University of Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study program/ Specialization: Computer Science	Spring semester, 2011 Open / Restricted access
Writer: Juncheng Li	…… (Writer's signature)
Faculty supervisor:Chunming Rong External supervisor(s):	
Titel of thesis: Multivariate statistical analysis with experimental data	
Credits (ECTS): 30	
Key words: Multivariate statistical analysis Heart rate Mussel Biosensor	Pages: ……42…… + enclosure: ……0…… Stavanger, ……14/06/2011…….. Date/year

Multivariate Statistical Analysis with Experimental Data

Juncheng Li

Department of Electrical and Computer Engineering

University of Stavanger

Abstract

Collected data from the sensors monitoring the environment in oil industry are various and raw, multivariate statistical analysis can turn these data into meaningful information. This paper would introduce some typical multivariate analysis methods, and investigate the data gathered in the Biota Guard exposed experiment by the means of some appropriate multivariate statistical analysis. Principal component analysis produces the principal components to represent the information of the multivariate in a reduced dimensional space; clustering analysis can group the observations of the multivariate into clusters in different ways; discriminant analysis can classifies new observations to existed clusters based on training data. These statistical analyses help us to understand the underlying information of the data from experiment and comparison of these analyses would distinguish the certain application of these methods in different situations and gives guidelines to further study.

1. Introduction

As the environmental concern and awareness of human beings increase, more and more oil companies have taken into account the environmental management, which is the major factor in the global competition among these companies. Efficient and real-time environmental management requires real-time monitoring data such as environmental measurement data and operation data. These data are collected by variety of sensors: chemical, physical sensors and even biosensors which are newly introduced. The Biota Guard chooses some mussels as biosensors to monitor the marine environment. Although these biosensors are interesting and reflect the changes of the environment in a different way, the raw data collected from the biosensors is hard to understand. Meanwhile the selection of the biosensor is also important before the deployment, since different biosensors acts differently. More sensitive mussels can react to the environmental variation more quickly and more precisely. So the study of these mussels with multivariate analysis can compare the behaviors of these biosensors at the same time, tell the different reactions of them to the surrounding changes and pick out the sensitive mussels as the biosensors in practical deployments.

Applied Multivariate Statistical Analysis (2nd_ed) (Springer, 2007 Wolfgang Härdle and Léopold Simar) presents various multivariate data analysis and introduces the reader to the wide selection of tools available for multivariate data analysis. Besides that, the book also applied these analyses to practical examples, however almost all the examples come from economical cases. *Introduction to Data Mining* only introduces the basic idea of some multivariate statistical analyses and illustrates these methods with the help of abundant figures. These two books detailedly introduced the basic idea and the algorithm of each multivariate analysis. This paper adopted the methods introduced in these books, and applied them to the practical data from experiment. The data we get is interesting and very special, the result of

the study of the data can help us to understand these biosensors-mussels and optimize the deployment of these sensors in future.

In section 2 this paper provides the methods of principal component analysis, clustering analysis and discriminant analysis, and introduces the basic ideas and algorithms of these methods; in section 3 real data is studied by the means of these different statistical analysis methods; according to results from section 3, the underlying information of the data and the performances of these methods are discussed in section 4; finally it gives the conclusion of the study in this paper.

2. Background

In the experiments of Biota Guard, there are two experiments at the same: exposed and controlled experiments. Each of them uses 7 mussels as the biosensors, as the nominal concentrations in the exposure tank varies; the heart rates of the mussels are collected every three seconds. Once observation records their heart beat rates from 7 variables, we can treat the observation as a multivariate which contains 7 variables. The variation of the nominal concentrations in the exposure tank influences the behaviors of these mussels in that thank, and the mussels in the control tank act as contrast.

Multivariate statistical analysis is a group of methods encompassing the simultaneous observation and analysis of more than one statistical variable; it helps us to understand the underlying information of the data in a simpler way and distinguish the useless data sources. The company can refine the deployment of the biosensors in return according to the results of the analysis. There are some typical methods for different aims: principal component analysis, cluster analysis, and discriminant analysis

2.1. Principal Components Analysis

Principal component analysis (PCA) has a basic objective that reducing the dimension of the multivariate data matrix, and achieves this in a different way: searching for linear combinations with the largest variances.

For a multivariate $\boldsymbol{X}$, $\mathbf{x} = \left(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p\right)^{\mathrm{T}}$ where $\mathbf{x}_1, \dots, \mathbf{x}_p$ are p variables. Low dimensional linear combinations of variables are often easier to interpret and can serve as an intermediate step in a more complex data analysis. However it is unreasonable to use one variable or the average of all the variables to represent the original multivariate. Simply reducing the dimension of the multivariate would lose the underlying information of the data matrix and fail to retrieve the nature of the multivariate. Here is a more flexible and logical way to reduce the dimension: give different weights to different variables of the multivariate and it is called as a standardized linear combination (SLC):

$$\delta^{\mathrm{T}}\boldsymbol{X} = \sum_{\mathrm{j}=1}^{\mathrm{p}} \delta_{\mathrm{j}}\boldsymbol{X}_{\mathrm{j}} \text{ and } \sum_{\mathrm{j}=1}^{\mathrm{p}} \delta_{\mathrm{j}}^2 = 1$$

Equation 2-1

We are interesting in the SLC which maximize the variance of the projection $\delta^{\mathrm{T}}\boldsymbol{X}$**:**

$$\max_{\{\delta:\ \|\delta\|=1\}} \mathrm{Var}(\delta^{\mathrm{T}}\boldsymbol{X}) = \max_{\{\delta:\|\delta\|=1\}} \delta^{\mathrm{T}}\mathrm{Var}(\boldsymbol{X})\,\delta$$

Equation 2-2

The direction of the interesting unit vector δ is given by the eigenvector γ_1 corresponding to the largest eigenvalue λ_1 of the covariance matrix $\Sigma = \mathrm{Var}(\boldsymbol{X})$.

The SLC which has the maximum variance is the first principal component (PC): $\mathrm{y}_1 = \gamma_1^{\mathrm{T}}\boldsymbol{X}$. The SLC with the second largest variance is the second principal component: $\mathrm{y}_2 = \gamma_2^{\mathrm{T}}\boldsymbol{X}$, where the eigenvector γ_2 is corresponding to the second largest eigenvalue $\lambda_2$and so on.

In order to obtain a zero mean PC variable ***Y***, we need to get the mean of the random

multivariate $\boldsymbol{X}$ and the eigenvectors from the formulas: $\mathrm{E}(\boldsymbol{X}) = \mu$ and $\mathrm{Var}(\boldsymbol{X}) = \Sigma = \Gamma\Lambda\Gamma^{\mathrm{T}}$ respectively, where $\Gamma = (\gamma_1, \gamma_2, \dots, \gamma_\mathrm{p})$ and $\Lambda = \mathrm{diag}(\lambda_1, \dots \lambda_\mathrm{p})$. $\lambda_1 \geq \cdots \geq \lambda_p$ are eigenvalues of the covariance matrix Σ with corresponding eigenvectors $\gamma_1, \dots, \gamma_p$. Then the variable $\boldsymbol{Y}$ is obtained:

$$\boldsymbol{Y} = \Gamma^{\mathrm{T}}(\boldsymbol{X} - \mu)$$

Equation 2-3

The quality of the first q PCs explain variation is given by a ration:

$$\psi_\mathrm{q} = \frac{\sum_{\mathrm{j}=1}^{\mathrm{q}} \lambda_\mathrm{j}}{\sum_{\mathrm{j}=1}^{\mathrm{p}} \lambda_\mathrm{j}} = \frac{\sum_{\mathrm{j}=1}^{\mathrm{q}} \mathrm{Var}(\boldsymbol{Y}_\mathrm{j})}{\sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{Var}(\boldsymbol{Y}_\mathrm{j})}$$

Equation 2-4

In practice, the data collected from the multivariate $\boldsymbol{X}$ which contains *p* variables after *n* observations form a data matrix $\mathcal{X}(n \times p)$. The mean μ and the covariance matrix Σ of the multivariate are replaced by the average $\overline{x}$ and empirical covariance matrix $\mathcal{S}$ respectively. That is to say, given the spectral decomposition of $\mathcal{S}$ $\mathcal{S} = \mathcal{G}\mathcal{L}\mathcal{G}^{\mathrm{T}}$, where $\mathcal{G} = (\mathcal{g}_1, \dots, \mathcal{g}_p)$ and $\mathcal{L} = \mathrm{dig}(\ell_1, \dots, \ell_p)$ the principal components are given by

$$\mathcal{Y} = (\mathcal{X} - 1_n\overline{x}^{\mathrm{T}})\mathcal{G}$$

Equation 2-5

And according to the Equation 2-4 the variance explained by the first *q* PCs is evaluated by

$$\hat{\psi} = \frac{\ell_1 + \cdots + \ell_q}{\sum_{j=1}^{p} \ell_j}$$

Equation 2-6

However variables always are measured on heterogeneous scales. Because the principal component technique utilizes the spectral decomposition of covariance matrix but not the correlation matrix, it is sensitive to scale changes. Different scales would result in different PCs, but standardization of the variables can give a robust

description of the underlying information in the data:

$$\mathcal{X}_{\mathcal{S}} = \mathcal{H}\mathcal{X}\mathcal{D}^{-1/2}$$

Equation 2-7

Where $\mathcal{D} = \mathrm{dig}\left(\mathcal{S}_{\boldsymbol{X}_1\boldsymbol{X}_1}, \dots, \mathcal{S}_{\boldsymbol{X}_\mathrm{p}\boldsymbol{X}_\mathrm{p}}\right)$ and $\mathcal{H}$ is centering matrix: $\mathcal{H} = \mathrm{I_n} - \mathrm{n}^{-1}1_\mathrm{n}1_\mathrm{n}^\mathrm{T}$

Note that $\overline{\mathrm{x}_{\mathcal{S}}} = 0$ and $\mathcal{S}_{\mathcal{X}_{\mathcal{S}}} = \mathcal{R} = \mathcal{G}_{\mathcal{R}}\mathcal{L}_{\mathcal{R}}\mathcal{G}_{\mathcal{R}}^\mathrm{T}$

Where $\mathcal{L}_{\mathcal{R}} = \mathrm{diag}\left(\ell_1^{\mathcal{R}}, \dots, \ell_1^{\mathcal{R}}\right)$ and $\ell_1^{\mathcal{R}} \geq \cdots \geq \ell_\mathrm{p}^{\mathcal{R}}$ are the eigenvalues of $\mathcal{R}$ with corresponding eigenvectors $\mathcal{g}_1^{\mathcal{R}}, \cdots, \mathcal{g}_\mathrm{p}^{\mathcal{R}}$. This PC transformations of the matrix $\mathcal{X}_{\mathcal{S}}$ are called the Normalized Principal Components (NPCs),$\mathrm{Z_j}$ is defined as

$$\mathcal{Z} = \mathcal{X}_{\mathcal{S}}\mathcal{G}_{\mathcal{R}} = \left(\mathrm{z}_1, \dots, \mathrm{z}_\mathrm{p}\right)$$

Equation 2-8

2.2. Cluster Analysis

When given a data set of observed individuals to a multivariate, we may want to know if there are some natural groups or classes of individuals. Cluster analysis develops tools and methods dealing with this case, group individuals that are "similar" according to some appropriate criterion. Once the clusters are obtained, we can use the previous analysis methods to study each group again and get a better understanding of the differences between the groups.

2.2.1. Hierarchical Clustering

Generally the groups or clusters should be as homogeneous as possible and the differences among the variance clusters should be as large as possible. The cluster analysis has two fundamental steps: choice of a proximity measure and choice of group-building algorithm.

The "similar" between the individuals are represented by the proximity; it is described by a matrix when given a data matrix $\mathcal{X}(\mathrm{n} \times \mathrm{p})$ with n individuals

(measurements) of p variables:

$$\mathcal{D} = \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{pmatrix}$$

The matrix $\mathcal{D}$ contains the measures of similarity or dissimilarity among the n individuals. If we choose the values d_{ij} as distance to measure the dissimilarity, the greater values means the less similar are the individuals; if we choose the values d_{ij} measure the proximity, the greater the proximity value means the more similar are the individuals. If the values of observations are binary, the proximity measure between two observations (x_i, x_j) where $x_i^T = (x_{i1}, \dots, x_{ip})$, $x_j^T = (x_{j1}, \dots, x_{jp})$ andx_{ik}, $x_{jk} \in \{0,1\}$ is:

$$d_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda(a_2 + a_3)}$$

Where

$$a_1 = \sum_{k=1}^{p} I(x_{ik} = x_{jk} = 1)$$

$$a_2 = \sum_{k=1}^{p} I(x_{ik} = 0, x_{jk} = 1)$$

$$a_3 = \sum_{k=1}^{p} I(x_{ik} = 1, x_{jk} = 0)$$

$$a_4 = \sum_{k=1}^{p} I(x_{ik} = x_{jk} = 0)$$

As shown in Table 1 the weighting factors δ and λ are given base choice of different algorithms:

Name	$\boldsymbol{\delta}$	$\boldsymbol{\lambda}$	$\boldsymbol{d_{ij}}$
Jaccard	0	1	$\frac{a_1}{a_1 + a_2 + a_3}$
Tanimoto	1	2	$\frac{a_1 + a_4}{a_1 + 2(a_2 + a_3) + a_4}$
Simple Matching	1	1	$\frac{a_1 + a_4}{p}$

Russel and Rao	-	-	$\frac{a_1}{p}$
Dice	0	0.5	$\frac{2a_1}{2a_1+(a_2+a_3)}$
Kulczynski	-	-	$\frac{a_1}{a_2+a_3}$

Table 1 The common similarity coefficients[i]

If the variables x_{ik} are continuous, the distance measures can be obtained by the $L_r-\text{norms}, r \geq 1$:

$$d_{ij} = \|x_i - x_j\|_r = \left\{\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r\right\}^{1/r}$$

Where x_{ik} denotes the value of the k-th variable in object i. The $L_2-\text{norms}$ is the common choice.

When the proximity measure is obtained, we turn to the building the groups. There are two basic types of clustering methods: hierarchical algorithms and partioning algorithms. The hierarchical algorithms can be divided into agglomerative and splitting procedures.

Given two objects or groups: P and Q, are united, the distance between the new group P+Q and other group R is defined:

$$d(R, P+Q) = \delta_1 d(R,P) + \delta_2 d(R,Q) + \delta_3 d(P,Q) + \delta_4 |d(R,P) - d(R,Q)|$$

The δ_j are weighting factors and vary depend on different agglomerative algorithms as shown in table 2, Where $n_p = \sum_{i=1}^{n} I(x_i \in P)$ is the number of objects in group P, same to the definition of n_Q and n_R.

Name	$\boldsymbol{\delta_1}$	$\boldsymbol{\delta_2}$	$\boldsymbol{\delta_3}$	$\boldsymbol{\delta_4}$
Single linkage	1/2	1/2	0	-1/2
Complete linkage	1/2	1/2	0	1/2
Average linkage (unweighted)	1/2	1/2	0	0
Average linkage	$\frac{n_P}{n_P+n_Q}$	$\frac{n_P}{n_P+n_Q}$	0	0

(weighted)				
Centroid	$\frac{n_P}{n_P + n_Q}$	$\frac{n_Q}{n_P + n_Q}$	$-\frac{n_P n_Q}{(n_P + n_Q)^2}$	0
Median	1/2	1/2	-1/4	0
Ward	$\frac{n_R + n_P}{n_R + n_P + n_Q}$	$\frac{n_R + n_Q}{n_R + n_P + n_Q}$	$-\frac{n_R}{n_R + n_P + n_Q}$	0

Table 2 Computations of group distances[ii].

In practice the Single and Complete linkages are frequently adopted, and a modified agglomerative algorithm has the following steps:

1. Construct the finest partition.

2. Compute the distance matrix: $\mathcal{D}$

3. Find the smallest (Single linkage)/ largest (Complete linkage) value (between objects m and n) in $\mathcal{D}$.

4. If m and n are not in the same cluster, combine the clusters m and n belonging to together, and delete the smallest value.

Back to step 3 until all clusters are agglomerated into $\mathcal{X}$ or the value in step 3 exceeds the preset level

In practice a linear search in the original distance matrix replace the computing new distance matrix in every step.

2.2.2. K-Means Clustering

K-means is a prototype-based clustering technique which defines a prototype in terms of a centroid; a centroid is usually the mean of group of points. The basic K-means algorithm is that:

The k in the term K-means is a user specified parameter which indicates the number of clusters desired. Firstly k initial centroids are chosen, each point is then assigned to the closest centroid. The group of points assigned to the same centroid forms a cluster, then the centroid of this cluster is updated based on the points in the cluster, and repeat this assignment and updating until those centroids remain the same. This

basic K-means algorithm is described as following[iii]:

1. Select K points as initial centroids.
2. Repeat
3. From K clusters by assigning each point to its closest centroid.
4. Recomputed the centroid of each cluster.
5. Until centroids do not change.

For data in Euclidean space, the proximity measure is Euclidean distance, and the objective function is to minimize sum of the squared L_2 distance of an object to its cluster centroid which is the mean of the cluster.

The initial centroids are very important; choosing initial centroids randomly may produce poor performance. Here are effective approaches: cluster the sample of points using a hierarchical technique, then k clusters are extracted from the hierarchical clustering and the centroids of those clusters are used as the initial centroids[iv].

Another approach is to choose the first point at random or the centroid of all points; select the point that is farthest from any of the initial centroids already selected as the successive initial centroid.

During the clustering, we can update centroids incrementally after each assignment of a point to a cluster. This guarantees that empty clusters are not produced since all cluster start with a single point, and if a cluster ever has only one point, then that point will always be reassigned to the same cluster.[v]

2.3. Discriminant Analysis

When known a priori about the clusters, the Discriminant analysis can be used to classify one or several observations into these known groups. Denote these groups or populations $\Pi_j, j = 1,2,\cdots,J$ and the set of methods and tools in Discriminant analysis is used to distinguish these populations and allocate an observation x to one

of these groups. We also define a set of regions R_j that if $x \in R_j$ it is identified as a member of population Π_j.

Denote the densities of each population Π_j by $f_j(x)$. The maximum likelihood Discriminant rule (ML rule) is given by allocating x to Π_j maximizing the likelihood $L_j(x) = f_j(x) = \max_i f_i(x)$. So the region R_j is defined as:

$$R_j = \{x: L_j(x) > L_i(x) \text{ for } i = 1, \dots, J, i \neq j\}$$

Theorem[vi] Suppose $\Pi_j = N_p(\mu_i, \Sigma)$.

The ML rule allocates x to Π_j, where $j \in \{1, \dots, J\}$ is the value minimizing the square Mahalanobis distance between x and μ_i:

$$\delta^2(x, \mu_i) = (x - \mu_i)^T \Sigma^{-1} (x - \mu_i), i = 1, \dots, J$$

In the case of J=2,

$$x \in R_1 \Leftrightarrow \alpha^T (x - \mu) \geq 0$$

Where $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ and $\mu = \frac{1}{2}(\mu_1 + \mu_2)$.

In practice, if the data come from multivariate normal distribution $N_p(\mu_j, \Sigma)$. And there are n_j observations in each group, the μ_j is estimated by $\bar{x}_j$, Σ is estimated by $\mathcal{S}_j$where $j \in \{1, \dots, J\}$. Then common covariance may be estimated by:

$$\mathcal{S}_u = \sum_{j=1}^{J} n_j \left(\frac{\mathcal{S}_j}{n - J}\right)$$

Where $n = \sum_{j=1}^{J} n_j$. According to the theorem, allocate a new observation x to the population Π_j which minimizes:

$$(x - \bar{x}_i)^T \mathcal{S}_u^{-1} (x - \bar{x}_i) \text{ for } i \in \{1, \dots, J\}$$

Fisher's linear discrimination function is another method in Discriminant analysis. Given a linear combination of observations $\mathcal{Y} = \mathcal{X}a$, then the total sum of squares of y, $\sum_{i=1}^{n}(y_i - \bar{y})^2$ is equal to $\mathcal{Y}^T \mathcal{H} \mathcal{Y} = a^T \mathcal{X}^T \mathcal{H} \mathcal{X} a = a^T \mathcal{T} a$. Where $\mathcal{H} = \mathcal{I} - n^{-1} 1_n 1_n^T$ is the centering matrix and $\mathcal{T} = \mathcal{X}^T \mathcal{H} \mathcal{X}$.

The Fisher's idea is to find the linear combination $a^T x$ which maximizes the ration of the between-group-sum of squares to the within-group-sum of squares.

Where the within-group-sum of squares is:

$$\sum_{j=1}^{J} \mathcal{Y}_j^T \mathcal{H}_j \mathcal{Y}_j = \sum_{j=1}^{J} a^T \mathcal{X}_j^T \mathcal{H}_j \mathcal{X}_j a = a^T \mathcal{W} a$$

And the between-group-sum of squares is:

$$\sum_{j=1}^{J} n_j \left(\bar{y}_j - \bar{y}\right)^2 = \sum_{j=1}^{J} n_j \left\{a^T \left(\bar{x}_j - \bar{x}\right)\right\}^2 = a^T \mathcal{B} a$$

Finally the projection vector maximizes the ratio: $\frac{a^T \mathcal{B} a}{a^T \mathcal{W} a}$ is the eigenvector of $\mathcal{W}^{-1}\mathcal{B}$ that corresponding to the largest eigenvalue.

3. Multivariate analysis with real data sets

Section 2 introduces some typical multivariate statistical analysis, in this section these methods are applied to result of the real experiments of the company Biota Guard. In the exposed experiment, 7 mussels are placed in the exposure tank and the tank is deposited in a manipulative environment. The nominal concentrations of the water where these mussels live varies every three days, meanwhile the heart rates of each mussel are recorded as the monitoring data. Other 7 mussels in control tank acts as a contrast and their heart rates also are measured in the control experiment.

Section 3.1 describes the data sets measured in the exposed experiment. Data set is represented by principal components in section 3.2. Section 3.4 tries to group the observations from different phases in experiment. Base in the clustering results in section 3.4, section 3.5 classifies data to the known clusters.

3.1. Description of data sets

As the Figure 3-1 shows, water in exposure tank is sampler every three days. The nominal concentration in the water starts from 0 mg/l, then rises slowly to 0.0125

mg/l, 0.06 mg/l, 0.125mg/l, 0.25mg/l, and finally reaches the peak 0.5mg/l. The nominal concentration in the last four days falls to 0 mg/l. According to the nominal concentrations of the water samples, the experiment involves seven phases with different concentrations.

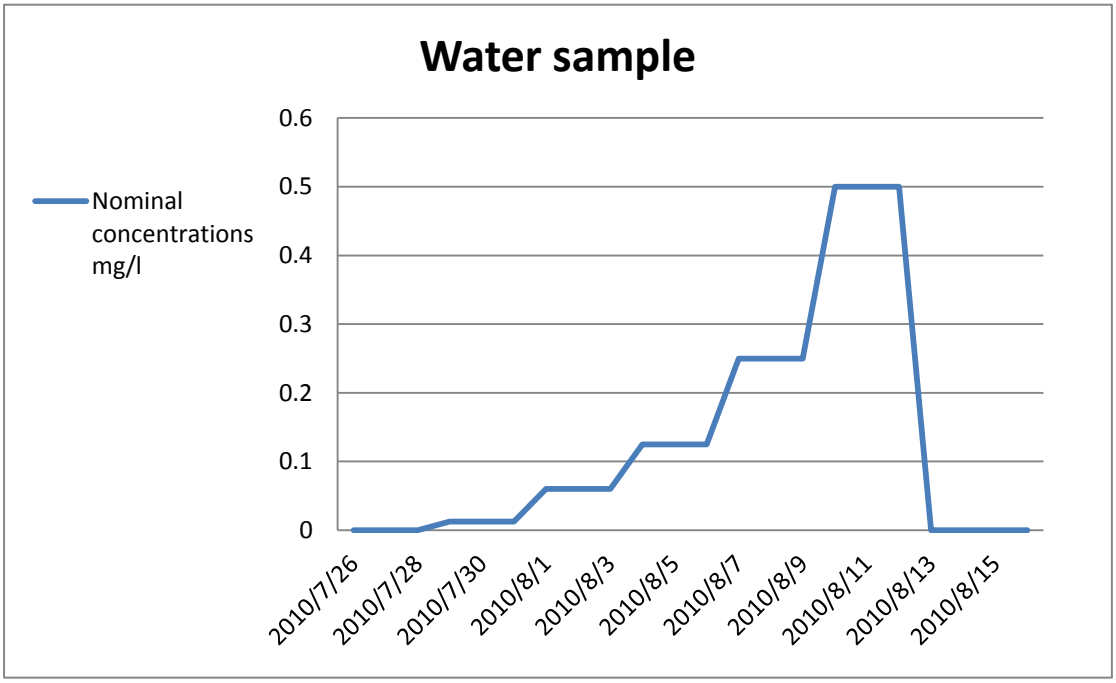

Figure 3-1 Nominal concentration in exposure tank

Besides the water is sampled, the raw heart beat signals of each mussel in the experiments are collected and then calculated as the heart rates of mussels. Every three seconds the heart rates of seven mussels in the same experiment are recorded; these seven variables form a multivariate and once record is an observation of the multivariate. These two experiments last 22 days, all the observations during these days are stored in documents. Dealing with all the observations in one experiment as a data matrix is unreasonable and it is hard to investigate the huge data matrix. Considering that the living environment of these mussels in the exposure tank varies during the exposed experiment and goes through seven phases according to the nominal concentration in water samples. It is obviously that the heart rates of these mussels change due to the variation of the living condition, and then we can group these observations into different data matrix. Partitioning these observations from

the same environmental condition into a group is possible, and these observations in the same group form to a data matrix. Finally we get seven data matrices.

3.2. Comparison of data sets

This paper starts with some descriptive techniques, trying to give a short outline of the data sets. 7 variables represent the heart rates of each mussel, and comparing these variables with each other can tell the difference of these mussels. Figure 3-2 shows the distributions of each mussel in the exposed experiment when the nominal concentration is 0 mg/l. We can find the physiological properties of these mussels:

- Mussel 4, and 6 have higher heart rate medians than the others which are 8.662 and 8.58 separately;
- Mussel 2, 5, and 7 have lower heart rate medians than others which are 6.385, 6.3 and 6.118 separately;
- Mussel 5 and 6 is more spread out mussel than the others;
- Mussel 1 and 3 has the most outliers among them.

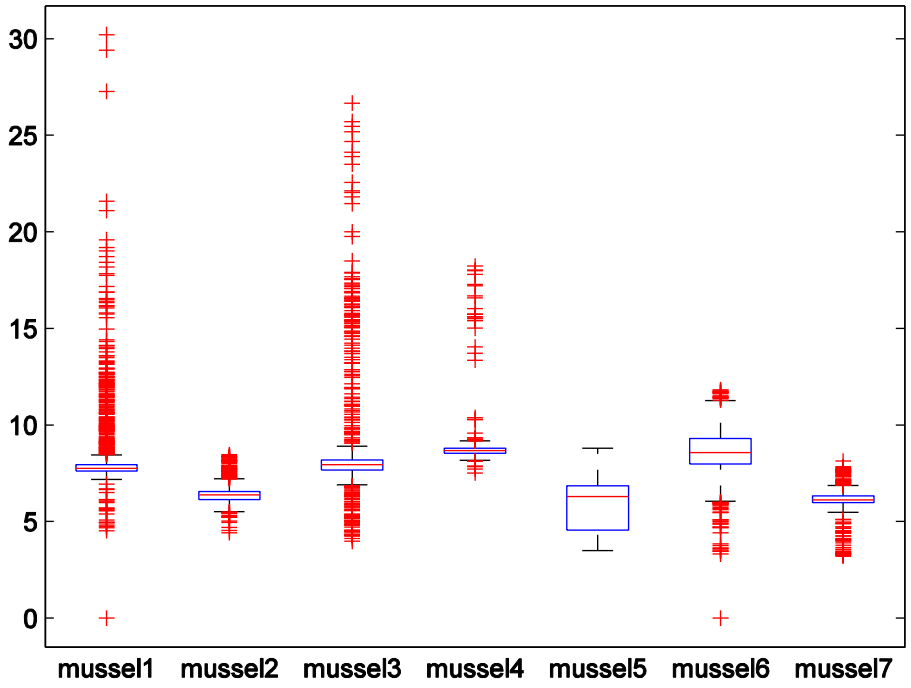

Figure 3-2 The heart rates of 7 mussels in exposed experiment phase 1

When the nominal concentration rises to 0.0125 mg/l, situation changes and is shown in Figure 3-3:

- Almost all the mussels become more spread out, especially the mussel 3;
- The mussel 1, 3 and 5 rise their heart rate medians, and the medians of the others descend;
- Mussel 1 and 4 get more outliers, but mussel 3 has no outlier since it is extremely spread out.

m

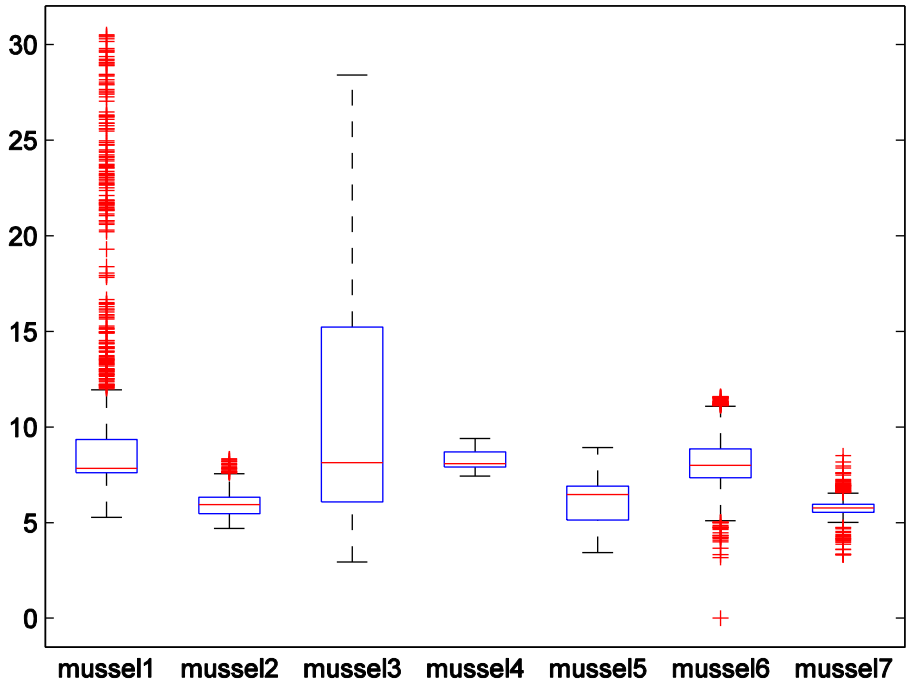

Figure 3-3 The heart rates of 7 mussels in exposed experiment phase 2

As the living environment goes worse, Figure 3-4 shows these mussels act in a different way:

- The heart rate median of mussel 1 goes up to 12.614; it is particularly higher than the value in the previous phase. Although there is no outlier in the data set, it is highly spread out;
- Mussel 3, 4 and 6 get more outliers, while mussel 7 gets less outlier;
- The heart rate medians of each mussels change as their own trends: these mussels which raise the medians in previous phase have higher median and these mussel which descend their medians in previous phase have lower medians.

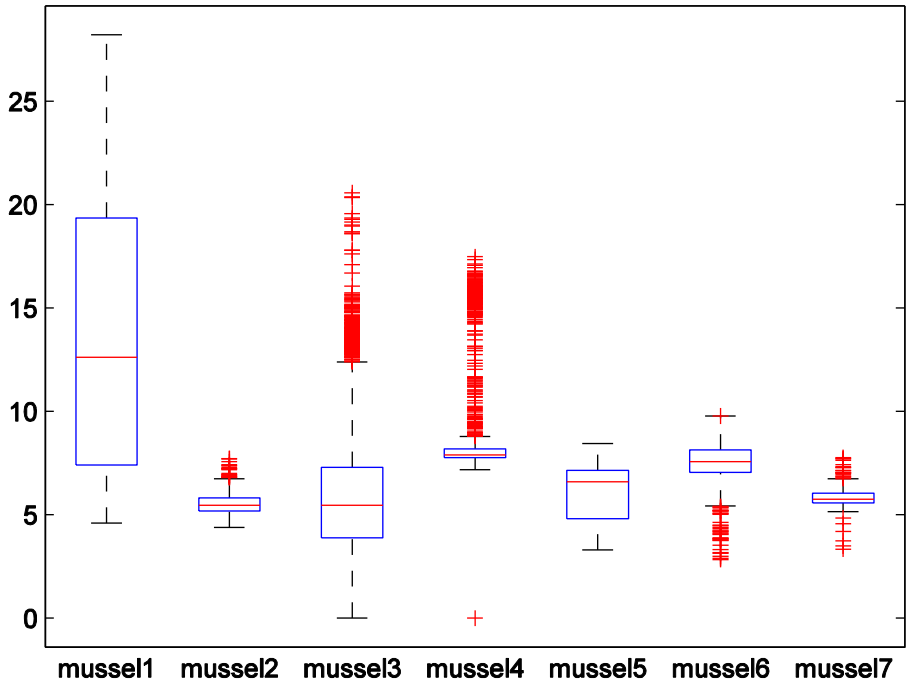

Figure 3-4 The heart rates of 7 mussels in exposed experiment phase 3

Figure 3-5 represents the situation when the nominal concentration gets to 0.125 mg/l; the data set has the following properties:

- The heart rate medians of the mussel 1 and 5 continue to get higher, and the heart rate medians of the mussel 3 and 4 continue to get down;
- The mussel 2, 6 and 7 act abnormally since the medians stop declining and rise slightly;
- The mussel 3 and 4 get less outliers and the mussel 5 and 7 get more outliers.

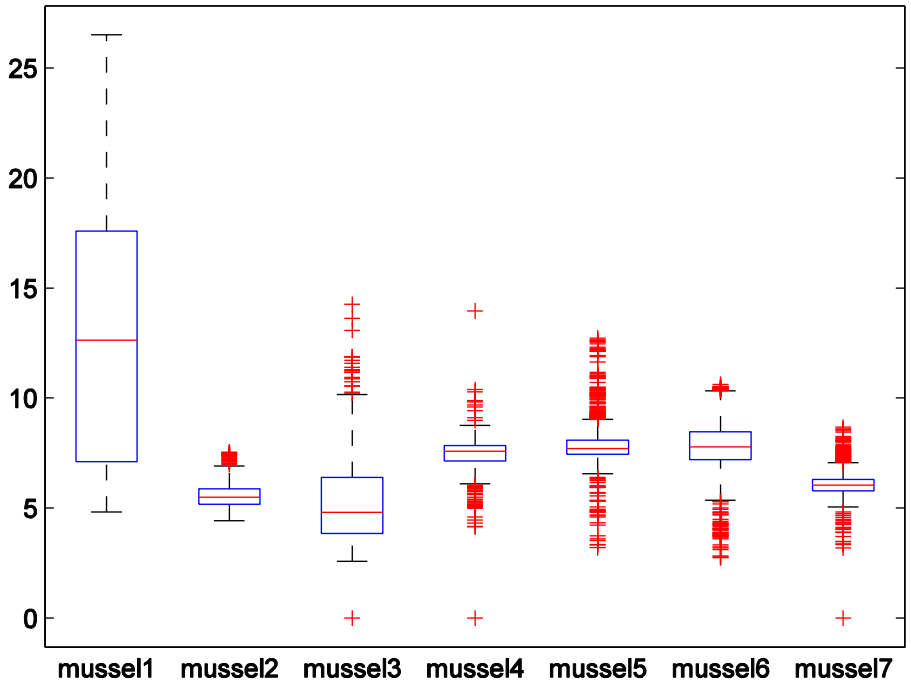

Figure 3-5 The heart rates of 7 mussels in exposed experiment phase 4

When the nominal concentration gets to 0.25 mg/l, the data set has the following properties and that is shown in Figure 3-6:

- The heart rate medians of the mussel 1 and 5 continue to get higher, and the heart rate medians of the mussel 3 and 4 continue to get down;
- The heart rate median of the mussel 2 declines slightly, it seems that the its behavior tend to be stable;
- The mussels 6 and 7 have higher medians, this fellow their trend in the previous phase.

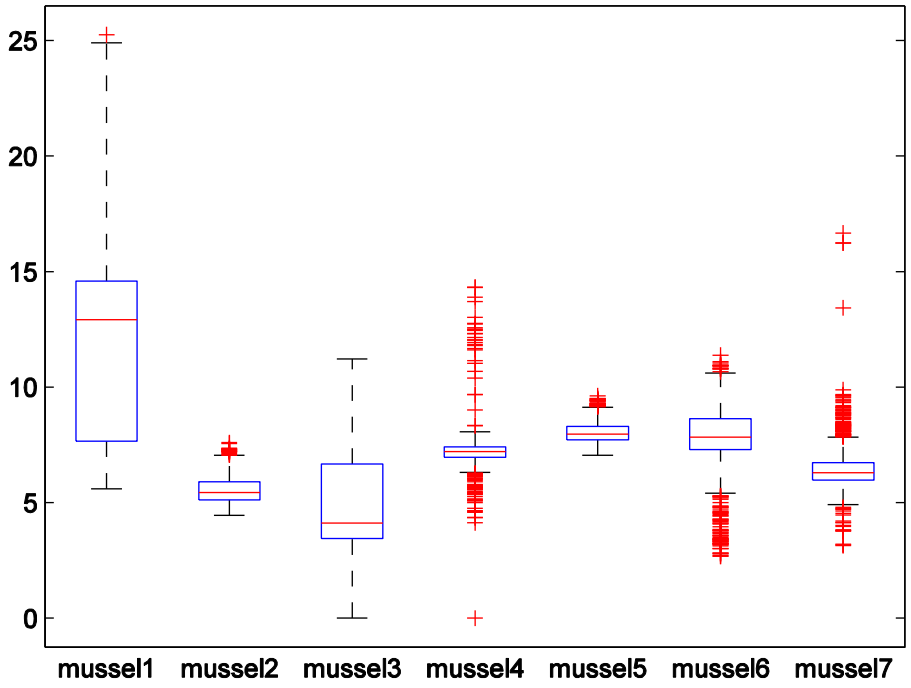

Figure 3-6 The heart rates of 7 mussels in exposed experiment phase 5

Then the nominal concentration gets to the peak 0.5 mg/l, Figure 3-7shows the distribution of the data set in phase 6:

- The heart rate medians of the mussel 1 and 5 continue to get higher, especially the median of mussel 5 rises from 6.294 to 7.895;
- Except the mussel 4, all the medians of the other mussels get higher. The median of mussel 4 always decline as the environment gets worse;

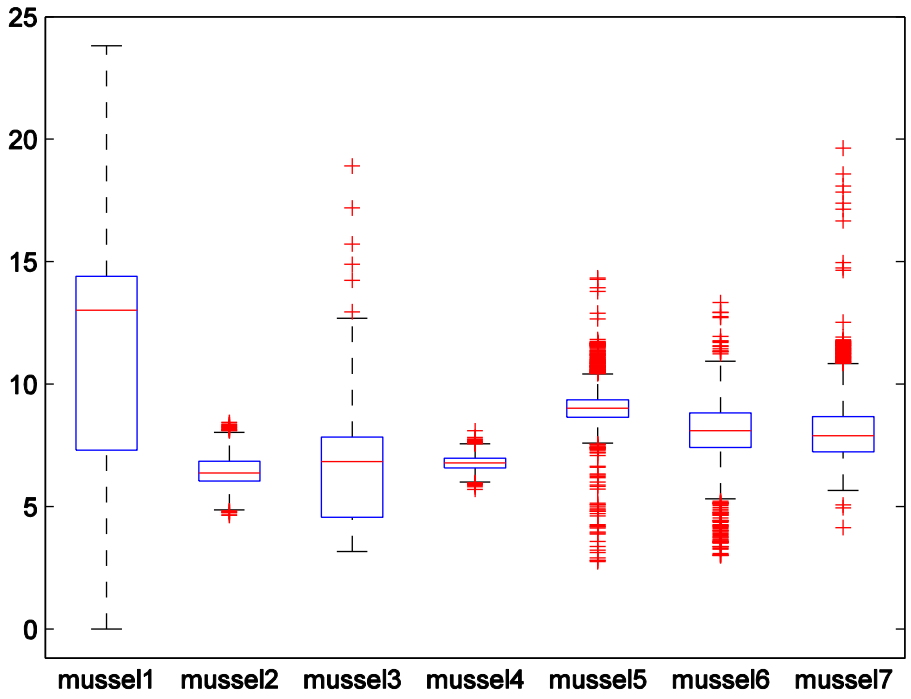

Figure 3-7 The heart rates of 7 mussels in exposed experiment phase 6

Finally, the circumstance returns to normal and the nominal concentration in the water is 0 mg/l. we can find that in Figure 3-8:

- Except the mussel 3, all other mussels get lower heart rate median;
- The mussel 3 reacts intensively, its heart rate median rise from 6.829 to 15.721 and the data is more spread out;
- As the living environment stops getting worse and finally returns to normal, these mussels get more outliers and that maybe a reaction to the change of nominal concentration.

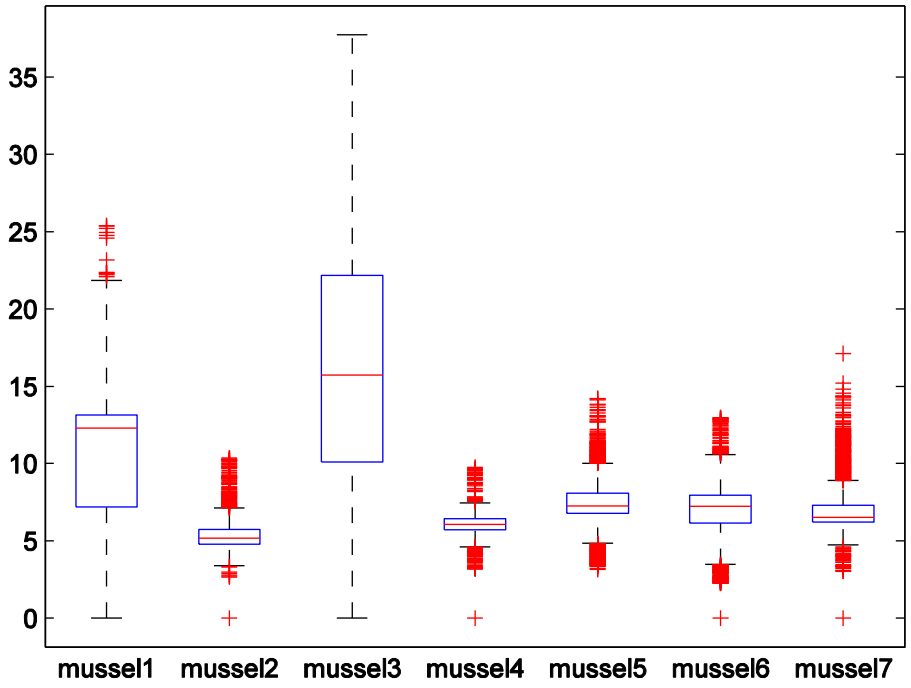

Figure 3-8 The heart rates of 7 mussels in exposed experiment phase 7

3.3. Principal component analysis

Now turn to multivariate methods. By the means of the multivariate statistical analysis we can study the variables at the same time. These variables are treated as a measurement to the system we are studying. In this paper, the marine environment is the system under investigation and the 7 mussels are the special means which reflect the variation of the system. Each observation of the multivariate represents the current state of the system. However if we try to use these variables to describe the system, we may find it is hard to plot these observations since the system is measured in a 7-dimensional space. PCA helps us to find a lower dimensional way to descript the system.

In order to investigate the principal components when the marine environment varies, two data sets in the two continuous phases are selected to estimate the PCs.

At first, the data sets from the phases where the nominal concentration in the water is 0 mg/l and 0.0125mg/l are selected. All these 7 variables measure the heart rates of mussels in the same unit, it is naturally that implement the PC analysis with these two data sets. Recall that in section 2.1 the vector of eigenvalues of the covariance matrix of these two data sets is

$$\ell = (59.8875, 31.3092, 3.6435, 2.9817, 0.9838, 0.7168, 0.4776)^{\mathrm{T}}$$

And first three corresponding $\mathcal{g}_i$ are given by the columns of the matrix

$$\mathcal{G} = \begin{pmatrix} 0.0774 & 0.9946 & 0.0580 \\ -0.0219 & -0.0334 & 0.1060 \\ 0.9949 & -0.0795 & 0.0313 \\ -0.0245 & -0.0206 & 0.0372 \\ -0.0192 & 0.0423 & -0.5399 \\ -0.508 & -0.0335 & 0.8304 \\ -0.131 & -0.0053 & 0.0456 \end{pmatrix}$$

The first column

$$\mathcal{g}_1 = (0.0774, -0.0219, 0.9949, -0.0245, -0.0192, -0.508, -0.131)$$

Is the first eigenvector and gives the weights used in the linear combination of the original data in the first PC. The first PC is dominated by the third variable that is the third mussel.

The second column

$$\mathcal{g}_2 = (0.9946, -0.0334, -0.0795, -0.0206, 0.0423, -0.0335, -0.0053)$$

tells that the second PC is dominated by the first variable.

The third column

$$\mathcal{g}_3 = (0.0580, 0.1060, 0.0313, 0.0372, -0.5399, 0.8304, 0.0456)$$

shows that the third PC is described by the difference between the sixth variable and the fifth variable. With the PCs these observations are plotted in a new coordinate system. Figure 3-9 shows the observations onto the first three principal components and the variability explained by each principal component. In these plots the observations from the phase 1 are marked by the sign '+' and the rest form the phase 2 are marked by the sign 'o'. The first three principal components explained

94.8402% variability of these two original data sets.

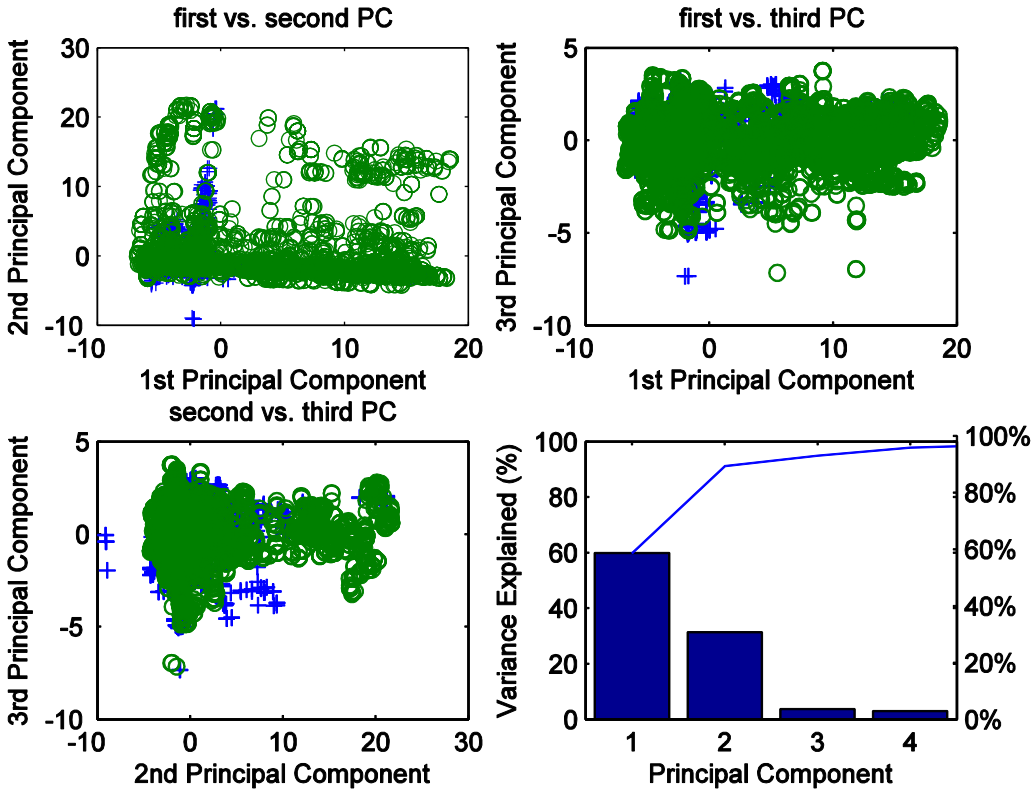

Figure 3-9 Principal components of the data sets in phase 1&2

As the experiment gets to phase 3, two data sets from phase 2 and 3 are merged to perform the principal component analysis. The vector of eigenvalues of the covariance matrix of these two data sets is

$$\ell = (48.9388, 41.7382, 4.5559, 2.1742, 1.9387, 0.4062, 0.2481)^{\mathrm{T}}$$

And first three corresponding $\mathcal{g}_i$ are given by the columns of the matrix

$$\mathcal{G} = \begin{pmatrix} 0.9353 & 0.3530 & 0.0082 \\ -0.0137 & 0.0028 & -0.0067 \\ -0.3530 & 0.9342 & 0.0482 \\ 0.0086 & -0.0473 & 0.9969 \\ 0.0144 & -0.0054 & 0.0548 \\ -0.0095 & -0.0179 & 0.0192 \\ 0.0043 & -0.0071 & 0.0168 \end{pmatrix}$$

And we can find that during the experiment shifting from phase 2 to phase 3 the first three PCs are dominated by the first variable, third variable and fourth variable separately. These three PCs explained 95.2329% variability of these two original data sets. Figure 3-10 shows the observations are plotted in the new system.

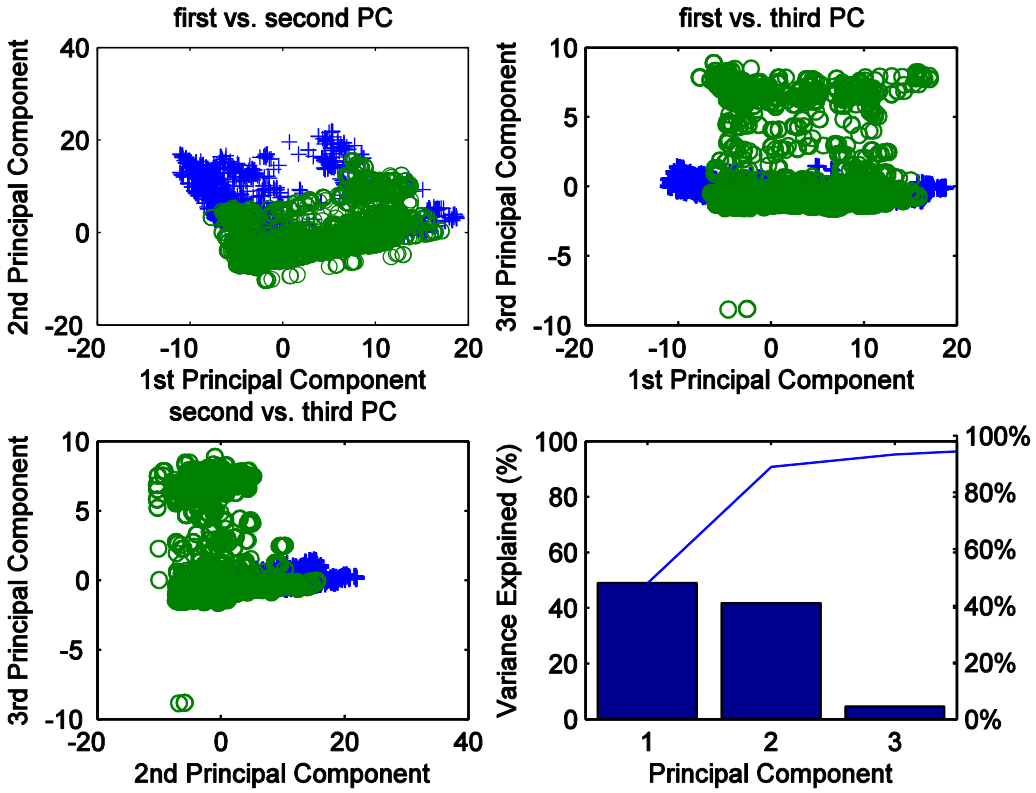

Figure 3-10 Principal components of the data sets in phase 2&3

The same analysis is applied to the data sets from phase 3 and 4. The vector of eigenvalues of the covariance matrix of these two data sets is

$$\ell = (70.0470, 14.9497, 8.2788, 3.4448, 2.3006, 0.5259, 0.4533)^{\mathrm{T}}$$

And first three corresponding g_i are given by the columns of the matrix

$$\mathcal{G} = \begin{pmatrix} 0.9947 & -0.1008 & 0.0113 \\ 0.0046 & 0.0228 & -0.0107 \\ 0.1012 & 0.9896 & -0.0877 \\ -0.0053 & 0.0782 & 0.9675 \\ -0.0126 & -0.0508 & -0.2360 \\ 0.0145 & -0.0224 & -0.0110 \\ -0.0032 & -0.0276 & -0.0112 \end{pmatrix}$$

It shows the similar result of the previous analysis: the first three PCs are dominated by the first variable, third variable and fourth variable separately and 93.2755% variability is explained. Figure 3-11 shows the result of the principal component analysis.

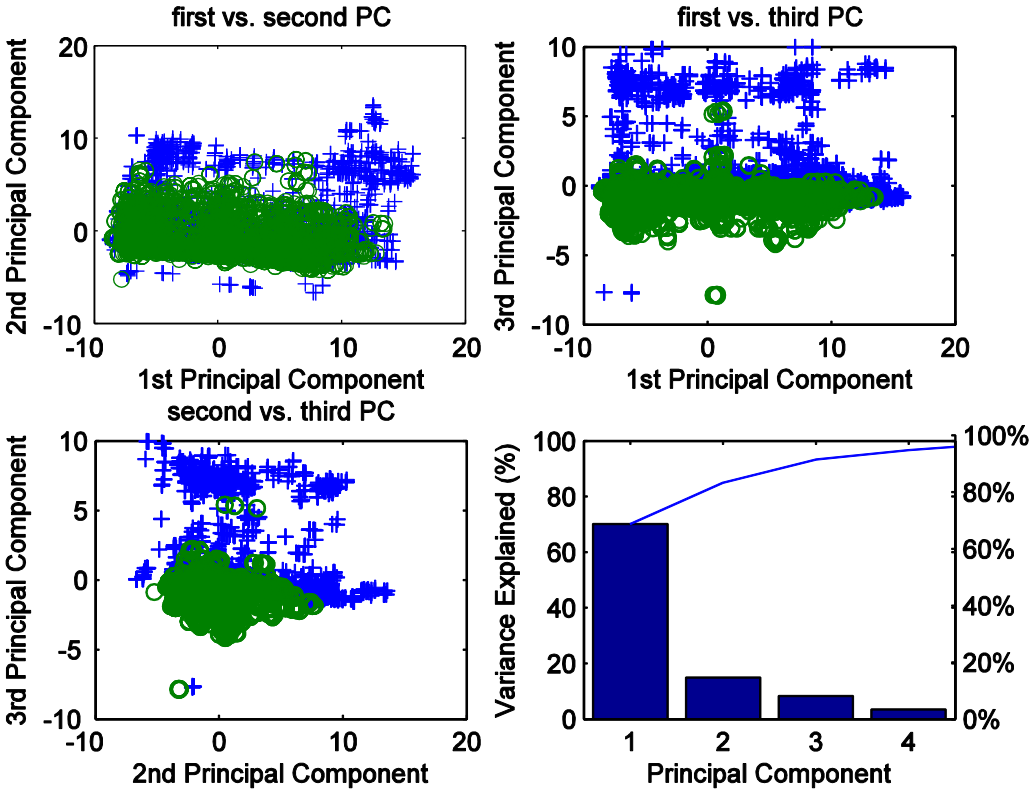

Figure 3-11 Principal components of the data sets in phase 3&4

Go on with the principal component analysis, the vector of eigenvalues of the covariance matrix of these two data sets from the phase 4 and 5 is

$$\ell = (76.9533, 11.9110, 4.7217, 2.4202, 1.9662, 1.1743, 0.8534)^{\mathrm{T}}$$

And first three corresponding $\mathcal{g}_i$ are given by the columns of the matrix

$$\mathcal{G} = \begin{pmatrix} 0.9999 & -0.0107 & -0.0069 \\ -0.0046 & 0.0192 & -0.0342 \\ 0.0100 & 0.9942 & -0.0909 \\ -0.0010 & 0.0455 & 0.0383 \\ -0.0071 & -0.0198 & 0.0165 \\ 0.0079 & 0.0912 & 0.9928 \\ 0.0002 & 0.0178 & -0.0568 \end{pmatrix}$$

There is a difference in the result: the third principal component is dominated by the sixth variable. 93.5859% variability is explained and is shown in the Figure 3-12.

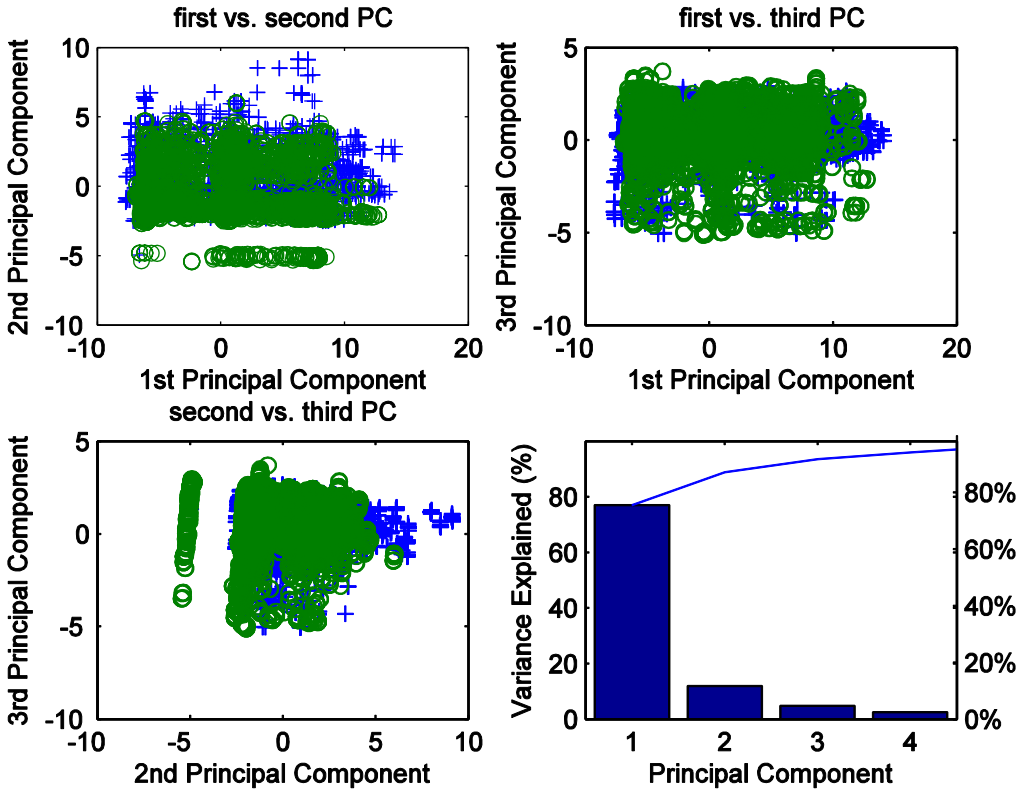

Figure 3-12 Principal components of the data sets in phase 4&5

When the environment goes to the worst condition, the vector of eigenvalues of the covariance matrix of these two data sets from these consequent phases is

$$\ell = (60.1945, 22.5148, 6.5491, 5.6506, 2.1723, 1.7149, 1.2038)^{\mathrm{T}}$$

And first three corresponding $\mathcal{g}_i$ are given by the columns of the matrix

$$\mathcal{G} = \begin{pmatrix} 0.9998 & 0.0102 & 0.0034 \\ -0.0164 & 0.0947 & 0.2063 \\ -0.0095 & 0.9357 & -0.3300 \\ -0.0061 & -0.0471 & -0.0492 \\ -0.0036 & 0.1159 & 0.3009 \\ -0.0085 & 0.1105 & 0.6025 \\ 0.0044 & 0.2959 & 0.6266 \end{pmatrix}$$

The situation for the first two principal components is the same, but the third principal component is described by the difference between the third variable and the sum of the last three variables. 89.2584% variability is explained by the three PCs and it is shown in the Figure 3-13.

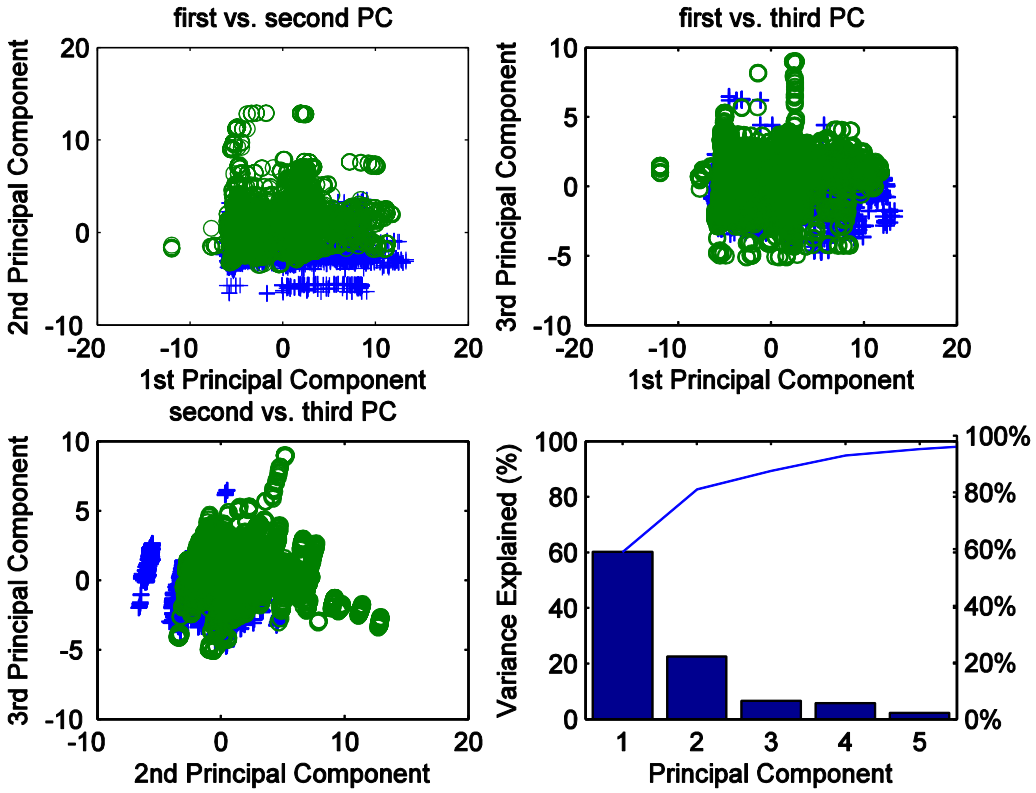

Figure 3-13 Principal components of the data sets in phase 5&6

Finally the experiment ends with nominal concentration is 0 mg/l once again. The vector of eigenvalues of the covariance matrix of these two data sets is

$$\ell = (73.7749, 14.3179, 5.8716, 3.3349, 1.2345, 0.9955, 0.4707)^{\mathrm{T}}$$

And first three corresponding $\mathcal{g}_i$ are given by the columns of the matrix

$$\mathcal{G} = \begin{pmatrix} -0.0076 & 0.9957 & 0.0411 \\ -0.0382 & 0.0187 & 0.1254 \\ 0.9952 & 0.0134 & 0.0370 \\ -0.0295 & 0.0266 & 0.0743 \\ -0.0642 & 0.0508 & -0.0219 \\ -0.0303 & -0.0463 & 0.9873 \\ -0.0466 & 0.0509 & 0.0222 \end{pmatrix}$$

This result is that the first three principal components are dominated by the third variable, first variable and the sixth variable separately and 93.9645% variability is explained. Figure 3-14 shows the details of the result.

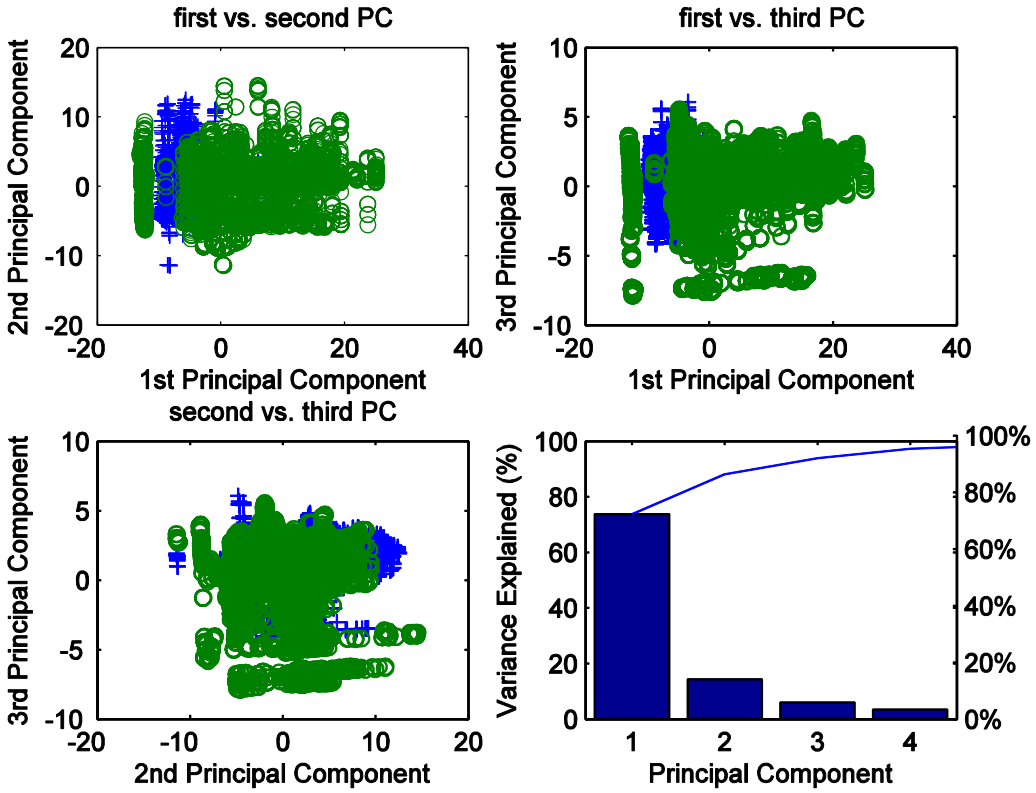

Figure 3-14 Principal components of the data sets in phase 6&7

3.4. Cluster analysis

By now, the writer achieved to find the principal components for these variables during the exposed experiment. Consider that the variables represent the environment and we are interesting in how well the variables reflect the variation of the marine circumstance, the next point in this paper is to investigate whether the observations from different phases in the experiment can form into clusters naturally based on the changing environmental conditions.

At first, the writer tries to group these data sets with the hierarchical clustering. Since the hierarchical clustering group data by creating a cluster tree or dendrogram, it has an advantage that the data can be grouped over a variety of scales since the multilevel hierarchy is specified by users. Although this advantage is attractive and gives flexible solutions to the problem, hierarchical takes time and memory space to compute the

similarity/dissimilarity between the observations and the proximity between these new objects. When the writer study two datasets together and attempt to employ the hierarchical method to cluster the datasets, it is disappointing that the lack of computer memory becomes the barrier for the further investigation.

In the experiment, each dataset in an experimental phase contains over 20,000 observations. When more than one datasets in different phases are selected to study, them emerge to a larger dataset which totally contains over 40,000 observations. In the first step of the hierarchical clustering analysis, the distance or similarity between these observations are calculated and stored in a vector. For a data matrix $\mathcal{X}(n \times p)$ with n measurements or objects of p variables, the proximity or similarity among objects is described by a matrix $\mathcal{D}(n \times n)$. The matrix $\mathcal{D}$ is a symmetric matrix and these diagonal components are equal 0, so in practice the reduced distance information can be stored in a vector and it contains $\frac{(n-1)\times n}{2}$ components. However more than 40,000 observations are selected in once analysis, and then over 8×10^8 components are created in a vector. In most personal computers, the operation system cannot support such a large data. In the trial hierarchical clustering analysis, the MATLAB function terminates unexpectedly and throws an exception 'out of memory'. This requirement of enough memory for computation forces the writer to give up investigating these data sets with the hierarchical clustering analysis.

Then this paper turns to k-means clustering. Although unlike hierarchical clustering, k-means clustering creates only a single level of clusters, it operates on actual observations rather than the larger set of dissimilarity measures. This avoids the failure of 'out of memory' during the investigation. By now the k-means clustering analysis is suitable for the scale of these large data sets, and then the problem is that if there is distinct difference between these observations collected from experiment phases and this method can group these observations into clusters which consistent with their original phases.

As the nominal concentration of the water varies continuously and impacts the

behaviors of these mussels, their physiological characteristic may also change. Unlike a physical or chemical sensor, the heat rate of a mussel cannot accurately reflect the tiny variation of the environment. That is physiological characteristic of the mussels change slowly during the exposed experiment and there is no obvious distinction between the observations in two continuous phases.

At first I consider the data sets from phase 1 and 5, there may be sufficient difference between the observations form the original phases. There are 20,880 observations in phase1 and 25,920 observations in phase5. In MATLAB k-means function provides several users-define parameters to specify how to cluster the objects. Generally the desired number of clusters and the distance measure of objects are defined for the analysis. Since data sets come from two phases, it is reasonable to set the desired number of clusters is two. And specify the 'city block' distance measure for test.

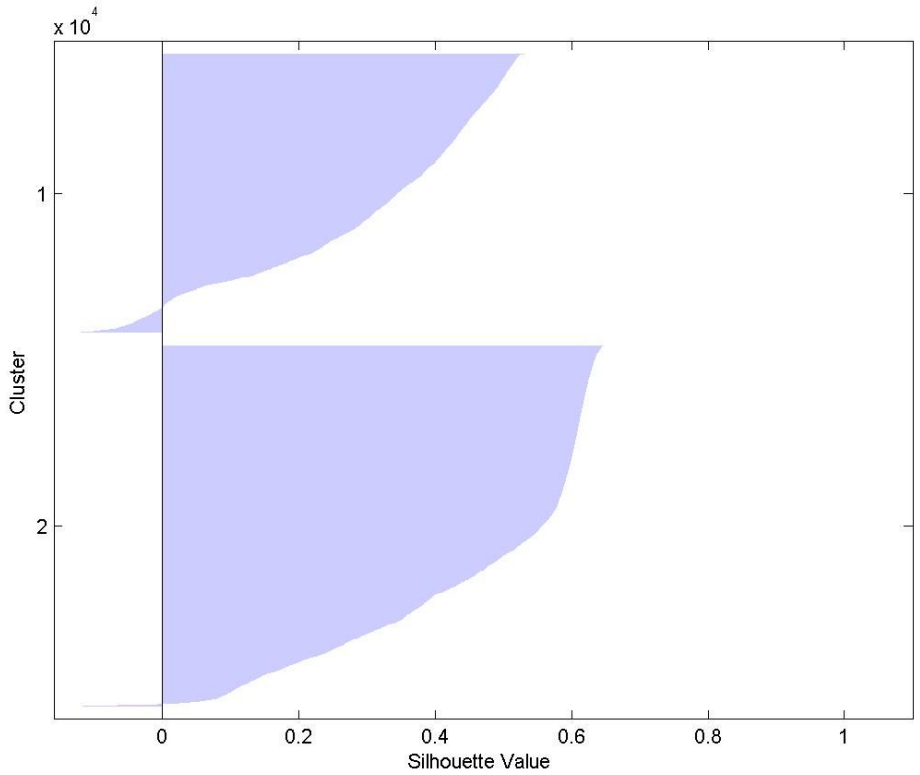

Figure 3-15Silhouette for 2 clusters measured by city block

The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure ranges from +1, indicating points

that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster.[vii] In the Figure 3-15, most points in each cluster have a value less than 0.6, indicating that the clusters are not well separated to each other. Especially the first cluster contains many points with low silhouette values and a few points with negative values, indicating that those two clusters are not well separated. 'Correlation' distance measure treats the observations as sequences of values and the clusters minus the sample correlation between points. Figure 3-16 shows the result of clustering by the 'correlation' measure.

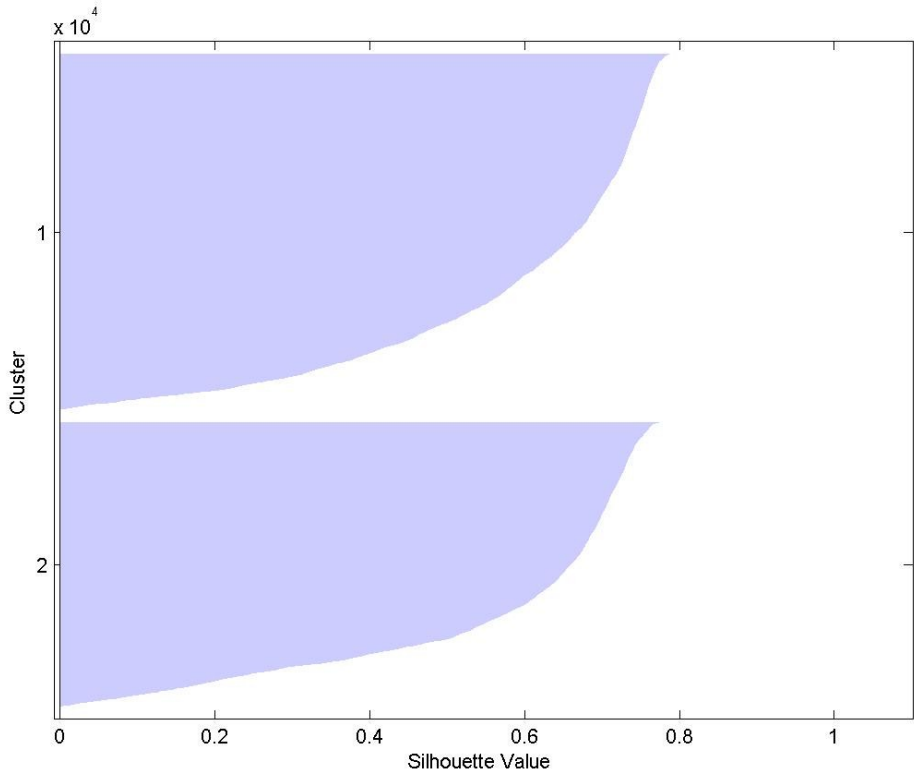

Figure 3-16Silhouette for 2 clusters measured by correlation

There are more than half of points in each cluster have a value greater than 0.6 and no negative value, it shows that clustering these data by the 'correlation' measure is better than 'cityblock' measure. If consider each observation as a vector in the multidimensional space, 'cosine' measure minus the cosine of the angle between points.

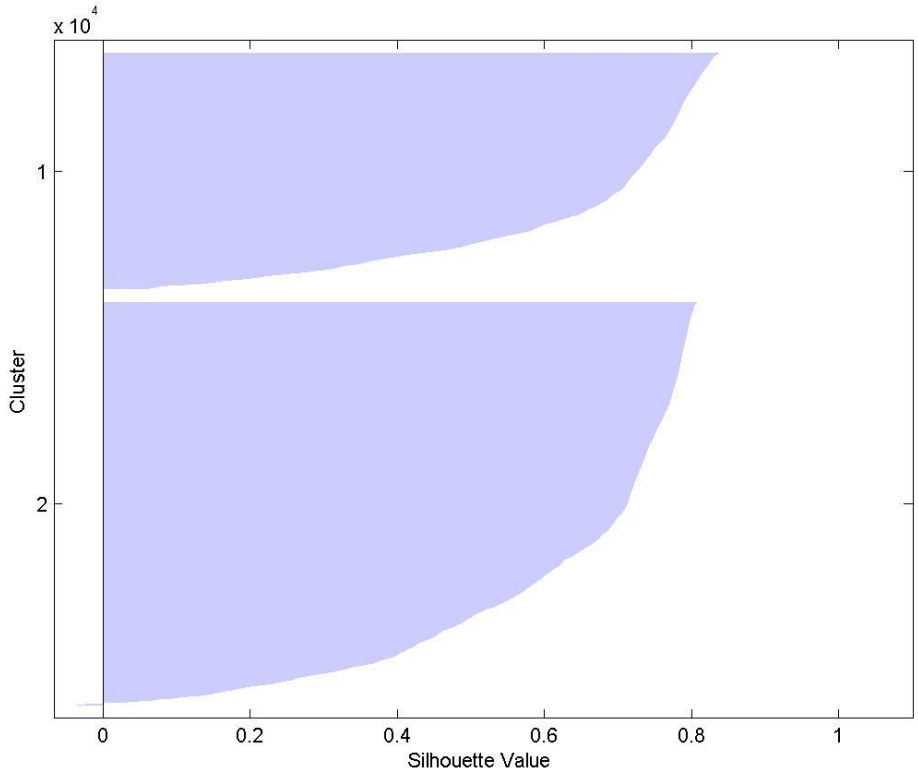

Figure 3-17Silhouette for 2 clusters measured by cosine

The clustering by the 'cosine' measure shown in Figure 3-17 seems similar to the result of clustering by the 'correlation' measure, but which one provides a better solution? The quantitative way to measure the solution is to check the average silhouette values for the two cases. The mean of the silhouette value of the 'correlation' measure solution is 0.5847 and that of the silhouette value of the 'cosine' measure solution is 0.6360, that is to say clustering the data by the 'cosine' measure than by 'correlation' measure. Don't jump to conclusions; it is hastily to say clustering by 'cosine' measure is the best solution since Squared Euclidean distance is a possible measure for the distance between observations. Figure 3-18 plots the silhouette for each cluster after clustering analysis by 'sqEuclidean' measure.

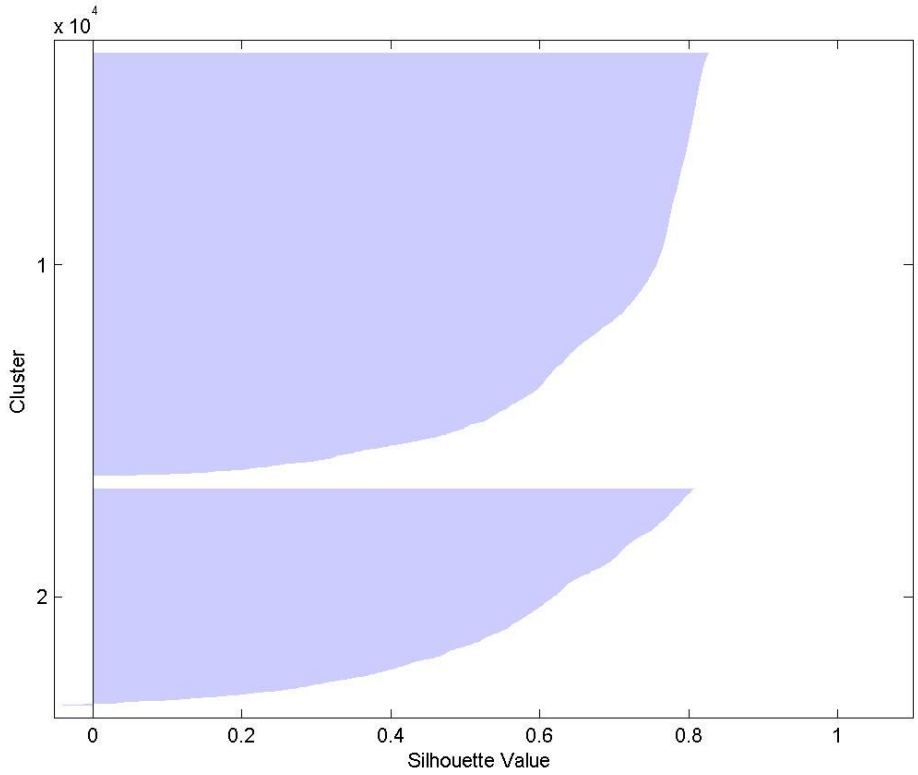

Figure 3-18 Silhouette for 2 clusters measured by sqEuclidean

The mean of the values of all the points in two clusters is 0.6497; it is greater than the means of these previous solutions. After the comparison of mean values of different solutions, clustering by 'sqEuclidean' is the most suitable way for the data sets in the exposed experiment. Does the number of clusters affect the quality of clustering?

Increase the number of clusters to see if k-means clustering can find a better grouping of the data. The result of the clustering the data sets into 3 groups by the 'sqEuclidean' measure is shown in Figure 3-19, almost all the values of points in cluster3 is less than 0.6 and some points in cluster 1 and 3 have a negative values. The mean value of points in 3 clusters is 0.4744; this is fairly less than that of above clustering.

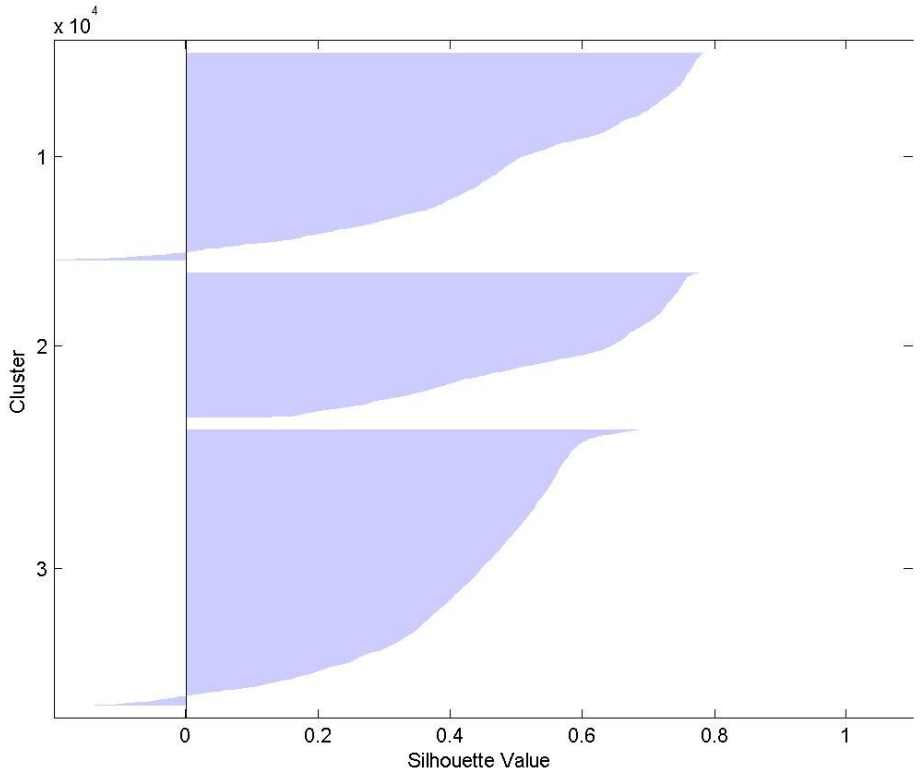

Figure 3-19 Silhouette for 3 clusters measured by sqEuclidean

By now, we can find that for two data sets from two different experimental phases measuring the squared Euclidean distance between the observations and partition them into 2 clusters is the best choice to clustering the data.

In the exposed experiment, the heart rates of these seven mussels are treated as a multivariate, during the clustering analysis the computation of the distance between these observations from the multivariate costs time since these observations locate in 7-dimensinal space. Considering section 3.3 gives a method to represents the multivariable in a reduced dimensional space, it may save time to calculate the distances of the observations in a reduced dimensional space. Continue the investigation with the data sets from phase 1 and 5; we work out the principal components of the data at the first step. The first three eigenvectors of the covariance matrix of the data are given in $\mathcal{G}$.

$$\mathcal{G} = \begin{pmatrix} 0.9059 & 0.4059 & 0.1040 \\ -0.0715 & 0.0553 & 0.1183 \\ -0.3553 & 0.8825 & -0.2620 \\ -0.1206 & 0.1232 & 0.3299 \\ 0.1694 & -0.1852 & -0.5324 \\ -0.0648 & 0.0625 & 0.6973 \\ 0.0214 & 0.0025 & -0.1673 \end{pmatrix}$$

Now the original data set can be represented in a three dimensional space using the first three principal components, the locations of the observations from two distinct phases are plotted in a three dimensional space in Figure 3-20 where observations from phase 1 in red and others from phase 2 in blue.

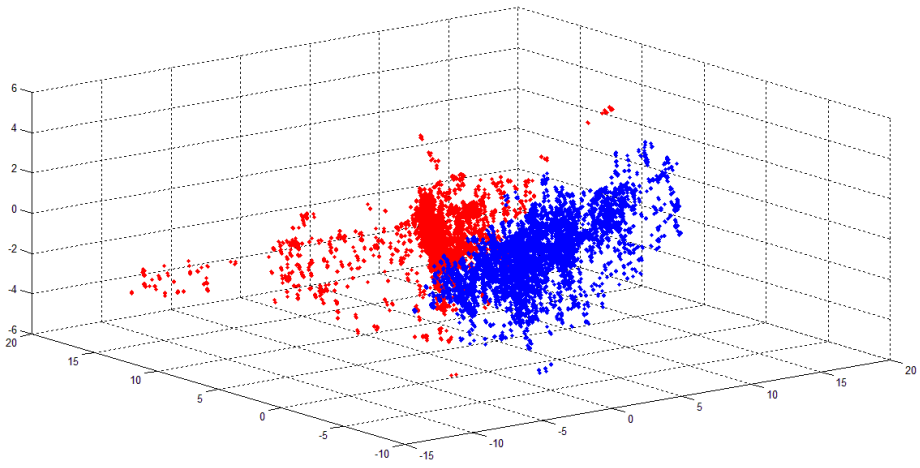

Figure 3-20Observations from phase 1 and 5

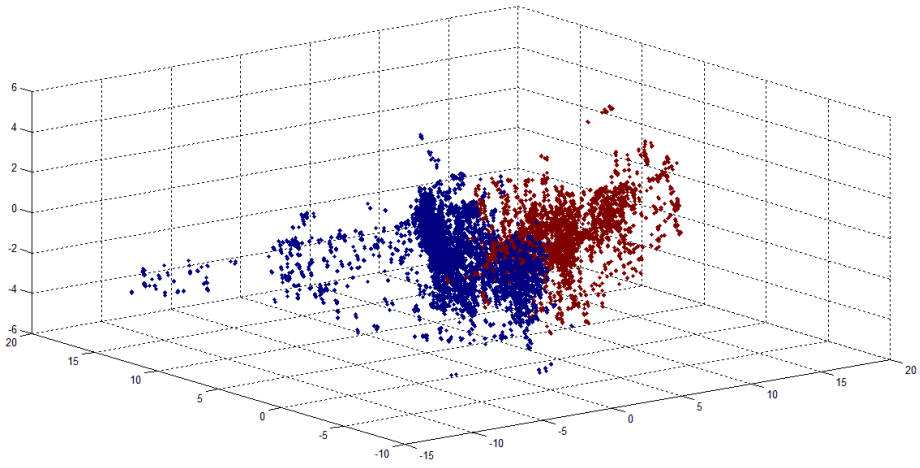

Figure 3-21 Observations clustered by the original data sets

After k-means clustering the original datasets into 2 groups, the observations in

group 1 are presented by PCAs in blue and the other points in brown shown in Figure 3-21 from group 2. Compare these two figures, we can find that the groups generated by the k-means largely match the original case where observations come from different phases. Only a bank of observations belong to phase 5 are allocated to a wrong group and are separated from the other observations from phase 5. Clustering the observations according to their reduced information would save time, and the result is shown in Figure 3-22.

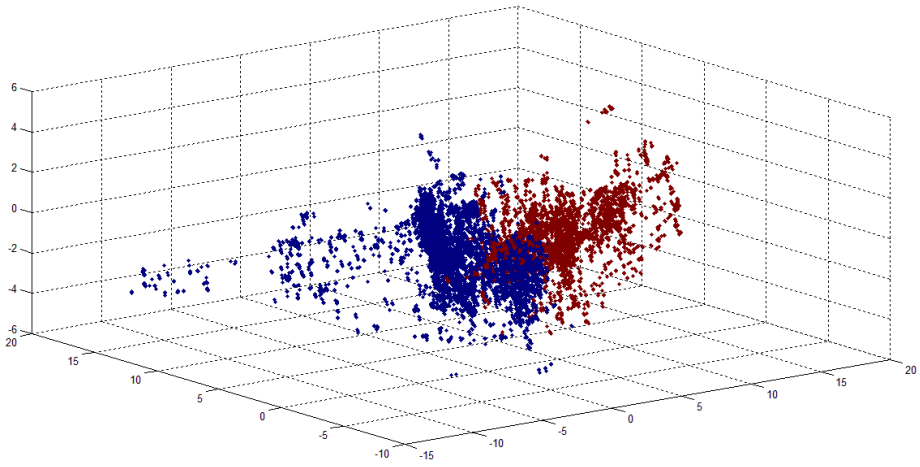

Figure 3-22 Observations clustered by the PCAs

Base on the reduced information of the original data, the observations are grouped into two clusters in the same way as clustering with the original data sets according the comparison of Figure 3-21 and Figure 3-22.

3.5. Discriminant Analysis

The data collected from the exposed experiment can be treated as training data to estimate the parameters of discriminant functions of the predictor variables. Base on the estimated parameters, Discriminant functions determine boundaries in predictor space between various classes and classify a new data to one of the various classes. Go on the investigation with data sets from phase 1 and 5, in section 3.4 k-means

clustering analysis group the data from these two data sets into two clusters 1 and 2. Now these two data sets can be treated as training data and the grouping of them is known. Given a new data, it can be classified to one of these two groups. Consider other data sets from the rest phases in the experiment, as the nominal concentration continuously varies the data sets in the phase which is close to phase 1 should be classified to cluster 1, similarly the data sets in the phase which is close to the phase 5 should be classified to cluster 2.

MATLAB function 'classify' classifies each row of the data in sample into one of the groups in training. Consider the data set from phase 2, the nominal concentration is 0.0125 mg/l, it close to the nominal concentration in phase 1 and is far from the concentration 0.25 mg/l in phase 5. The behaviors of these mussels in phase 2 should not have great variation compare to that in phase 1 since the nominal concentration just slightly rises. That is the observations of the multivariate in phase 2 should be similar to that collected in phase 2 and have distinct difference to that collected in phase 5 because of the massive changes in environment. Using the 'classify' function to classify the observations in phase 2, and we except that all or most of these observations are classify to the natural cluster of phase 1.

At first, two data sets from phase 1 and 5 are selected to be training data and observations of them are clustered into their natural groups. The type of discriminant function can be specified in the 'classify'.

- Linear: this type of discriminant function fits a multivariate normal density to each group, with a combined estimate of covariance, 22479 of 25920 observations are classified to the cluster of phase 1.
- Diaglinear: this type of discriminant function also fits a multivariate normal density to each group, but with a diagonal estimate of covariance, 22366 observations are classified to the cluster of phase 1.

- Quadratic: this function fits multivariate normal densities with covariance estimates stratified by group, 23528 observations are classified into the desired cluster.
- Diagquadratic: it function fits multivariate normal densities with diagonal covariance estimates stratified by group, 23462 observations are classified into the desired cluster.
- Mahalanobis: it uses Mahalanobis distances with stratified covariance estimates, 23396 observations are classified into the desired cluster.

Apply the analysis to the data sets from phase 4, sine it is a neighboring phase to phase 5, the observations in phase 4 are supposed to classified into the natural cluster of phase 5. 'Linear' discrimination function classifies 24165 of 25920 observations from phase 4 into the cluster of phase 5; 'Diaglinear' function correctly classifies 24170 observations; 23523 observations in phase 4 are grouped into cluster 2 by 'Quadratic' function; 23498 observations are identified by 'Diagquadratic' function; finally the default 'Mahalanobis' function classifies 23705 out of 25920 observations.

Compare the results of the discriminant analysis with data sets 2 and 4, it seems that 'Mahalanobis' discriminant function gives a balanced performance to these two data sets. Other discriminant functions also produce acceptable results. Continue the analysis in section 3.4, the clustering of the data sets from phase 1 and 5 is created by the k-means clustering analysis. Using the clustering result instead of the natural grouping and the data sets from phase 1 and 5 still are the training data.

22571 of 25920 observations in phase 2 are classified into the cluster that belongs to data from phase 1 and 15013 out of 25920 observations in phase 4 are classified to the cluster that belongs to the data from phase 5. The result of discriminant analysis to data set from phase 4 is unsatisfactory, since more than third of observations are classified to the cluster that belongs to data from phase 1 by mistake. Recall to the

Figure 3-20 and Figure 3-21, we can find that part of the observations from phase 5 are assigned to the cluster which belongs to the data from phase 1. In the result of the k-means clustering analysis, all the observations from phase 1 and part of observations from phase 5 are grouped into a new cluster and the rest of the observations from phase 5 forms to another cluster. This imperfect training data and its clustering lead to poor performance of the discriminant analysis.

4. Discussion and Conclusion

The data we collected are from the real-time monitoring sensors. The set of sensors is composed of biosensors, physical sensors and chemical sensors. The measurements of the physical environment consist of the current speed of the seawater, the direction of the seawater, the conductivity of the seawater, and the pressure in the seawater of the sensors station. These 5 monitoring data are obtained simultaneously. Chemical sensors monitor the turbidity of the seawater, content of oxygen in seawater and the content of chlorophyll in seawater which indicates if the food for the mussels is available. In general, these sensors form a measurement of the sea environment. However these three types of sensors measure the different aspects of the environment but performance differently. Physical sensors give accurate measurements but cannot reflect the condition of pollution directly; although the chemical sensors give the accurate measurements of the pollution in the sea but it takes time to get the result of the sensors. The heart rates of these mussels can be collected immediately, and the results of the statistical analysis of the data can be produced quickly and be used to reflect situation of the environment.

At first phase, the values of the heart rate of almost all the mussel are centralized, and some extremely lower and higher values of their heat rates are treated as outliers for each mussel. As the nominal concentration rise, especially the values of

the heart rates of the mussel 1 and 3 vary widely, so the observed data from their heart rates spread out in a wide range. Compare to mussel 1 and 3, other mussels seems not sensitive to the environmental deterioration, the values of their heart rates rate bit more diffused than before, but more outliers occur. This phenomenon apparently occurred in the mussel 5. The results of the principal component analysis can also represent the variation of the characteristics of their heat rates. Take account of the overall PCs for the experiment, the first and second principal components are always represented by the values of heart rates of the mussel 1 and 3 respectively. This is because that the heat rates of these two mussels have the largest variance during the experiment. As the condition of the environment in the experiment changes, more outliers of the heart rates of mussel 4 and 5 occur. This lead to the increment of the variance of these mussels' heart rates, so the third principal components is represented by the value of the heart rate of mussel 5 or 4 or the combination of them.

In the section 3.4 data sets from two phases are chose to test the clustering analysis. Two kinds of clustering analysis are considered, and the hierarchical clustering analysis is found to be inappropriate for large data set. However the k-means clustering analysis did not performance well, it grouped part of the data from phase 5 to the cluster of phase 1 by mistake. The poor performance is related to the algorithm and the feature of the data sets. As the figures show in section 3.2, the values of the first two PCs-mussels 1 and 3 spread out in a large range, and most parts of ranges are overlapped that is most observed values of them in these two phases are approximate. The algorithm of the analysis is to minimize the sum of the distances of the observations to the two centroids of two clusters; two approximate values from two different mussels may by group into one cluster by mistake. From this point, we can find that the selection of the mussels is very important. Look into the mussel 1 and 3, they seem very active during the experiment, but they act in an irregular way. The large variance of data observed from the mussel 1 and 3 helps these

two mussels to be principal component, but it did not reflect the variation of the environmental condition exactly. In the further study or practical deployment of these biosensors, the optimal choice is that the median value of the mussel varies as the environmental change; and the range of the observed values should not be significantly large because this would reduce the impact of the overlapping of data. Finally several this kind of mussels form to a multivariate will perform better.

Discriminant analysis is very useful to classify a new observed data. However its excellent results are based on good training data set. Poor or wrong clustering causes to group a new data into an undesired cluster. If we choose optimal mussels as multivariate, the data we collected in the experiment can be easily group by clustering analysis and the results can be used as training data directly for discriminant function.

5. Further Comments

The multivariate analysis consists of many methods for different purposes, and the data is numerous as time goes. Choosing sample data and analyzing them manually would lose some information and waste time. As in the paper, all the analyses are implemented in the MATLAB, how these functions cannot be applied to real-time monitoring. Since the PI system provides the collection, storage and access of the data, it facilitates we retrieve the data. More over PI system also provides an add-in tool ACE (Advanced Computing Engine) for computation, the computation is managed by the PI system and the result can be written back to the system. In the further work, we can turn to PI system, write some computations in VB, and deploy the computation to the data.

6. REFERENCES

[i] Table11.1 *Applied Multivariate Statistical Analysis* (2nd_ed) (Springer, 2007 Wolfgang Härdle and Léopold Simar)

[ii] Table11.2 *Applied Multivariate Statistical Analysis* (2nd_ed) (Springer, 2007 Wolfgang Härdle and Léopold Simar)

[iii] Page497 *Introduction to Data Mining* (Pang-Ning Tan, Michigan State University, Michael Steinbach, University of Minnesota Vipin Kumar, University of Minnesotav)

[iv] Page503 *Introduction to Data Mining* (Pang-Ning Tan, Michigan State University, Michael Steinbach, University of Minnesota Vipin Kumar, University of Minnesotav)

[v] Page508 *Introduction to Data Mining* (Pang-Ning Tan, Michigan State University, Michael Steinbach, University of Minnesota Vipin Kumar, University of Minnesotav)

[vi] Theorem 12.2 *Applied Multivariate Statistical Analysis* (2nd_ed) (Springer, 2007 Wolfgang Härdle and Léopold Simar)

[vii] http://www.mathworks.com/help/toolbox/stats/bq_679x-18.html