CHAPTER 3

RESULTS AND DISCUSSIONS

3.1 Principal Component Analysis, PCA for classification/grouping of data sets

In this study, the experiments were concerning to the study of PCA in order to get the PCA procedure that can be applied with some of the analytical problems. After that, the PCA procedure was applied for classification of spectrum data of food colorants and for screening of thalassemia in the flow injection experimental work.

3.1.1 Classification of food colorants by PCA

In order to study the discriminate ability of the PCA procedure, the spectral data of the five food colorants from the sequential injection experimental work were performed by the PCA procedure. The results from each step of the procedure are shown in these parts. The new set of variation in the final step is revealed in the suitable two dimensional plots which are the plot of the first two components.

3.1.1.1 The spectra of the food colorant solutions

The classification of the five food colorants, yellow from Tartrazine, shrimp from Sunset yellow FCF, red from Carmoisine, green from the mixture between Tartrazine and Brilliant blue FCF, and orange from the mixture Tartrazine and

Carmoisine, by the PCA procedure were investigated. In this study, the yellow, shrimp, red, green food colorants were brought from the market near the university and the orange food colorant was obtained as the mixture solution of the yellow and red food colorants. Concentration of each food colorant solution was 10 ppm and prepared in sodium acetate buffer solution (pH 4.5). The visible spectra were recorded over the range of 375 to 725 nm.

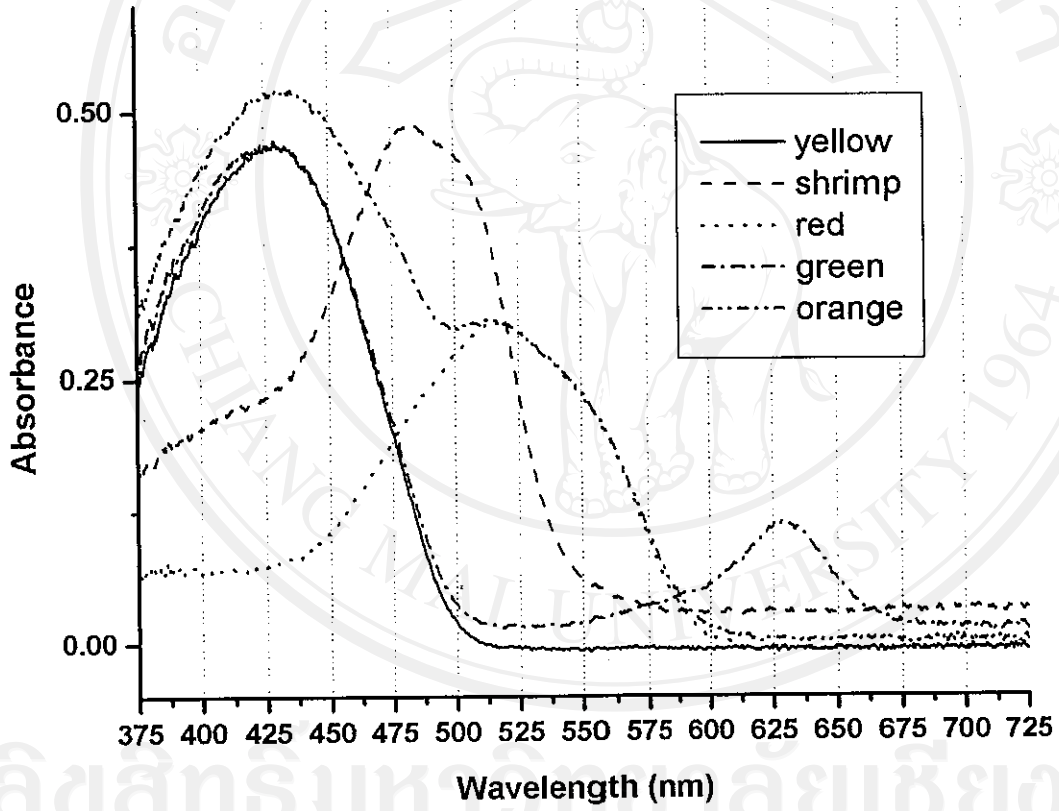


Figure 3.1 The spectra of the five food colorants: yellow; shrimp; red; green; orange.

From Figure 3.1, the spectra of the solutions of yellow, shrimp, and red food colorants show absorption maxima at 428, 482 and 519 nm respectively. It can be observed that the green food colorant solution which is a mixture of yellow and blue food colorants exhibits the double peaks which should belong to the yellow and blue respectively. Similarly, the orange solution exhibits double peaks which should be due to the yellow (428 nm) and the red (519 nm).

3.1.1.2 The food colorant spectral data in data matrix

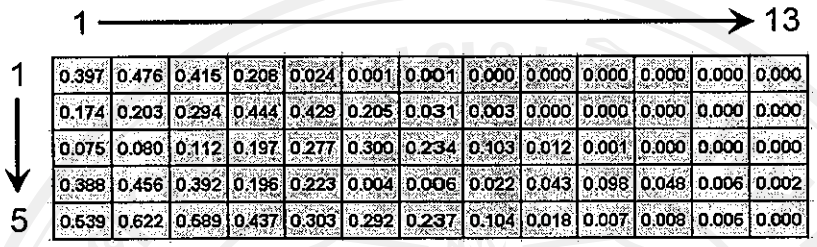
In order to perform the PCA procedure, the absorbance of the five food colorants in intervals of 25 nm from 400 to 700 nm obtained from Figure 3.1 are represented in Table 3.1. The numbers of the variables in this study were selected to be 13 variables.

Table 3.1 The absorbance of the five food colorants in intervals of 25 nm from 400 to 700 nm.

Food colorants	Absorbance at the wavelength (nm)												
	400	425	450	475	500	525	550	575	600	625	650	675	700
Yellow	0.397	0.476	0.415	0.208	0.024	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
Shrimp	0.174	0.203	0.294	0.444	0.429	0.205	0.031	0.003	0.000	0.000	0.000	0.000	0.000
Red	0.075	0.080	0.112	0.197	0.277	0.300	0.234	0.103	0.012	0.001	0.000	0.000	0.000
Green	0.388	0.456	0.392	0.196	0.223	0.004	0.006	0.022	0.043	0.098	0.048	0.006	0.002
Orange	0.539	0.622	0.589	0.437	0.303	0.292	0.237	0.104	0.018	0.007	0.008	0.006	0.000

For the suitable form of performing the mathematical operations, the absorption data in Table 3.1 were represented into a matrix system, row-wise system,

where: each row is donated for each food colorant and each column is donated for the absorbance of the wavelength at 25 nm interval as shown in Figure 3.2.



0.397	0.476	0.416	0.208	0.024	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
0.174	0.203	0.294	0.444	0.429	0.205	0.031	0.003	0.000	0.000	0.000	0.000	0.000
0.075	0.080	0.112	0.197	0.277	0.300	0.234	0.103	0.012	0.001	0.000	0.000	0.000
0.388	0.456	0.392	0.196	0.223	0.004	0.006	0.022	0.043	0.098	0.048	0.006	0.002
0.539	0.622	0.589	0.437	0.303	0.292	0.237	0.104	0.018	0.007	0.008	0.006	0.000

Figure 3.2 The data matrix of the absorbance of the five food colorants from 400 to 700 nm in 25 nm intervals.

In the Figure 3.2 shows the data matrix of the absorption from the five food colorants. It can be seen that the data matrix has the size of 5 rows and 13 columns which were obtained from the five food colorants (yellow, shrimp, red, green and orange). There are thirteen variables (13 wavelength of intervals of 25 nm from 400 to 700 nm) were selected.

3.1.1.3 The covariance matrix

To study the relationship between the variables, the covariance between each of the two variables was calculated and organized into a matrix called covariance matrix shown in Figure 3.3.

0.1397	0.1646	0.1270	0.0200	-0.0403	-0.0348	-0.0067	-0.0001	0.0043	0.0085	0.0053	0.0018	0.0001
0.1646	0.1941	0.1492	0.0216	-0.0506	-0.0451	-0.0114	-0.0017	0.0049	0.0102	0.0063	0.0021	0.0002
0.1270	0.1492	0.1224	0.0390	-0.0197	-0.0212	-0.0058	-0.0013	0.0025	0.0044	0.0033	0.0016	0.0001
0.0200	0.0216	0.0390	0.0693	0.0539	0.0410	0.0139	0.0026	-0.0030	-0.0090	-0.0037	0.0002	-0.0002
-0.0403	-0.0506	-0.0197	0.0539	0.0874	0.0590	0.0234	0.0080	0.0000	-0.0024	-0.0009	0.0001	-0.0001
-0.0348	-0.0451	-0.0212	0.0410	0.0590	0.0887	0.0641	0.0248	-0.0027	-0.0143	-0.0065	-0.0001	-0.0003
-0.0067	-0.0114	-0.0058	0.0139	0.0234	0.0641	0.0601	0.0254	-0.0001	-0.0083	-0.0035	0.0002	-0.0002
-0.0001	-0.0017	-0.0013	0.0026	0.0080	0.0248	0.0254	0.0112	0.0007	-0.0019	-0.0007	0.0002	0.0000
0.0043	0.0049	0.0025	-0.0030	0.0000	-0.0027	-0.0001	0.0007	0.0013	0.0028	0.0014	0.0002	0.0001
0.0085	0.0102	0.0044	-0.0090	-0.0024	-0.0143	-0.0083	-0.0019	0.0028	0.0074	0.0036	0.0004	0.0002
0.0053	0.0063	0.0033	-0.0037	-0.0009	-0.0065	-0.0035	-0.0007	0.0014	0.0036	0.0017	0.0002	0.0001
0.0018	0.0021	0.0016	0.0002	0.0001	-0.0001	0.0002	0.0002	0.0002	0.0004	0.0002	0.0000	0.0000
0.0001	0.0002	0.0001	-0.0002	-0.0001	-0.0003	-0.0002	0.0000	0.0001	0.0002	0.0001	0.0000	0.0000

Figure 3.3 The 13 X 13 dimensional covariance matrix of the five food colorant spectral data matrix.

The another point is that the covariance matrix is a square matrix which the number of the matrix dimension depends on the number of the studied variables and the fact that *cov(X,Y)* was equal to *cov(Y,X)*; where *X* and *Y* are variables, make the matrix symmetrical about the main diagonal.

3.1.1.4 The eigenvalue matrix and the eigenvector matrix

In this step, the eigenvectors which are the vectors that transform a square matrix or a transformation matrix into an multiple integer of the original vector or eigenvalues were calculated from the previous covariance matrix in order to compress the size of the covariance matrix were shown in Figure 3.4. Also, the eigenvalue which were the essential by product of the eigenvector calculation were shown in Figure 3.5, respectively.

0.3053	0.4501	0.1946	0.0828	0.0403	0.0216	0.3311	-0.0785	0.4726	-0.1270	-0.0900	-0.0899	0.5347
-0.3267	0.2416	-0.0653	-0.0298	0.0561	0.0916	-0.3346	0.0806	-0.5239	-0.1066	-0.0869	-0.0756	0.6330
0.0618	-0.7561	-0.2439	-0.0035	-0.0534	0.0771	0.2218	-0.0467	0.0791	0.0961	0.1118	-0.2059	0.4843
0.0087	0.0817	0.2883	-0.0706	-0.1041	-0.3747	-0.3259	0.0612	0.1190	0.4364	0.4502	-0.4829	0.0592
-0.0441	0.1374	-0.2103	-0.1030	0.1262	0.0451	0.2771	-0.0481	-0.1573	-0.5652	0.4671	-0.4802	-0.1854
0.0162	0.0122	0.0808	0.3387	-0.0761	0.6240	-0.1313	0.0194	0.0370	0.1392	-0.3162	-0.5579	-0.1834
-0.0414	-0.0789	-0.1376	0.0492	0.2715	-0.6088	0.0252	-0.0018	0.0010	-0.1068	-0.6060	-0.3768	-0.0709
0.0320	-0.0776	0.2822	-0.7649	-0.4360	0.0909	0.0193	-0.0024	-0.0134	-0.1685	-0.2827	-0.1438	-0.0203
0.0336	0.0565	-0.2112	0.4061	-0.8152	-0.2484	-0.0599	0.0135	-0.0154	-0.2353	-0.0128	0.0089	0.0148
0.0198	-0.3553	0.6554	0.2730	0.1273	-0.0482	-0.2541	0.0562	0.0313	-0.5219	0.0558	0.0726	0.0346
-0.0573	-0.0084	-0.4153	-0.1794	0.0924	0.1024	-0.5906	0.1173	0.5859	-0.2554	0.0242	0.0287	0.0209
0.8868	-0.0032	-0.1389	-0.0530	0.0816	0.0008	-0.2628	0.0637	-0.3326	-0.0320	-0.0016	-0.0040	0.0063
0.0012	-0.0020	-0.0039	-0.0035	0.0027	-0.0041	-0.2047	-0.9786	-0.0187	-0.0106	0.0013	0.0017	0.0006

Figure 3.4 The eigenvector matrix that was calculated from the covariance matrix of the five food colorant spectral data matrix.

-6.92E-17	0	0	0	0	0	0	0	0	0	0	0	0
0	-2.45E-17	0	0	0	0	0	0	0	0	0	0	0
0	0	-1.13E-17	0	0	0	0	0	0	0	0	0	0
0	0	0	-2.72E-18	0	0	0	0	0	0	0	0	0
0	0	0	0	-6.04E-19	0	0	0	0	0	0	0	0
0	0	0	0	0	4.02E-18	0	0	0	0	0	0	0
0	0	0	0	0	0	3.87E-17	0	0	0	0	0	0
0	0	0	0	0	0	0	4.18E-17	0	0	0	0	0
0	0	0	0	0	0	0	0	7.40E-17	0	0	0	0
0	0	0	0	0	0	0	0	0	0.02018	0	0	0
0	0	0	0	0	0	0	0	0	0	0.07632	0	0
0	0	0	0	0	0	0	0	0	0	0	0.20730	0
0	0	0	0	0	0	0	0	0	0	0	0	0.47943

Figure 3.5 The eigenvalue matrix that was calculated from the covariance matrix of the five food colorant spectral data matrix.

PCA procedure applies this transformation for transforming the data matrix into a new set of orthogonal variables named as eigenvector. Figure 3.5 is the eigenvalue matrix that was calculated from the covariance matrix of the five food colorant spectral data matrix. It can be investigated that the eigenvalue matrix are the diagonal matrix and the members in the matrix are formed in the diagonal position of the matrix. The eigenvalue, $X_{i,j}$; i = j, is the eigenvalue of the eigenvector column vector number i of the eigenvector matrix. The usefulness of the eigenvalue is that this value can be used for indicating the significant of its pair eigenvector. It can be

founded that the eigenvector that has the greatest eigenvalue contains can contribute the greatest variation of the data and so on.

3.1.1.5 Transformation of the new coordinated data

Following the PCA procedure of transforming the data described by Eq. (1.2) in the introduction section, the original data matrix will be transformed into the new coordination data that contains the variation data of the food colorant data, called components as shown in Figure 3.6.

Principal Components, PCs

7.75E-17	1.87E-17	-4.73E-17	-7.10E-17	3.03E-17	-1.00E-16	-1.60E-16	6.41E-17	2.54E-16	8.69E-02	-3.33E-02	2.57E-01	2.12E-01
1.47E-17	-7.32E-17	-4.73E-17	-2.59E-17	-2.53E-17	-3.33E-17	8.69E-18	-3.95E-17	9.96E-17	3.15E-02	2.09E-01	-1.12E-01	-2.39E-01
-9.20E-18	-6.02E-17	-1.61E-17	4.04E-18	2.16E-17	-2.85E-17	5.03E-17	-8.48E-17	3.28E-17	-1.09E-02	-1.56E-01	-7.60E-03	-4.78E-01
7.84E-17	5.80E-17	6.25E-17	7.06E-17	-4.13E-17	3.07E-17	2.59E-16	-4.77E-17	-2.90E-16	-1.07E-01	5.01E-02	1.76E-01	1.50E-01
-6.21E-17	1.04E-16	6.67E-17	7.48E-18	9.66E-18	4.00E-17	-1.29E-18	7.42E-17	8.06E-17	-1.18E-04	-7.00E-02	-3.13E-01	3.55E-01
PC_{13}	←								...	PC_3	PC_2	PC_1

Figure 3.6 All possible components generated from of the five food colorant spectral data matrix using all possible eigenvector.

In Figure 3.6, the first transformation data set, a column vector, named as component number one that exhibits the greatest amount of the variation is the new coordination data transformed from the most significant eigenvector. The second transformation data set named as component number two that exhibits the second greatest amount of the variation is the new coordination data transformed from the second most significant eigenvector, and so on to the last component that can be

possible generated, component number thirteen that contains the least variation of the data.

In this experiment, since the significant eigenvalue in Figure 3.5 are the first four eigenvalues which are 0.47943, 0.20730, 0.07632 and 0.02018 comparing to the eigenvalue number five which is much less significant (7.40E-17), the significant eigenvectors that can be the constant column vectors that can be used for transforming almost all variation of the data from the original data should be the four first column vectors of the eigenvector matrix.

However, though there are four significant eigenvalues, only the two first column vectors that have significant eigenvalues (0.47943 and 0.20730, respectively) are much enough to be used for expressing the variation of the data and it is more convenient to interpret to data in two dimensions than in the higher dimensions. In the case of keeping both of the eigenvectors with the largest eigenvalues for the data transformation, the transformation data that are transformed from the selected eigenvector are usually called principal components, PCs and the plot of the first two transformation data of the five food colorant spectral data are shown in Figure 3.7.

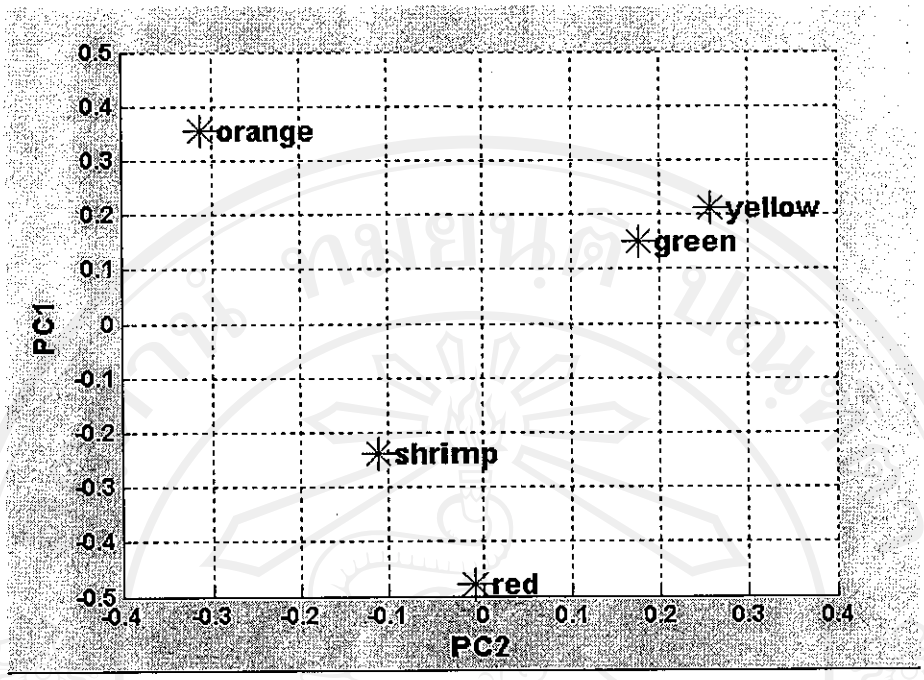


Figure 3.7 The PCA score plot of the first two PCs of the five food colorant spectral data.

Inspecting of the first two PCA score plot in Figure 3.7, the result shows that by the PCA procedure, each of the food colorant spectral data that have thirteen dimensions of the selected absorption variables can be reduced into the two new latent variables in the forms of principal component number one, (PC1) and component number two, (PC2). This plot of the two new variables shows the location of each food colorant and it indicates that PCA groups the data that have the similar patterns. The spectra between the yellow and green food colorants which show the overlap region from 375 to 500 nm in Figure 3.1 result in the closer distance in the PCA score plot. The spectrum of the shrimp food colorant that has the maximum absorption

closer to that of the red food colorant also shows the location in the PCA score plot closer to the red food colorant.

For the purpose of the studies, the data have been transformed into the principal component. The pattern of the data is dependent on the contribution of all of the data themselves. The usefulness is that the data can be classified out of the combination of the variation that is the contribution from all of the data themselves by using each of those of the PCs. Comparing to the spectral data in the Figure 3.1, the hold absorbance data of each spectrum data that have to use in order to explain the difference among the food colorants can be reduced into only a few new variables that express the difference of each data in the term of variation of each food data on that new variables

3.1.2 PCA for screening test for thalassemia by flow injection

The PCA procedure applied for screening test for thalassemia using flow injection [9] was studied. The FI-grams involved the osmotic fragility of the blood sample in the hypotonic buffer solution were reported in Figure 3.8. There are 30 blood samples including 15 samples from negative thalassemia blood samples and 15 samples from positive thalassemia blood samples studied.

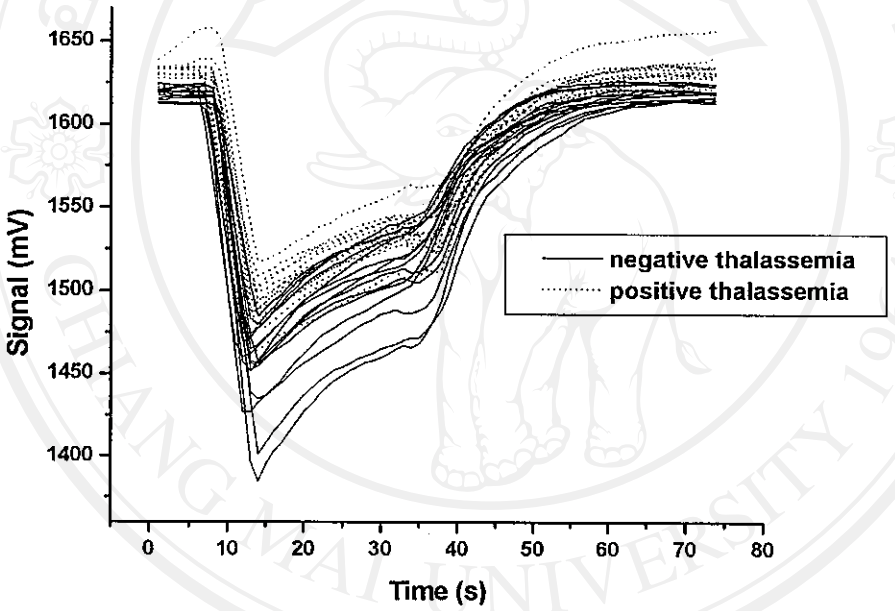


Figure 3.8 FIA-grams; —— negative thalassemia blood sample; ---- positive thalassemia blood sample.

In the data pretreatment step, the data of each FI-gram from 16 to 31 seconds were used for calculating the slope at every 6 s interval for obtaining the new type of data that depended on the change of the osmotic fragility of blood sample between the times shown in Table 3.2.

Table 3.2 The data that depend on the rate change of the osmotic fragility of blood samples of the time (s).

Sample No.	The data that depend on the rate change of slope of the time (s)									
	16	17	18	19	20	21	22	23	24	25
1	5.12	4.60	4.32	3.84	3.72	3.28	3.08	3.00	2.72	2.64
2	4.48	4.12	4.00	3.72	3.64	3.40	3.52	3.44	3.36	3.28
3	4.84	4.64	4.36	4.08	3.72	3.32	3.16	2.96	2.92	2.72
4	4.08	3.68	3.20	2.80	2.68	2.56	2.28	1.92	1.84	1.68
5	3.92	3.72	3.24	3.08	2.84	2.56	2.36	2.04	1.84	1.44
6	5.04	4.88	4.52	3.80	3.28	2.68	2.36	2.36	2.16	2.00
7	3.96	3.76	3.56	3.36	3.12	3.00	3.12	3.16	3.08	2.88
8	5.80	5.36	5.04	4.48	4.04	3.40	2.92	2.72	2.40	2.08
9	6.32	5.84	5.88	5.56	4.96	4.44	4.04	3.40	2.92	2.48
10	5.92	5.44	4.68	3.96	3.32	2.72	2.40	2.28	2.24	2.08
11	3.48	3.56	3.52	3.48	3.12	2.80	2.48	2.28	2.12	2.04
12	4.92	5.12	5.12	4.52	4.00	3.88	3.40	2.80	2.64	2.64
13	4.24	4.04	3.88	3.52	3.20	2.80	2.72	2.52	2.52	2.28
14	4.64	4.60	3.88	3.64	3.20	3.00	2.56	2.64	2.20	1.76
15	4.12	4.08	3.84	3.40	3.04	2.72	2.28	2.24	1.96	1.68
16	2.40	2.16	1.92	2.24	2.40	2.44	2.64	3.00	2.88	2.60
17	2.84	2.64	2.36	2.48	2.48	2.36	2.28	2.40	2.32	2.20
18	2.48	2.36	2.04	2.00	1.72	1.72	1.72	1.64	1.88	1.80
19	3.68	3.24	3.16	2.84	2.52	2.32	2.16	2.08	1.84	1.96
20	3.76	3.56	3.40	3.16	3.00	2.96	3.04	2.88	2.84	2.64
21	3.12	2.72	2.84	2.88	2.52	2.40	2.40	2.04	1.96	2.20
22	3.40	3.20	2.96	3.08	2.92	2.48	2.44	2.44	2.08	1.88
23	4.04	3.76	3.80	3.32	2.80	2.56	2.24	1.88	1.96	1.84
24	2.64	2.80	3.04	2.72	2.52	2.84	2.52	2.12	2.24	2.08
25	4.08	3.40	3.48	3.32	2.80	2.60	2.80	2.64	2.48	2.44
26	3.28	3.32	3.20	3.12	2.92	2.88	2.48	2.44	2.08	2.16
27	2.84	2.48	2.40	2.40	2.12	1.96	2.20	2.24	2.00	2.20
28	3.16	3.44	3.64	3.60	3.44	3.36	3.12	2.84	2.64	2.48
29	2.48	2.44	2.40	2.28	2.32	2.16	2.04	1.96	2.08	1.92
30	2.32	2.12	2.08	2.16	2.52	2.36	2.04	2.00	2.16	1.88

To reveal the patterns of the data from Table 3.2, the ten new variables that were calculated from the slope from 16 to 31 seconds at every 6 s interval of each blood sample were summarized into the two latent variables or the two PCs using the PCA procedure and shown in Table 3.3. The score plot of the first two PCs resulting from the PCA procedure was shown in Figure 3.9.

Table 3.3 The scores of the first two PCs of both negative and positive thalassemia blood samples.

Sample No.	Negative thalassemia		Sample No.	Positive thalassemia	
	PC1	PC2		PC1	PC2
1	1.6327	0.3043	16	-1.8201	1.4081
2	1.6372	1.3236	17	-1.6984	0.4111
3	1.8179	0.3999	18	-3.0092	-0.3858
4	-0.8558	-0.6828	19	-1.1370	-0.4688
5	-0.6243	-0.7165	20	0.1891	0.7323
6	0.9832	-0.8768	21	-1.3463	-0.0611
7	0.5881	0.9645	22	-0.8054	-0.0463
8	2.4506	-0.5619	23	-0.3142	-0.7992
9	4.6490	0.1714	24	-1.0868	0.0871
10	1.3829	-1.0591	25	-0.1019	0.2628
11	-0.1063	-0.2181	26	-0.4690	0.0281
12	2.7732	-0.0283	27	-2.0934	0.1056
13	0.4866	-0.0173	28	0.6503	0.6642
14	0.7034	-0.4940	29	-2.1744	-0.0073
15	0.0469	-0.7260	30	-2.3485	0.2862

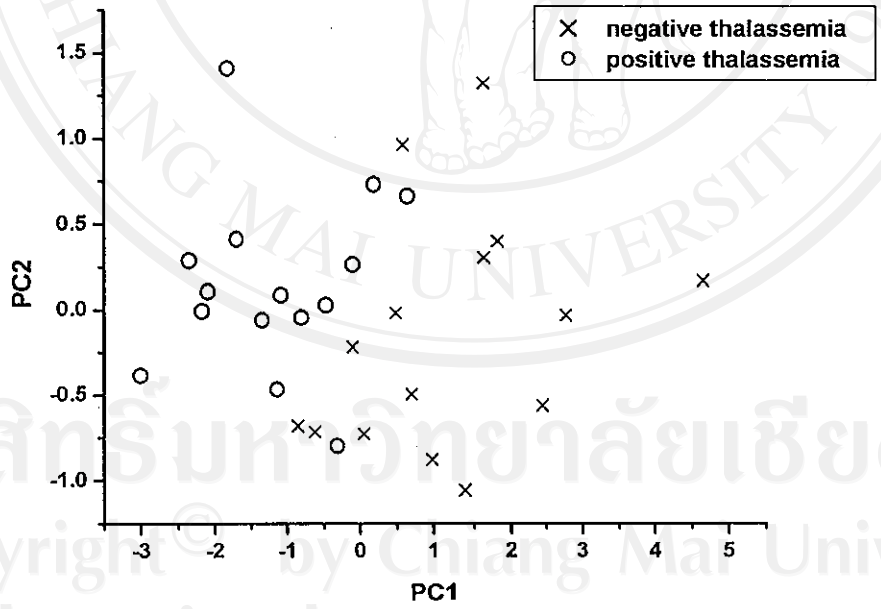


Figure 3.9 The score plot of the first two PCs of the thalassemia blood samples.

Based on the result of PCA score plot shown in Figure 3.9, the rate change of the osmotic fragility of the blood sample in the hypotonic buffer solution data from 16 to 31 s are summarized into the two PCs and it can also be expressed in the two dimensional plot. A pattern of the data could be noticed. Although the data cannot clearly be grouped in to the two groups of the negative and positive thalassemia blood samples. In Figure 3.9, the left side area, belonging to the circle marks, is the zone of positive thalassemia blood samples and the right side area, belonging to the cross marks, is the zone of negative thalassemia blood samples. From, the PCA score plot indicates that the blood sample number 9 which has 4.6490 and 0.1714 scores on PC1 and PC2, respectively, could be classified as a negative thalassemia blood sample. On the other hand, the blood sample number 16 which has -1.8201 and 1.4081 scores on PC1 and PC2, respectively, would be classified as a positive thalassemia blood sample. However, due to the overlap area between the circle and cross marks, the blood samples which have the locations in this area could not clearly be classified.

By using the PCA procedure, it shows the advantage that complicate data can be used in the screening process. Besides, the graphical result can be shown in the two dimensional space so that it is easier to classify the characteristic of the samples.

However, in this study, the conditions of the data pretreatment are not optimized. The parameters such as the area of the data and the time range should be studied for the better screening.

3.2 Principal Component Regressions, PCR

3.2.1 Determination of food colorant by PCR

To study the predictive ability of the PCR procedure for determining of the analyte in the spectrum containing overlapping peaks of interference, The PCR procedure was applied to determine the concentration of a yellow food colorant in the mixture solution of yellow and red food colorants. In this study, the PCR procedure was performed follow the process in the section 2.4.2 and the food colorant spectrum data were collected from the sequential injection experimental work.

3.2.1.1 The training set of the food colorant

In order to obtain the data for constructing the food colorant PCR models, the training set was prepared. The solutions of the mixture between the yellow and red food colorants were prepared in sodium acetate buffer (pH 4.5). The concentrations of the food colorants in the mixture solutions for the training set are shown in Table 3.4. The absorbance measured over the range of 400 to 650 nm is shown in Figure 3.10.

Table 3.4 The concentrations of the yellow and the red food colorants in the mixture solutions of the training set.

Sample names	Concentrations (ppm)	
	Yellow food colorant	Red food colorant
a	2.00	8.00
b	4.00	4.00
c	6.00	6.00
d	8.00	2.00
e	10.00	10.00
f	12.00	16.00
g	14.00	14.00
h	16.00	12.00

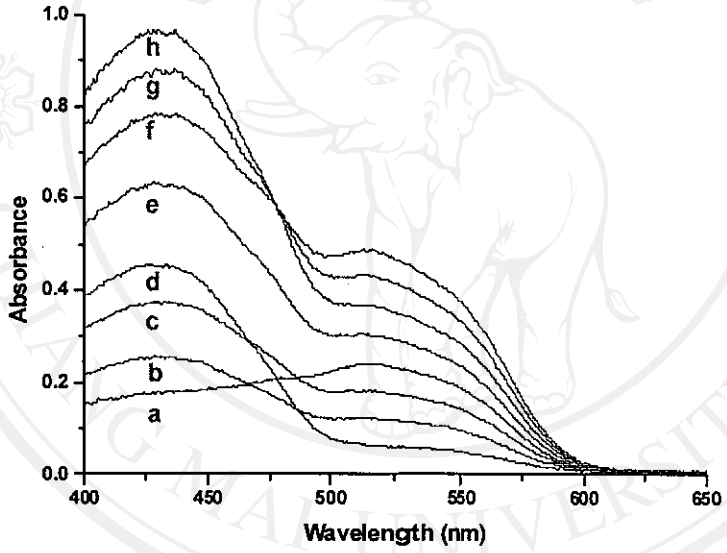


Figure 3.10 The spectra obtained from the mixture solutions of the training set measured over the range of 400 to 650 nm; (a) 2.00 ppm yellow and 8.00 ppm red; (b) 4.00 ppm yellow and 4.00 ppm red; (c) 6.00 ppm yellow and 6.00 ppm red; (d) 8.00 ppm yellow and 2.00 ppm red; (e) 10.00 ppm yellow and 10.00 ppm red; (f) 12.00 ppm yellow and 16.00 ppm red; (g) 14.00 ppm yellow and 14.00 ppm red; (h) 16.00 ppm yellow and 12.00 ppm red food colorants.

The results in Figure 3.10 show that the spectra of the mixtures contain two overlapping peaks. From the previous experiment in the section 3.1.1, the results can be concluded that the higher peaks that have the maximum absorption around 425 nm should be the peaks of the yellow food colorants and the smaller peaks that have the maximum absorption around 520 nm should be the peaks of the red food colorants.

To develop the calibrations, the training set, the data set containing measurements of a set of the known samples, was prepared. This set consists of a matrix of features for the independent variables, an absorbance matrix, and a matrix of features for the dependent variables, a concentration matrix. The relationship between both data sets is calculated in order to produce the model parameters. The most important criteria for designing the training set sample is that the concentration used in the training set must cover all of the possible concentration of the real samples.

However, PCR is the technique that performs a least-square regression on the principal components which are the compressed forms of the data by the PCA process so it is essential that the absorbance data shown in Figure 3.10 were assembled into a matrix and then transformed into new orthogonal variables, PCs, shown in Table 3.5.

Table 3.5 The scores of the first ten components transformed from the absorbance data of the training set data.

Sample names	Principal Component No.									
	1	2	3	4	5	6	7	8	9	10
a	4.0474	-2.0059	0.0112	-0.0014	0.0216	-0.0069	0.0051	-0.0036	-0.0036	-0.0013
b	4.1660	-0.1385	0.0034	-0.0039	0.0065	-0.0222	-0.0247	-0.0036	-0.0036	-0.0013
c	6.1430	-0.2314	0.0087	-0.0063	-0.0020	0.0211	-0.0155	-0.0036	-0.0036	-0.0013
d	6.3208	1.6577	-0.0029	-0.0146	-0.0104	-0.0098	0.0052	-0.0036	-0.0036	-0.0013
e	10.3570	-0.3383	-0.0007	-0.0564	0.0223	-0.0044	-0.0098	-0.0036	-0.0036	-0.0013
f	13.6090	-1.7569	0.0189	-0.0182	-0.0249	-0.0109	-0.0077	-0.0036	-0.0036	-0.0013
g	14.4470	-0.5701	-0.0420	0.0053	0.0115	-0.0046	-0.0082	-0.0036	-0.0036	-0.0013
h	15.0920	0.6729	0.0397	0.0039	0.0229	-0.0063	-0.0080	-0.0036	-0.0036	-0.0013

Table 3.5 shows the first ten PCs from the spectrum data in Figure 3.10. It can be seen that the few first components show the more significant score than in the latter components. These results show the advantage of using the PCA procedure that by this technique, it is possible that the spectrum data of each food colorant mixture can be express almost all of the variation data by using only the first few components.

3.2.1.2 The validation set of the food colorant

To evaluate the calibration's performance, the validation set which is the set that contains the known samples of the mixtures between the yellow and the red food colorants in sodium acetate buffer (pH 4.5) were prepared. The concentrations of the mixtures are shown in Table 3.6. The spectra measured over the range of 400 to 650 nm are shown in Figure 3.11.

Table 3.6 The concentrations of the yellow and the red food colorants in the mixture solutions of the validation set.

Sample names	Concentrations (ppm)	
	Yellow food colorant	Red food colorant
a	2.00	16.00
b	4.00	14.00
c	6.00	12.00
d	8.00	10.00
e	10.00	8.00
f	12.00	6.00
g	14.00	4.00
h	16.00	2.00

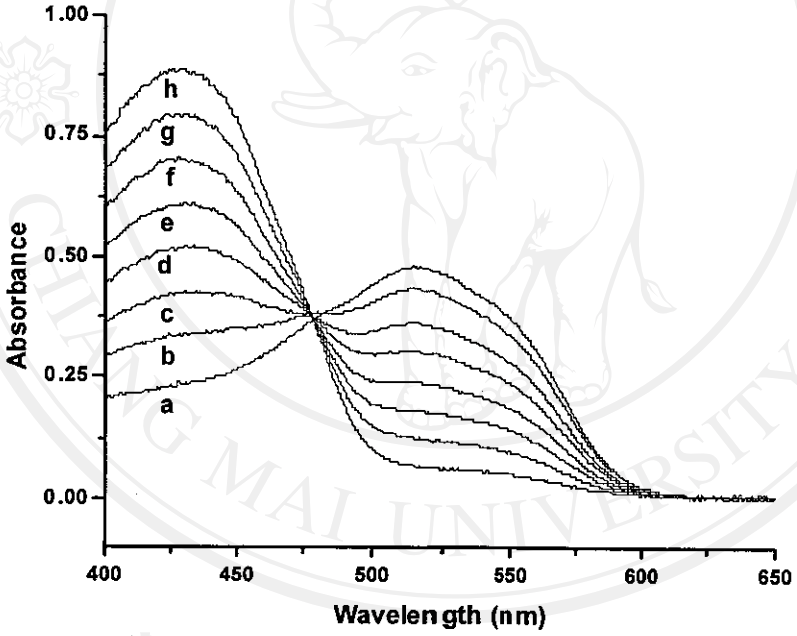


Figure 3.11 The spectra obtained from the mixture solution of the training set measured form 400 to 650 nm; (a) 2.00 ppm yellow and 16.00 ppm red; (b) 4.00 ppm yellow and 14.00 ppm red; (c) 6.00 ppm yellow and 12.00 ppm red; (d) 8.00 ppm yellow and 10.00 ppm red; (e) 10.00 ppm yellow and 8.00 ppm red; (f) 12.00 ppm yellow and 6.00 ppm red; (g) 14.00 ppm yellow and 4.00 ppm red; (h) 16.00 ppm yellow and 2.00 ppm red food colorants.

The validation set is the data set that consists of the known concentration data and the absorbance data of the known concentration samples shown in Table 3.6 and Figure 3.11, respectively. The absorbance data of the validation data will be applied with the PCR models as the absorbance data of the sample set for obtaining the predicted concentrations. Then, those of the obtained results are compared to the known concentrations, expected concentration, to determine the performance of the PCR models. Importantly, it is essential that the validation samples should have as much as it is possible the same characteristics as the real samples, such as the range of the concentrations in order that it can be used for predicting the model performance precisely. In this study, the parameter used for measuring the prediction abilities of the models is the predicted residual error sum-of-square or PRESS that is the summation of the square of the difference between the predicted and the expected concentrations. The fewer value of PRESS the model obtains, the better performance of the model is.

3.2.1.3 Generating the calibration

Form the Eq. (1.6) in the introduction section, the regression coefficients, the constant matrices of the models, which were used for predicting the concentration of an unknown samples, were calculated from the training set data. The regression coefficients resulting from the different number of the principal components used are shown in Table 3.7.

Table 3.7 The regression coefficients of the PCR models obtained from the principal components used from the first one to ten components.

Regression coefficient for the number of PCs used									
1	2	3	4	5	6	7	8	9	10
0.9791	1.0162	1.0949	2.3132	-1.1530	2.2110	0.2136	-11.723	10350	-3895.4
	1.0095	1.0164	1.1403	2.2757	-1.1087	2.1998	-0.6249	2767.9	-8597.2
		1.0090	1.0152	1.1513	2.2107	-1.1027	1.8193	-0.6249	3135.8
			1.0113	1.0148	1.1704	2.2019	-0.9001	1.8193	-0.6249
				1.0119	1.0143	1.1730	1.9047	-0.9001	1.8193
					1.0128	1.0142	1.2604	1.9047	-0.9001
						1.0129	1.0118	1.2604	1.9047
							1.0173	1.0118	1.2604
								1.0173	1.0118
									1.0173
[1 x 1]	[2 x 1]	[3 x 1]	[4 x 1]	[5 x 1]	[6 x 1]	[7 x 1]	[8 x 1]	[9 x 1]	[10 x 1]
Size of the matrix									

The regression coefficient matrices which are the matrices that contain the relationship data between the absorbance data and the concentration data of the training set from the ten first PCR models are shown in Table 3.7. The difference among the models is that the number of the components used in the calibration step. It is observed that the samples in this section are the single samples, yellow food colorant solution, so that the column of the resulting coefficient matrix has one dimension, and the number of dimension each row is equal to the number of the components used.

3.2.1.4 Prediction of the concentrations of the yellow food colorant of the validation samples by PCR

In the prediction step, the compressed data set of the validation samples was applied with the regression coefficients by using Eq. (1.7). There are ten PCR models studied. The difference among each model is that the number of components used in the generating the calibration steps. The results of the predictions of the PCR models are shown in Table 3.8.

Table 3.8 The predicted results of the yellow food colorants in the mixture solutions of the yellow and the red food colorants using the PCR models.

The expected concentrations	The predicted concentrations of the yellow food colorants for the number of PCs used									
	1	2	3	4	5	6	7	8	9	10
2.00	6.39	1.91	1.93	1.94	1.92	1.96	1.96	1.98	9.38	-4.02
4.00	7.39	4.06	4.07	4.08	4.06	4.07	4.07	4.08	-7.11	29.29
6.00	7.95	6.01	6.01	5.99	5.99	6.04	6.04	6.04	0.43	20.51
8.00	8.73	8.05	8.05	8.04	8.04	8.06	8.06	8.05	5.05	40.01
10.00	9.41	10.05	10.04	10.01	10.02	10.08	10.06	10.09	18.02	-4.15
12.00	10.15	12.04	12.04	12.01	12.01	12.04	12.04	12.03	26.17	16.53
14.00	10.91	14.04	14.03	13.98	13.99	14.02	14.02	14.03	38.85	12.89
16.00	11.66	16.04	16.03	16.01	16.03	16.03	16.03	16.03	45.79	-2.95
PRESS	67.31	0.0218	0.0187	0.0122	0.0126	0.0179	0.0183	0.0220	1987	2491

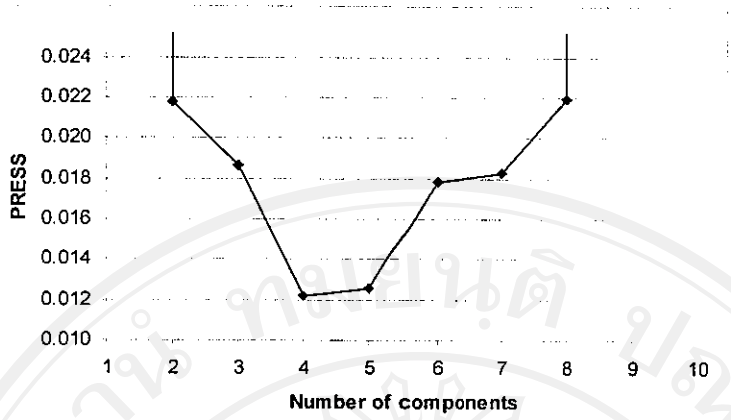


Figure 3.12 Plot of PRESS against the number of components used in the calibration step of the yellow food colorant determination.

In this study, PRESS is the criterion to evaluate the predictive abilities of the models. The number of components giving a minimum PRESS is the right number of the components for the model that can give the optimum prediction. Inspecting of the result in Table 3.8, the predicted concentrations of the yellow food colorants and the PRESS of each PCR model are shown. The results show that the PCR models using the first two to the first eight components obtain the lower values of PRESS than the PCR model using only the first component and the models using the first nine and the first ten components.

In Figure 3.12, the results show that the PCR model obtaining the lowest PRESS is the model that uses the first four components so that this model is the model that obtains the best predictive ability here in this experiment.

Considering the Eq. (1.1) in the introduction path and the PRESS results of each model in Figure 3.12, it can be implied that only the first four principal components can contain almost all of the useful information from the absorbance data.

On the other hand, in the later components, those of the components start to contain the information that is not correlated with the dependent data or noises. The PRESS is starting to increase when the number of the components used is more four components. Furthermore, the PCR models using the first nine and the first ten components obtain much values of PRESS because when the numbers of the variables used are more than the number of the samples used in the matrix calculations, the matrix inversions can cause the problems.

The results can be indicated that the PCR procedure can be applied to determine the concentrations of the yellow food colorant in the mixtures of the prepared solutions although the spectrum is the overlapping peak with the red food colorant

3.2.1.5 Simultaneous analysis of the yellow and the red food colorants of the validation samples by PCR

To study the abilities of the PCR procedure for simultaneous analysis of the components in a sample with the overlapping peaks. The PCR procedure was applied to determine both of the concentrations of the yellow and red food colorants in the mixture solutions. The training set and the validation set used in this experiment are the data sets in the section 3.2.1.1 and 3.2.1.2, respectively. The results of the predicting and the plot of the PRESS of the PCR models are shown in Table 3.9 and Figure 3.13, respectively.

Table 3.9 The simultaneous predicted results of the yellow and red food colorants in the mixture solutions using the PCR models.

The expected concentrations		The predicted concentrations of the yellow and red food colorants for the number of factors used																			
		Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red
Yellow	Red	1		2		3		4		5		6		7		8		9		10	
2.00	16.00	6.39	6.24	1.91	15.91	1.93	15.94	1.94	15.93	1.92	15.93	1.96	15.94	1.96	15.94	1.98	15.94	9.38	20.14	-4.02	0.61
4.00	14.00	7.39	7.21	4.06	14.37	4.07	14.39	4.08	14.39	4.06	14.38	4.07	14.38	4.07	14.38	4.08	14.38	-7.11	16.86	29.29	23.02
6.00	12.00	7.95	7.77	6.01	11.96	6.01	11.96	5.99	11.97	5.99	11.97	6.04	11.98	6.04	11.98	6.04	11.98	0.43	15.15	20.51	9.54
8.00	10.00	8.73	8.53	8.05	10.00	8.05	10.01	8.04	10.01	8.04	10.01	8.06	10.02	8.06	10.02	8.05	10.02	5.05	15.48	40.01	39.84
10.00	8.00	9.41	9.18	10.05	7.81	10.04	7.81	10.01	7.82	10.02	7.83	10.06	7.84	10.06	7.84	10.09	7.84	18.02	11.52	-4.15	-20.86
12.00	6.00	10.15	9.91	12.04	5.84	12.04	5.83	12.01	5.85	12.01	5.85	12.04	5.86	12.04	5.86	12.03	5.86	26.17	15.65	16.53	1.13
14.00	4.00	10.91	10.65	14.04	3.91	14.03	3.90	13.98	3.92	13.99	3.92	14.02	3.93	14.02	3.93	14.03	3.93	38.85	14.43	12.89	13.58
16.00	2.00	11.66	11.38	16.04	1.93	16.03	1.91	16.01	1.92	16.03	1.92	16.03	1.93	16.03	1.92	16.03	1.92	45.79	16.52	-2.95	-35.76
PRESS		67.31	310.3	0.0218	0.2256	0.0187	0.2416	0.0122	0.2214	0.0126	0.2144	0.0179	0.2096	0.0183	0.2090	0.0220	0.2094	1988	490.3	2492	3588
∑PRESS		377.6		0.2474		0.2602		0.2336		0.2270		0.2275		0.2273		0.2314		2478		6080	

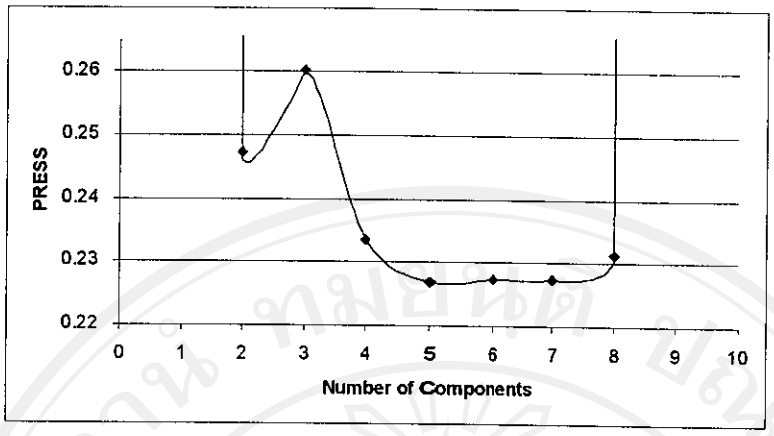


Figure 3.13 Plot of PRESS against the number of components of the simultaneous determination of the yellow and red food colorants.

Like the previous experiment, the different characteristic of the PCR models is the numbers of the components used in the calibration step. The predictive ability of each model is expressed in term of the PRESS value. The difference between the single and the simultaneous PCR analysis is that in the single analysis, the training concentration is only from the yellow food colorant but in the simultaneous analysis, the training concentrations are from both of the yellow and the red food colorants.

The results in the Figure 3.13 show that the PCR model that obtains the lowest PRESS is the model using the first 5 principal components. However, the analysis of variance with F-test probabilities of 0.75 (α = 0.25) is calculated and the results indicate that the models using the first four, five, six, seven and eight PCs can obtain the results that are not significantly different from the best one.

Those of the results can be indicated that the PCR procedure can be applied to determine the concentrations of the yellow and red food colorants in the mixture

samples of the food colorants singly and simultaneously although the spectra are the overlapping peaks and any chemical separations are not required.

Comparing to the other simple calculation methods, it can be seen that the PCR prediction is quicker and more convenient because all of the steps are performed though the computer programming automatically, and the advantages of using PCR procedure are that the full spectrum data can be used in the analysis procedure so that all of the useful information in the data are enabled in the predictions, and by using PCA procedure, some noises containing in the data can be negligible. In these studies, the known sample concentrations can be well-predicted from the spectral data without any data pretreatments.

3.3 Partial Least-Square Regression, PLS

3.3.1 Determination of the food colorants by PLS

To study the predictive ability of the PLS procedure for determining of the analyte in the spectrum containing overlapping peaks and of the simultaneous analysis, The PCR procedure was applied to determine the concentration of a yellow food colorant in the mixture solution of the yellow and a red food colorants and the concentration of both of the yellow and red food colorants of the same samples. In these studies, the PLS procedure was performed follow the process in the section 2.4.3 and the food colorant spectrum data were collected from the sequent injection experimental work which are the same sets as the section 3.2.1.

3.3.1.1 Prediction of the concentrations of the yellow food colorants of the validation samples by PLS

To study the ability of PLS for determining of the component in the samples with the overlapping peak conditions. The PLS procedure was applied for determining the concentration of a yellow food colorant in the presence of a red food colorant. The training set and the validation set used are the same sets as the previous experiments. The results of the predicting are shown in Table 3.10. The PRESS of each model is shown in Figure 3.14.

Table 3.10 The predicted concentration (ppm) of the yellow food colorants in the mixture solutions of the yellow and the red food colorants using PLS models.

The expected concentrations	The predicted concentrations of the yellow food colorants for the numbers of PCs used											
	1	2	3	4	5	6	7	8	9	10	11	12
2.00	6.36	1.89	1.92	1.93	1.93	1.94	1.94	1.94	1.93	1.93	1.93	1.94
4.00	7.35	4.05	4.05	4.06	4.05	4.06	4.06	4.06	4.06	4.05	4.05	4.06
6.00	7.90	5.99	6.00	5.99	6.01	6.02	6.01	6.01	6.01	6.01	6.01	6.01
8.00	8.66	8.03	8.04	8.04	8.04	8.05	8.04	8.04	8.04	8.04	8.04	8.03
10.00	9.32	10.03	10.03	10.02	10.04	10.04	10.04	10.04	10.04	10.04	10.04	10.02
12.00	10.05	12.03	12.02	12.01	12.02	12.02	12.02	12.02	12.02	12.02	12.02	12.01
14.00	10.79	14.02	14.01	13.99	14.01	14.00	14.00	14.00	14.00	14.01	14.01	13.98
16.00	11.53	16.03	16.01	16.01	16.03	16.02	16.02	16.02	16.02	16.02	16.02	16.01
PRESS	68.80	0.0183	0.0132	0.0107	0.0121	0.0120	0.0117	0.0117	0.0117	0.0112	0.0112	0.0096

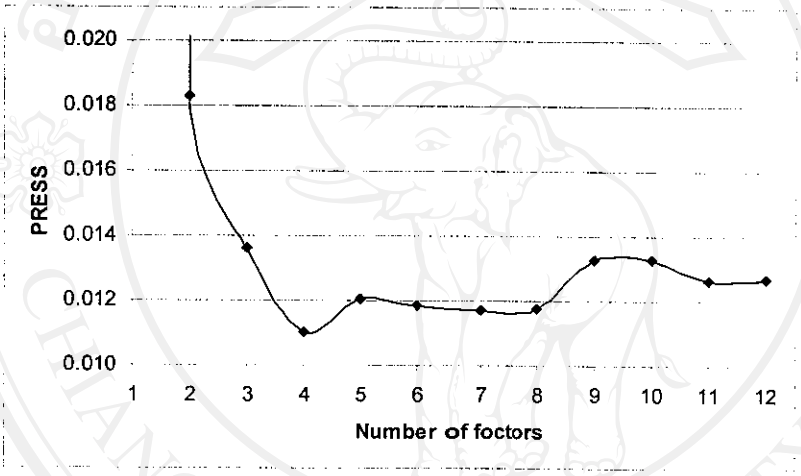


Figure 3.14 Plot of PRESS against the number of the factors of the determination of the yellow food colorants.

Investigating to the Table 3.10, the PLS models are applied to the same data set as the previous experiments in order to benchmark the performance of the predictive ability with the overlapping peak condition of the PLS models. Also, the difference of each model is the number of the factors used in the calibration steps. The results in Figure 3.14 show that the models obtaining the best PRESS is the model using the first 4 factors. However, the models using the first four, five, six, seven and

eight components are acceptable due to F-ratio tests ($\alpha = 0.25$). The results can be implied that there are many models can be used for predicting the sample concentrations, but one advantage of the models that uses the less number of the factors is that the less time consuming is needed in the computational calculations.

3.3.1.2 Simultaneous analysis of the yellow and the red food colorants of the validation samples by PLS

To study the abilities of PLS models for simultaneous analysis of components in the sample with the overlapping peaks, The PLS procedure was applied to determine the concentrations of the yellow and the red food colorants in the mixture solutions. The training set and the validation set used in this experiment are the same sets as the previous experiments. The results of the predicting and the plot of the PRESS of each model are shown in Table 3.11 and Figure 3.15, respectively.

Table 3.11 The simultaneous determination of the yellow and red food colorants in the mixture solutions using the PLS models.

The expected		The predicted concentrations of the yellow and red food colorants for the number of PCs used																							
concentrations		Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red	Yellow	Red
Yellow	Red	1		2		3		4		5		6		7		8		9		10		11		12	
2.00	16.00	6.36	6.68	1.89	15.91	1.92	15.94	1.93	15.93	1.93	15.93	1.94	15.93	1.94	15.94	1.94	15.94	1.93	15.93	1.93	15.93	1.93	15.94	1.94	15.93
4.00	14.00	7.35	7.55	4.05	14.38	4.05	14.39	4.06	14.38	4.05	14.38	4.06	14.38	4.06	14.38	4.06	14.38	4.06	14.38	4.05	14.38	4.05	14.39	4.06	14.38
6.00	12.00	7.90	8.03	5.99	11.96	6.00	11.97	5.99	11.97	6.01	11.98	6.02	11.98	6.01	11.98	6.01	11.98	6.01	11.98	6.01	11.98	6.01	11.98	6.01	11.98
8.00	10.00	8.66	8.70	8.03	10.01	8.04	10.01	8.04	10.01	8.04	10.02	8.05	10.02	8.04	10.02	8.04	10.02	8.04	10.02	8.04	10.02	8.04	10.02	8.03	10.02
10.00	8.00	9.32	9.28	10.03	7.81	10.03	7.81	10.02	7.82	10.04	7.83	10.04	7.83	10.04	7.84	10.04	7.84	10.04	7.83	10.04	7.83	10.04	7.83	10.02	7.84
12.00	6.00	10.05	9.92	12.03	5.84	12.02	5.84	12.01	5.85	12.02	5.86	12.02	5.86	12.02	5.86	12.02	5.86	12.02	5.86	12.02	5.86	12.02	5.86	12.01	5.87
14.00	4.00	10.79	10.58	14.02	3.91	14.01	3.90	13.99	3.92	14.01	3.93	14.00	3.93	14.00	3.93	14.00	3.93	14.00	3.94	14.01	3.93	14.01	3.93	13.98	3.95
16.00	2.00	11.53	11.23	16.03	1.93	16.01	1.92	16.01	1.92	16.03	1.93	16.02	1.93	16.02	1.92	16.02	1.92	16.02	1.93	16.02	1.93	16.02	1.93	16.01	1.93
PRESS		68.80	291.48	0.0183	0.2244	0.0132	0.2327	0.0107	0.2192	0.0121	0.2101	0.0120	0.2101	0.0117	0.2092	0.0117	0.2090	0.0117	0.2050	0.0112	0.2086	0.0112	0.2123	0.0096	0.2106
∑PRESS		360.28		0.2427		0.2459		0.2299		0.2222		0.2221		0.2209		0.2208		0.2167		0.2178		0.2235		0.2202	

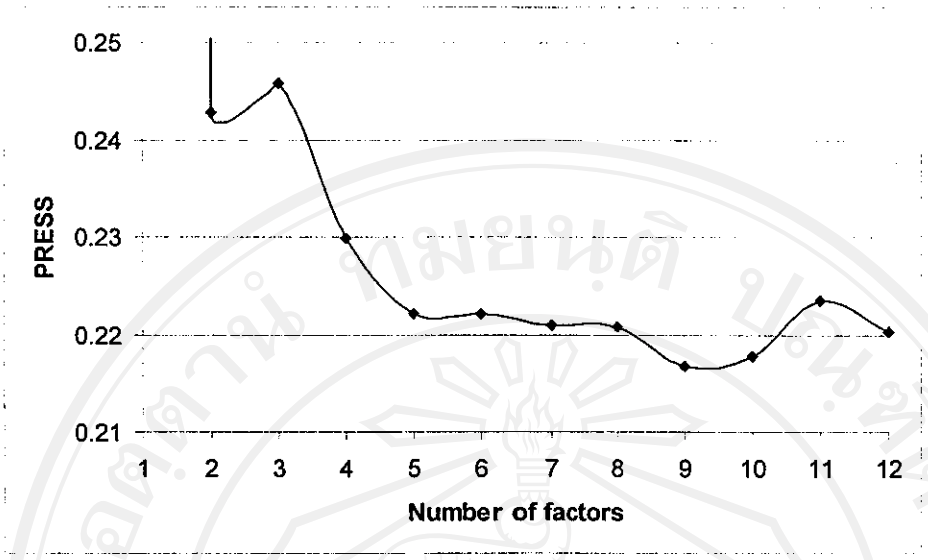


Figure 3.15 Plot of PRESS against the number of the factors used in the calibrating of the PLS models.

The results of the simultaneous determination of the yellow and red food colorants by the PLS models using the factors from one to twelve in Table 3.11 show that the PLS model that obtains the lowest value of the PRESS number is the model with nine factors. However, by using F-statistic comparison ($\alpha = 0.25$), except the first model, the values of PRESS of the models are not significantly greater than the minimum one.

The results indicated that the PLS procedure can be used for determining the single concentration of the yellow food colorant in the mixture samples while the red food colorant is donated as the interference, and the simultaneous analysis of both of the yellow and the red food colorants in the mixture samples although the spectrum are the overlapping peaks of those food colorants and any chemical separations and data pretreatments are not needed in the processes.

Both of PCR and PLS, these full-spectrum multivariate calibrations enable almost all of the useful information of the spectrum in the form of compressed variables in order to predict the concentrations. The procedures built here can be applied for determining any numbers of the analytes in the samples easily depending on the number of the analytes containing in concentration matrix that are used as the training set.

The difference between PCR and PLS is that PLS still obtains the appreciate value of PRESS when the number of the factors used in the calibrating steps are more than the number of the samples used in the training set. Theorically, PLS seem to predict better than PCR because in PLS decomposition, the inner relationship between the absorption data and the concentration data is improved. However, according to the F-test, the abilities of the PCR and the PLS to predict the concentrations of the food colorants are not different because the independent and the dependent data used in these studies are good relationship with each other. Although the peaks in the spectra are overlapping, the absorbance that is measured is in good assumption of Beer-Lambert law.

However, Training set is the important intergradient for establishing both of the procedures. The number of the known samples and their features are needed in order to calibrate the models. The good training set must contain the concentration data that span the full range of the concentrations that will be present in the unknown samples. The concentrations of each analyte in the training samples need to be studied for obtaining enough information in order to calibrate the models when there is more than an analyte in the sample.

3.4 Determination of a Bradford protein in a cow milk sample by PCR and PLS

In this part, the PCR and PLS procedures were applied for estimating of the Bradford protein concentration in cow milk samples from the peak results that were broaden from pH gradient. The measurement involves the reaction between the protein and the Bradford reagent. The data were obtained from the sequential injection experimental work [10] and the absorption patterns of the product were observed by the spectrophotometric detector.

The schematic diagram of simple sequential injection analysis with spectrophotometric detector system is shown in Figure 3.16. The SIA system was assembled from a syringe pump (XP-3000 Syringe Pump with a valve; Cavro), 10 ports multi-position valve, holding coil (PTFE tubing, 0.76 mm i.d., 1.59 mm o.d., 2500 μl) and mixing coil (PTFE tubing, 0.76 mm i.d., 1.59 mm o.d., 100 cm). The SI-grams were collected by the spectrometer detector (USB2000 Fiber Optic Spectrometer, Ocean Optics). The data collection and instrument controlled were done by using FIAlab-3000 with a personal computer.

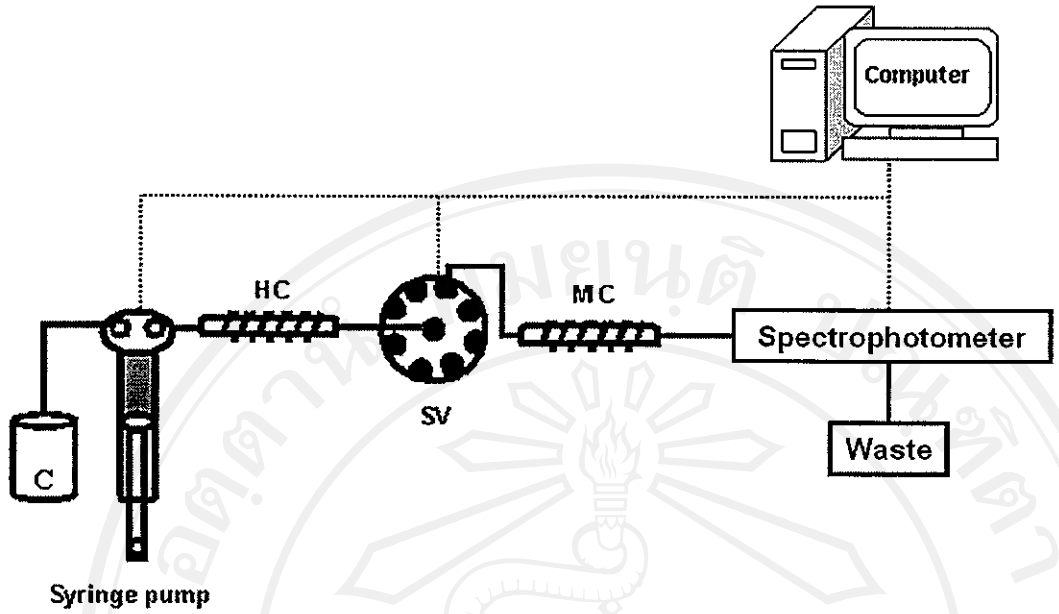


Figure 3.16 A schematic diagram of SI system C: water carrier, HC: holding coil, 2500 µl, MC: mixing coil, 100 cm, SV: 10 ports selection valve.

3.4.1 The training set

In order to construct the models, the training set samples, 0.10, 0.30, 0.50, 0.70, and 0.90 mg/ml of BSA, were prepared in deionized water. The Bradford solution was used as the reagent. The deionized water was used as the carrier. The products, protein-Bradford complexes, were monitored by the spectrophotometric detector at 595 nm. The SI-grams obtained from the training set samples are shown in Fig. 3.17.

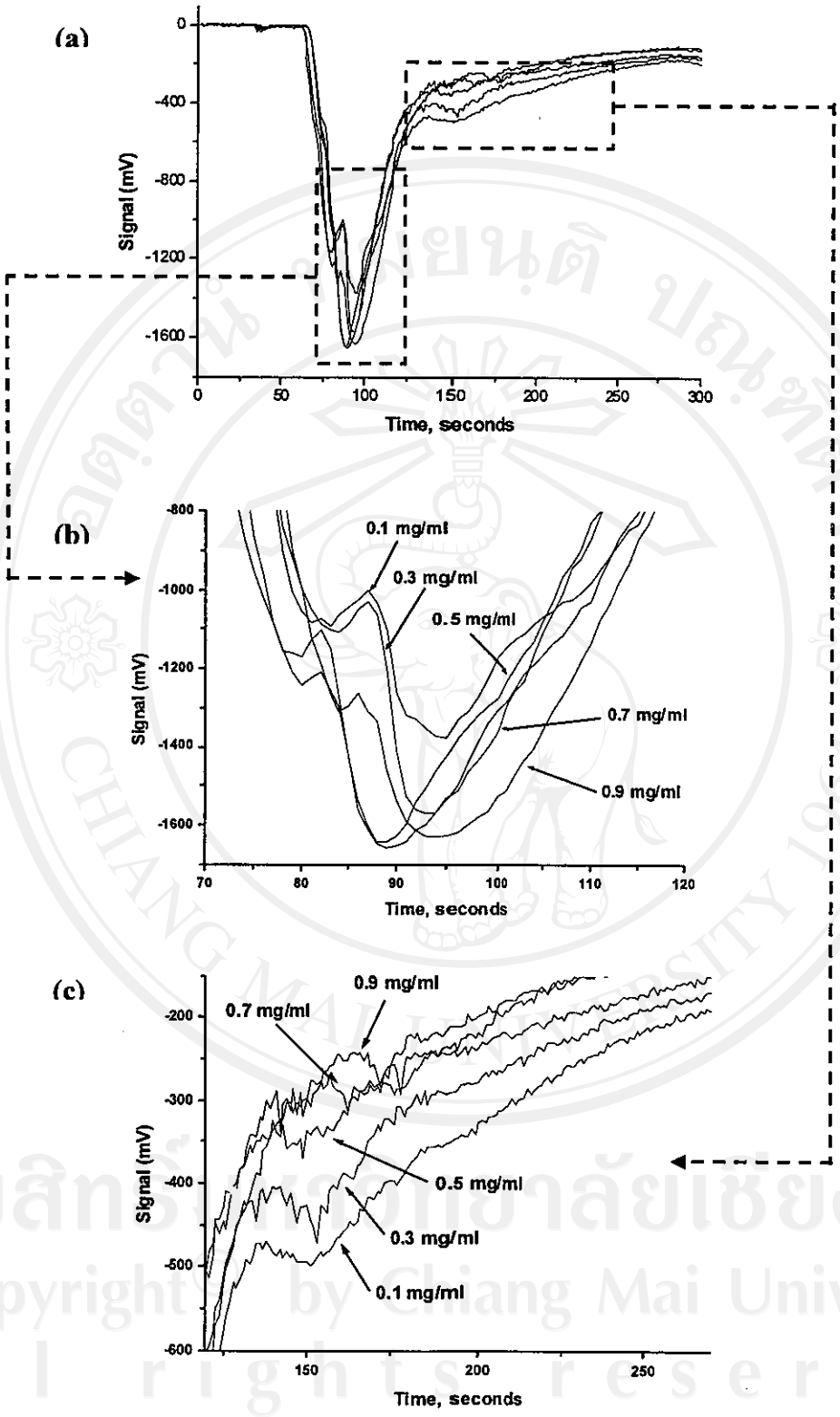


Figure 3.17 The overlay plots of the SI-grams obtained from the training samples.

The overlay plots of the SI-grams of Bradford-protein complexes are shown Figure 3.17. In Figure 3.17(b), the results can be conducted that the peaks seem to have two overlapping peaks and there are the difference among the peak height resulting from the different BSA concentrations. Figure 3.17(c) shows that the peaks are broaden due to the pH gradient. There is the difference of pH between the reactants. The Bradford reagent was prepared in the concentrated phosphoric acid (95 % H_3PO_3) and the flow directly influences the dispersion in the channel. The inter-diffusion of the sample zone and the carrier affects the size of the change in pH. The result shows that the profiles of SI-grams are changed when the concentrations of BSA are varied.

3.4.2 The validation set

The validation set samples, 0.20, 0.40, 0.60, 0.80 and 1.0 mg/ml of BSA, were prepared. These validation samples were examined as the real samples in order to measure the performance of the models. The SI-grams of the training samples are shown in Figure 3.18.

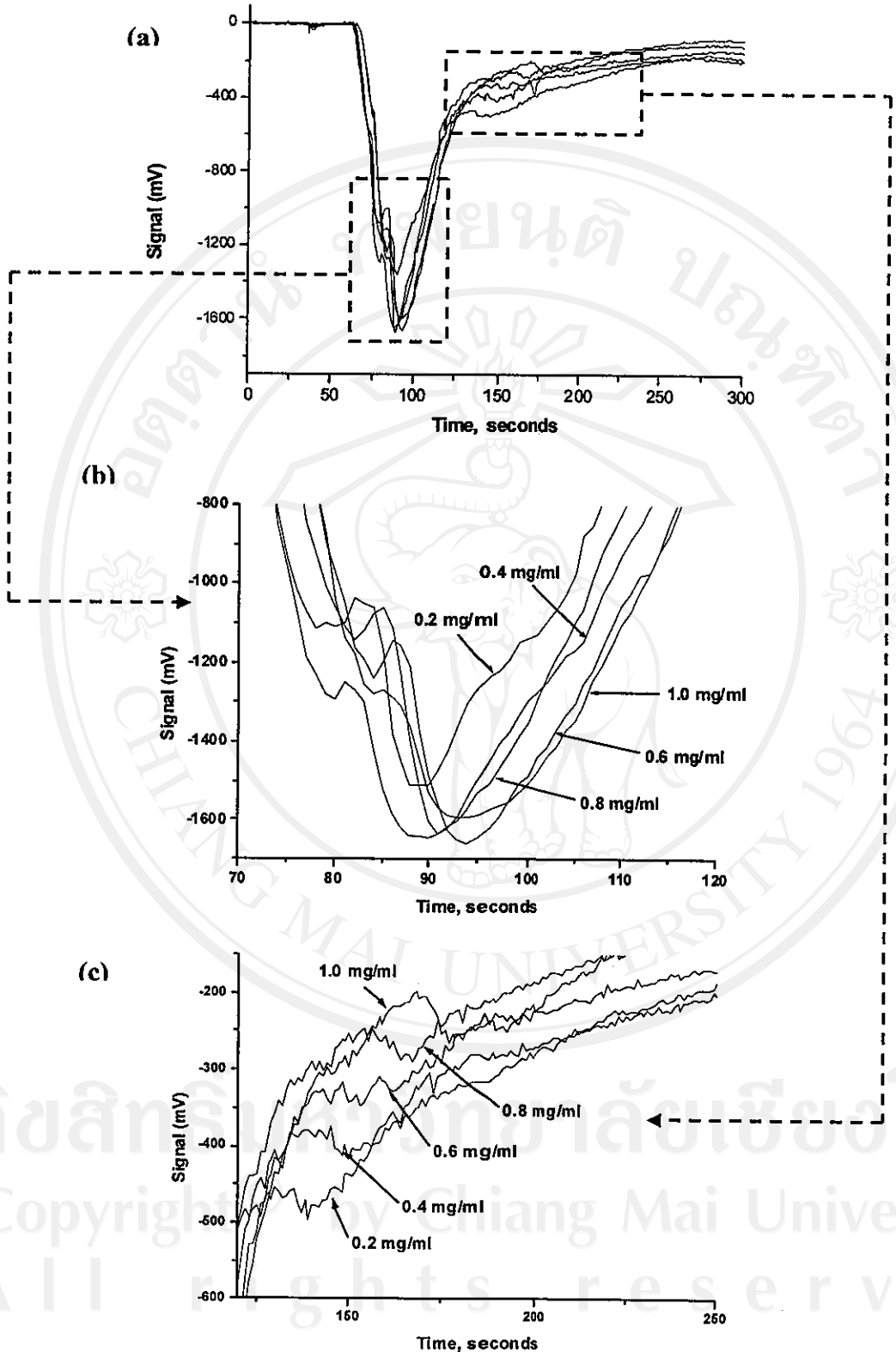


Figure 3.18 The overlay plots of the SI-grams obtained from the training set samples.

The overlay plots of the SI-grams of the validation samples are shown in Figure 3.18. The result also shows that there are the differences among the patterns of the BSA-Bradford complexes when the concentrations of the BSA are changed.

3.4.3 The calibrating of the models

3.4.3.1 The selected data in the training set

From the SI-grams in Figure 3.17, the data were investigated in order to find the suitable data that can be used for calibrating the right models. The areas of the SI-grams were selected and each of the selected area was studied as a training set in order to establish the PCR and PLS models. Then, the models were performed the Cross-Validation technique for measuring the predictive abilities. The pictures of the studied data are shown in Figure. 3.19.

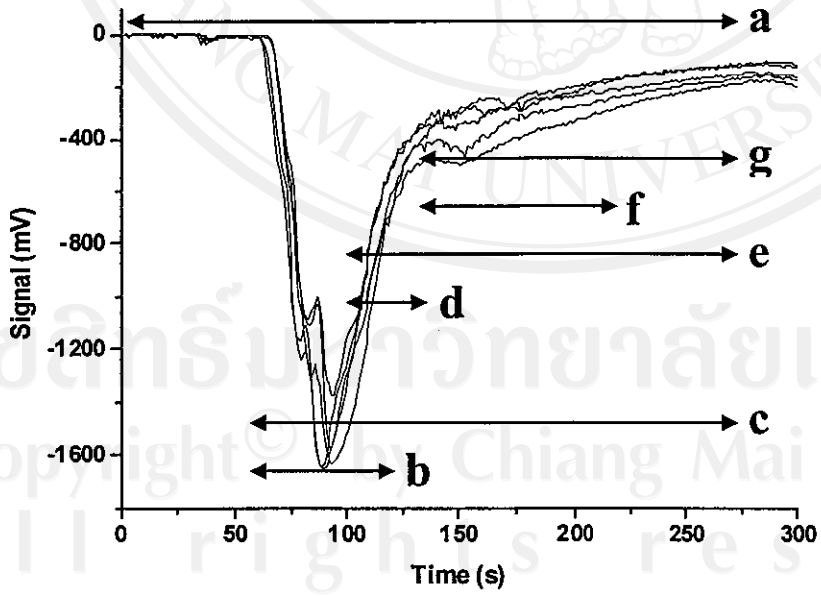


Figure 3.19 The studied areas separated from the SI-grams of the training samples.

In order to find the data those obtain the right models. Some of the data were selected. The models were generated from the selected zone of the data. Figure 3.19 shows the seven zones of the studied data.

Where:

a is a zone of a hold peak of the data (from 1 to 270 s).

b is a zone of the peak of the data (from 60 to 120 s).

c is a zone of the peak of the data and the zone of pH gradient (from 60 to 270 s).

d is the back of the peak (from 90 to 120 s).

e is the back of the peak and the zone of pH gradient (from 90 to 270 s).

f is the zone of the only pH gradient (from 120 to 220 s).

g is the zone of pH gradient (from 120 to 270 s).

3.4.3.2 The validation of PCR and PLS models by Cross-validation

When performing the PCR and PLS procedures, there are three sets of the standard samples required. The first one is used for calibrating the models. The second one is used for selecting the right number of the factors form those of the generated models. The last one is used for testing the performance of the selected models before using with the real samples. It seems that a large numbers of the standard samples are needed through the procedures but the possible number of the standard samples is restricted. Therefore, it is necessary that the standard samples are used efficiently.

In this work, Cross-Validation technique, CV, was applied to the PCR and PLS models. The procedure of CV is stated in Appendix C. The training samples themselves will be acted as they are the validation samples. First, a training sample will be taken out. Then, the calibration is performed from the remaining samples in order to estimate the concentration of the taken out sample. The process will be automatic performing by the computer programming until the last sample. The final results are the set of the predicted concentrations. After that, PRESS value is calculated from the difference between the originate concentrations and the predicted concentrations resulting from the CV process. The CV-PRESS of the PCR and PLS models estimated from the different training data are shown in Table 3.12 and Table 3.13, respectively.

Table 3.12 The CV-PRESS values for the PCR models with the different studied areas.

Number of factors	Studied area						
	a	b	c	d	e	f	g
1	0.4642	0.4345*	0.4642	0.5884	0.9353	1.0639	1.0666
2	0.1665	0.7735	0.1664	0.6227	0.0628	0.2577	0.3233*
3	0.0945*	0.7656	0.0943*	0.5130*	0.0571*	0.2457*	0.3341
4	0.2340	18.5710	0.1399	3.1104	6.0662	55.1490	2.4841
5	2.51E+00	2.38E+27	1.68E+28	6.47E+26	7.26E+28	2.24E+02	2.95E+02
6	1.24E+02	3.97E+30	3.61E+00	5.65E+27	1.78E+27	7.75E+03	5.55E+01

*the minimum value of PRESS of the models

Table 3.13 The CV-PRESS values for the PLS models with different studied areas.

Number of factors	Studied area						
	a	b	c	d	e	f	g
1	0.2648	0.3958	0.2649	0.5051	0.1379	0.1580	0.1361
2	0.0936	0.0951*	0.0935	0.4856	0.0991	0.0632	0.0561
3	0.0937	0.1066	0.0933	0.3953	0.1042	0.0639	0.0562
4	0.0952	0.1150	0.0973	0.3899	0.1048	0.0577	0.0525
5	0.0889*	0.1223	0.0965	0.4126	0.0691	0.0256*	0.0509
6	0.0956	0.1183	0.1051	0.4117	0.0701	0.0345	0.0485*
7	0.0993	0.1099	0.0796	0.3888	0.0711	0.0382	0.0495
8	0.1004	0.1099	0.0793*	0.3313	0.0782	0.0394	0.0525
9	0.1046	0.1079	0.0859	0.3300	0.0643	0.0392	0.0561
10	0.1039	0.1109	0.0846	0.2897*	0.0518*	0.0336	0.0552
11	0.1051	0.1123	0.0853	0.3349	0.0601	0.0287	0.0545
12	0.0925	0.1165	0.0835	0.3410	0.0578	0.0318	0.0572

*the minimum value of PRESS for the studied areas

Table 3.12 and Table 3.13 show The CV-PRESS values obtained by optimizing the calibration samples of the studied areas of PCR and PLS models. The minimum values of CV-PRESS of each model are shown in the marked (*) blocks. Investigating in the overlay plot of the SI-grams obtained from the training sample In Figure 3.17, it appears that the absorbance data contain the deviation due to the noises. In this case, the CV-PRESS values in Table 3.12 can be implied that the first few components (first two or three components) are contained the information that are correlated with the concentration data but when the number of the factors used are more than three factors, the variation of noises start to contain in the components. The result is that the CV-PRESS starts to increase when the later components are used. However, the CV-PRESS of the models with the factors more than 4 factors is unusually increased because of the over-fitting problem.

The other predominant advantage of PLS is that the relationship between the absorbance data and the concentration data is increased. Although the absorbance data contain the noise information, the PLS calculation can calibrate that uncorrelated data as a parameter of the independent data so that when the factors used in PLS are increased, the models can still obtain the good PRESS.

3.4.3.3 The validation of PCR and PLS models by the validation samples

In this section, the models obtaining the minimum PRESS of each selected areas are measured the predictive abilities. The validation samples in section 3.4.2 were applied with the optimized model of each studied area of the procedures. The PRESS resulting from the predictions is shown in Table 3.14.

Table 3.14 PRESS values of the selected models from the validation samples.

Selected Areas (s)	PCR		PLS	
	Number of factors	PRESS	Number of factors	PRESS
1-270	3	0.0831	5	0.0459
60-120	1	0.7041	2	0.1465
60-270	3	0.0833	8	0.0593
90-120	3	0.2252	10	0.1750
90-270	3	0.0694	10	0.0228
120-220	3	0.0485*	5	0.0201
120-270	2	0.0917	6	0.0083*

*the minimum value of PRESS for the procedures

Table 3.15 The predicted concentrations (mg/ml) of the validation samples from the best models of each chemometric technique.

[BSA] (mg/ml)	PCR			PLS		
	predicted concentrations	error	%error	predicted concentrations	error	%error
0.20	0.16	-0.04	-21.04	0.12	-0.08	-40.65
0.40	0.32	-0.08	-19.70	0.38	-0.02	-4.91
0.60	0.62	0.02	4.00	0.63	0.03	5.44
0.80	0.60	-0.20	-24.80	0.79	-0.01	-1.17
1.00	0.98	-0.02	-2.36	1.01	0.01	1.36

Expecting in the Table 3.14, the predictive abilities of the selected models from the Table 3.12 (for PCR) and Table 3.13 (for PLS) are measured in the term of PRESS of the validation samples. The PCR and PLS models that obtain the minimum of PRESS are the model that uses the data from 120 to 220 s with 3 factors and the model that uses the data from 120 to 270 s with 6 factors, respectively. The predictive results of the models are shown in Table 3.15. The results conduct that those models are the most suitable models that are established from the training data.

3.4.4 The prediction of the protein concentrations in the cow milk samples

The cow milk samples were performed by the same procedure as the validation samples with the suitable models of both PCR and PLS procedures from the previous section in order to determine the protein concentrations. The milk samples were UHT milk and brought from the market near the university. Before aspirating, the samples were diluted for 50 times with deionized water. The results of the predicting are shown in Table 3.16.

Table 3.16 Determination of protein concentrations (mg/ml) in the cow milk samples.

Sample names	PCR (range from 120 to 220s; factors = 3)				PLS (range from 120 to 270s; factors = 6)			
	predicted	concentrations	mean	SD	predicted	concentrations	mean	SD
Country Fresh	0.36	17.76	18.99	1.11	0.60	29.76	29.14	1.07
	0.40	19.90			0.59	29.75		
	0.39	19.32			0.56	27.90		
Foremost	0.26	12.77	15.72	2.56	0.55	27.69	29.56	1.63
	0.34	17.01			0.61	30.72		
	0.35	17.37			0.61	30.27		
Nhong Pho	0.35	17.63	16.93	0.95	0.57	28.35	28.22	0.62
	0.32	15.85			0.58	28.77		
	0.35	17.31			0.55	27.55		

Table 3.17 The comparative concentrations (mg/ml) of the procedures with the concentrations (mg/ml) stated by the manufacturers.

Sample names	Labeled (mg/ml)	SIA-PCR		SIA-PLS	
		mean	SD	mean	SD
Country Fresh	40	18	1.11	29	1.07
Foremost	40	15	2.56	29	1.63
Nhong Pho	32	16	0.95	28	0.62

The predictive concentrations of the milk samples are shown in Table 3.16. In Table 3.17, the results were compared with the concentrations stated by the manufacturers. Although the chemometric models can obtain the good predictive abilities of the validation samples shown in Table 3.15, the results show that the predicted results from both of the models are different from each other and they are also different from the values stated by the manufacturers. The difference between the models may be due to the selected area in the calculation steps (from 120 to 220 s for PCR and from 120 to 270 s for PLS). These discordant selections may contain the source of the matrix interference that varied the predictive results.

The difference also may be due to the method used. In these studies, the Bradford reagent was used. This reagent has some specificity with some protein but in the standard method, Kjeldahl method, all of the nitrogen in the protein is converted to ammonia in order to measure the total protein concentration. Furthermore, the turbidity in the samples can be the source of the error in Bradford method. Although the milk samples were dilute for 50 times before aspirating, there are still detectable of the turbidity signal. It is possible that the chemometric procedures would be modified in the later work in order that the turbidity data can be also used in the prediction steps.