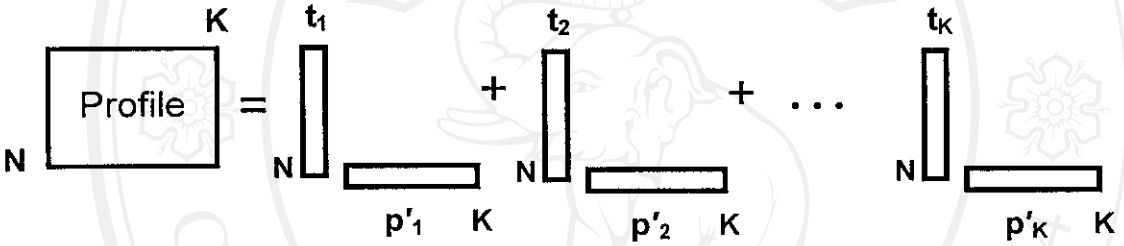

Figure 1.3 The compact matrix pictograph of the profile matrix and its vector combination.

From the visual appearance in Figure 1.3, the score and the loading vectors and the residuals of a profile matrix can then be written in vector algebra:

$$\mathbf{X}_{\text{profile}} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \mathbf{t}_3\mathbf{p}_3' + \ldots + \mathbf{E}_{\text{residual}} \quad (1.1)$$

and this may be written in more compact matrix form:

$$\mathbf{X}_{\text{profile}} = \mathbf{TP}' + \mathbf{E}_{\text{residual}} \quad (1.2)$$

Summarizing, the compact overview of PCA may be written:

$$\mathbf{X} = \overline{\mathbf{X}} + \mathbf{TP}' + \mathbf{E}_{\text{residual}} \quad (1.3)$$

1.4 Principal Component Regression, PCR [2, 3, 5, 6]

PCR is the technique that performs a least-squares regression of the latent variables that are the new coordinated variables or the components obtained from PCA approach. There are the predominant advantages of the use of PCA in data pretreatment such as the full spectrum analysis and the benefit of removing some noises.

- *Calibration steps of PCR*

In this step, the components are optimal for estimating the predictive results. The PCR model parameter that is regression coefficients or sensitivities of the model are calculated from the relationship between dependent and independent variables.

Following the Eq. (1.2), the compact matrix from of the score vectors or the components of the independent data of the training data set may be written:

$$\mathbf{T} = \mathbf{XP} \quad (1.4)$$

Where:

T is the matrix containing the score vectors or the components of the original independent variables of the training set.

X is the matrix that contains the original independent data of the training set.

P is the matrix containing the loading vectors of the original independent data of the training set.

To calculate the regression coefficient **F**, the regression coefficient F is supposed to be a linear combination of the matrix T:

$$\mathbf{C} = \mathbf{FT} \tag{1.5}$$

Where:

C is the concentration matrix (the approach for composition analysis).

F is the calibration coefficient matrix.

T is the matrix containing the score vectors or the components of the original independent variables of the training set.

To solve for the regression coefficient F, the T matrix must be eliminated from the right-side of the Eq. (1.5) by:

$$\mathbf{CT}' = \mathbf{FTT}'$$

$$\mathbf{CT}'[\mathbf{TT}']^{-1} = \mathbf{FTT}'[\mathbf{TT}']^{-1}$$

$$\mathbf{F} = \mathbf{CT}'[\mathbf{TT}']^{-1} \tag{1.6}$$

- *Prediction steps of PCR*

First, the data from the samples should be transformed into the new coordinate variables T_{unk} by the Eq. (1.2). Then these new coordinate variables are used together

with the regression coefficient F that was calculated from calibrating steps to predict the dependent data C_{unk}, by Eq. (1.7):

$$\mathbf{C}_{unk} = \mathbf{F}\mathbf{T}_{unk} \qquad (1.7)$$

1.5 Partial Least-Square Regression, PLS [2, 3, 6, 7]

PLS is an extension of the multiple linear regression models that can analyze data with strongly collinear, noise and numerous variables, and even incomplete variables in both dependent and independent data. In this study, the PLS model parameters (T, P, U, G, and B) was mainly performed by using the algorithm called Non-linear Iterative Partial Least Square (NIPALS) [6]. These model parameters can provide an intriguing connection between the data variables that can enhance the possibility in prediction of the results from the complicated data or the data that the relationship might or might not exist where ordinary regression was difficult or impossible to apply.

- *Calibration step of PLS*

In this step, PLS is very similar to that of PCR except the way that the factors are calculated. To describe the variation in the independent data, X, and dependent data, Y, in PLS, the columns of the X matrix are used for estimating the factors for the Y matrix. As the same time, the columns of the Y matrix are used for estimating the factors for the X matrix. From this change, the Eq. (1.4) in the calibration step of PCR can be rewritten:

$$\mathbf{T} = \mathbf{XW} \tag{1.8}$$

Where:

T is the matrix containing the factors corresponding to both of the independent and dependent data of the training set.

X is the matrix that contains the original independent data of the training set.

W is the matrix containing the loading vectors corresponding to both of the independent and dependent data of the training set.

and the resulting Eq. for the independent data:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \tag{1.9}$$

X is the matrix containing the original independent data of the training set.

T is the matrix containing the independent factors corresponding to both of the independent and dependent data of the training set.

P is the matrix containing the loading vectors of the original independent data of the training set.

E is the matrix containing the errors corresponding with modeling T with the PLS model.

and the resulting Eq. for the dependent data:

$$\mathbf{Y} = \mathbf{UG}' + \mathbf{F} \tag{1.10}$$

Y is the matrix that contains the original dependent data of the training set.

U is the matrix containing the dependent factors corresponding to both of the independent and dependent data of the training set.

G is the matrix containing the loading vectors of the original dependent data of the training set.

F is the matrix containing the errors corresponding with modeling U with the PLS model.

In the ideal situation, the source of the variation in X is exactly equal to the source of the variation in Y and the factor for T and C should be identical. However, in real application, X is usually varied in the ways that may not correlate to the variation in Y, therefore $T \neq U$.

So, in PLS, the additional factor is determined in order to adjust the variation for both of the factors. The relationship is given by:

$$\boldsymbol{U = bT} \tag{1.11}$$

Where:

U is the matrix containing the dependent factors corresponding to both of the independent and dependent data of the training set.

T is the matrix containing the independent factors corresponding to both of the independent and dependent data of the training set.

b is inner relationship between U and T.

In other words, from the Eq. (1.10) and Eq. (1.11), Y can be estimated from the factors of X by:

$$\mathbf{Y} = \mathbf{bTG}' + \mathbf{F} \qquad (1.12)$$

Where:

Y is the matrix that contains the original dependent data of the training set.

b is inner relationship between U and **T**.

T is the matrix containing the independent factors corresponding to both of the independent and dependent data of the training set.

G is the matrix containing the loading vectors of the original dependent data of the training set.

F is the matrix containing the errors corresponding with modeling U with the PLS model.

- *Prediction steps of PLS*

To perform prediction on an unknown samples, the factors of X and Y and the inner relationship are derived from the PLS model. The factors responding from the samples ($\mathbf{T}_{unk}$) also calculated. Then, these of the PLS model parameters are used to estimate the dependent data (Y_{unk}) by the Eq. (1.13):

$$\mathbf{Y}_{unk} = \mathbf{bT}_{unk}\mathbf{G}' + \mathbf{F} \qquad (1.13)$$

1.6 Cross-validation, CV [2, 8]

CV is a technique that estimates how well the model is going to perform on the future sample data by using some of the training set data. This validation technique is potentially used when there are insufficient samples to from a proper validation set especially in PCR or PLS calibration because these kinds of the techniques require at least three set of samples for establishing the model; one is used for making a calibration, other one is used for indicating how many factors should be retained, and the other one is used for determining the performance of the calibration.

In this study, leave-one-out cross validation (LOOCV) will be discussed. The procedure can be performed automatically though the computer programming. First, one of the training samples is taken out. Then, the calibration is established from the remaining training samples. Next, the established calibration is use to predict the samples that was left out in the previous step. After that, the procedure will be repeated with the next training sample until the last training samples. Finally, the obtained predicted results from the procedure will be compared with the expected values in order to measure the predictive abilities of the calibration.

1.7 Research Aims

The aims of these studies were to investigate the chemometric procedures such as Principal Component Analysis (PCA), Principal Component Regression (PCR) and Partial Least-Square Regression (PLS) by making use of the basic computer language and to apply them to classify the complicate data, to determine the analytes in the mixture samples and to interpret the results from the data that are from the un-optimized system.

The chemical analyses include:

1) a mixture of food colorants
2) flow injection analysis for thalassemia screening test

and 3) Bradford protein assay by sequential injection analysis.