VOT 75069

ROBUSTNESS OF STATISTICAL GRAPHS FOR THE DETECTION AND DIAGNOSIS OF SOME LUNG DISEASE

(KETEGAPAN GRAF STATISTIK UNTUK PENGESANAN DAN DIAGNOSIS BEBERAPA PENYAKIT PARU-PARU)

NORLIZA MOHD NOOR

RESEARCH VOTE NO: 75069

DEPT. OF ELECTRICAL ENGINEERING
DIPLOMA PROGRAM STUDIES
UTM CITY CAMPUS
JALAN SEMARAK
54100 KUALA LUMPUR

2005

The project was carried with the collaboration from Dr. Hamidah Shaban, Specialist Medical Consultant, Institute of Respiratory Disease, Kuala Lumpur Hospital, Assoc. Prof. Dr. Omar Mohd. Rijal from The Institute of Mathematical Science, Faculty of Science, University Malaya and his Bachelor of Science (Statistics) student, Ms. Ong Ee Ling while undertake the SJES 3483: 2 Semester Statistics Project (8 credit). A copy of the SJES 3483 project thesis that was submitted to the Institute of Mathematical Science, Faculty of Science, University Malaya, is enclosed. A paper titled 'A Feature Detection Method For Pulmonary Mycobacterium Tuberculosis Using Conventional Chest Radiographs' was presented at the 12th European Signal Processing Conference in Vienna, Austria in September 2004. The paper was published in the conference proceedings.

A Feature Detection Method For Pulmonary Mycobacterium Tuberculosis Using Conventional Chest Radiographs

Omar Mohd. Rijal*, Norliza Mohd. Noor†
* Institute of Mathematical Science, University of Malaya
†Diploma Program Studies, Universiti Teknologi Malaysia

Published in
The Proceedings of the 12th European Signal Processing Conference, 2004.

A FEATURE DETECTION METHOD FOR PULMONARY MYCOBACTERIUM TUBERCULOSIS USING CONVENTIONAL CHEST RADIOGRAPHS

Norliza Mohd. Noor, Omar Mohd. Rijal**, & Chang Yun Fah****

** Dept. of Electrical Engineering, Universiti Teknologi Malaysia City Campus*
***Institute of Mathematical Science, Faculty of Science, University Malaya*
****Faculty of Engineering, Multimedia University*

ABSTRACT

The success of eliminating the disease Mycobacterium Tuberculosis (MTB) depends on the detection capabilities of medical organizations. In Malaysia, the government hospitals perform the major part of this particular task. An important ingredient of the diagnostic process in government hospital is the visual interpretation of standard chest X-ray films. A previous study proposed an objective alternative; involving wavelets coefficient, as the feature vector of MTB. In this study, we proposed an Andrew's Curve graphical presentation of the feature vector of MTB.

INTRODUCTION

Due to economic considerations the conventional x-ray film is still an important ingredient in the diagnostic process despite rapid advances in medical imaging technology (see for e.g. Middlemiss) [1] and Moores [2]. A description on the reliability of chest radiography is discussed in Toman [3].

The authors in [4] looked at subsets of the digitized chest x-ray image of a confirmed MTB patient. In particular, the infected region seen visually as white cloudy or snowflakes is the region of interest in this study. We then take some sample of grey-level values or pixel values the infected areas. The samples are in the form of vertical lines defined between a given pair of adjacent ribs, which in turn is defined as the line profile.

Thirty line profiles were obtained. For each line profile, applying one-dimensional discrete wavelet [5] gave the corresponding approximate and detailed Daubechies Coefficients. In total, a vector of 26 coefficients represented each line profile. Hierarchical clustering techniques [6] were applied using Minitab [7] and SPSS [8].

THE PILOT STUDY

In the pilot study [4] we studied two sets of data, confirmed tuberculosis patients and none-MTB patients. The chest radiographs of the confirmed MTB patients were provided by the Respiratory Unit, Kuala Lumpur Hospital and none-MTB patients were provided by Selayang Hospital.

The medical expert on MTB identified the infected area. For each infected area, we sampled 30 lines profile. For each line profile, we obtained 26 Daubechies coefficients. Six hierarchical clustering techniques were applied to the 30 x 26 approximate Daubechies coefficients. The general result of clustering showed that most techniques separate the line profiles into two groups: the first set nearest to the 'wind-pipe' is regarded as primary infected area, whilst the other set may be considered the secondary infected area.

Figure 1(a) shows a chest X-ray of the confirmed MTB patients. Figure 1(b) shows the line profiles selected. As an example of clustering, Figure 2a(i) is the dendogram using complete linkage and Figure 2a(ii) a schematic representation of line profiles being separated into two groups. The grouping of line profiles is summarized in Table 1. We define the average of the profile vectors, for example for the complete linkage;

$$\bar{\underline{x}} = \frac{\underline{x}_1 + \underline{x}_2 + \underline{x}_3 + \underline{x}_{11} + \underline{x}_{12} + \underline{x}_{22} + \underline{x}_{23} + \underline{x}_{24} + \underline{x}_{25}}{9} \quad (1)$$

where $\underline{x}_j$ is the vector of approximate coefficients for the j^{th} line profile.

From Table 1, we could see that clustering using the complete linkage, Between Group Average and Within Group Average gave the same grouping of line profiles. We denote the average for these vectors as X_A. Also, X_B, X_C and X_D represent the average vector for Ward, Centroid and Median method,

respectively. The Euclidean's distance between these vectors are shown in Table 2. Table 2 indicates the $\bar{x}$-vector for all clustering method is similar. Henceforth the $\bar{x}$-values, (for example complete linkage) may be used as a feature to identify MTB.

In summary, clustering of line profiles, or equivalently clustering of vectors of approximate Daubechies wavelet coefficients may be used as a method to identify regions that are infected with MTB. The average value of the profile vectors, for example, for complete linkage is as follows:

$\bar{x}$ = (2.456, -14.723, 51.464, 151.881, 149.618, 148.203, 146.751, 148.195, 145.076, 144. 298, 145.446, 144.347, 146.718, 145.889, 146.373, 149.406, 146.237, 145.709, 147.076, 147.818, 150.784, 150.092, 149.977, 143.548, 170.145, 24.881)

Therefore $\bar{x}$ may be used as a feature characteristic of the MTB disease.

ANDREW'S PLOT

Whilst the $\bar{x}$-vectors given in equation (1) may be used as a feature to identify MTB, there is a need to be able to compare the $\bar{x}$-vectors, for example:

(i) To compare two X-ray films of a patient undergoing treatment after one month.

(ii) Comparing a 'new' patient with a confirmed MTB patient.

The visual comparison of two 26-dimensional vectors is clearly not appealing. Andrews [9] proposed a method of plotting a data point $\underline{x}_r^T = (x_{r1}, \ldots, x_{rp}), r = 1, \ldots, n$, which involves plotting the curve $\{t, f_{\underline{x}_r}(t)\}$ where

$$f_{\underline{x}_r}(t) = x_{r1}/\sqrt{2} + x_{r2}\sin t + x_{r3}\cos t + x_{r4}\sin 2t + x_{r5}\cos 2t + \cdots \quad (2)$$

for each "data point' $\underline{x}_r (r = 1, \ldots, n)$ over the interval $-\pi < t < \pi$. Thus, each data point (in this case our $\bar{x}$-vectors) will appear as a harmonic curve drawn in 2 dimensions. It may be shown that $\int_{-\pi}^{\pi} [f_{\underline{x}}(t) - f_{\underline{y}}(t)]^2 dt$ between two curves $\{t, f_{\underline{x}}(t)\}$ and $\{t, f_{\underline{y}}(t)\}$ is proportional to the square of Euclidean distance between $\underline{x}$ and $\underline{y}$.

CLUSTERING AND ANDREW'S PLOT

Since the $\bar{x}$-vector are very similar, we propose studying $\underline{v} = (\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_6)/6$ where the subscripts 1,2, ...,6 represent the six clustering methods for a given patient. This is done in accordance with a standard approach in statistical clustering, namely accept the clusters suggested by the majority of clustering methods. In this study the Andrew's Curve of vector $\underline{v}$ were considered.

A random sample of ten patients were selected; six confirmed MTB patients from The Respiratory Unit, Kuala Lumpur General Hospital and four non-MTB patients from The Selayang Hospital.

The Andrew's Curve for each of the ten patients or data points $\underline{v}_r (r = 1,2,\ldots,10)$ were compared for values of t between zero and six. These values of t were chosen solely for producing graphs that may be recognized and differentiated with ease.

One confirmed MTB patient (black curve) and one non MTB patient (blue curve) were randomly selected and compared in Figure 3. Both curves show similar trend, except that the amplitude of the black curve is larger than the blue curve.

To show that the shape of the curves in Figure 3 is not a chance occurrence, Figure 4 compares the same blue curve with all the six confirmed MTB patients. Clearly, the black curves are clustered in a group, distinct from the blue curve. Further, in Figure 5, one confirmed MTB patient is compared with the other four non MTB patient showing similar result.

DEVELOPMENT OF AN MTB FEATURE DETECTION SYSTEM

Both Conventional chest X-ray and digital chest x-ray were used in this study. The x-ray films were digitized using film scanner and transfer to a PC-based system. The MTB detection software was develop in MATLAB 6.1 GUI (Graphical User Interface) environment [10]. The image of digitized chest X-ray is first

displayed on the screen. Then the user will be prompt to select the area that need to be analyzed. The selected area will then be display on the screen, the user then will need to provide some data for testing by drawing with the cursor several lines at the suspected infected area. The software will then generate line profiles and calculate the Daubechies approximate and detail coefficient. The Daubechies approximate coefficients were used as feature vector in this research. All the feature vector of the lines profile was then subjected to 6 clustering methods: complex linkage, centroid method, median method, Ward's method, between group average and within group average.

The Dendogram (see Figure 2) and a schematic representation of the clusters are displayed. Finally the Andrew's plot may also be displayed for selected values of t and $f_v(t)$.

SUMMARY AND FURTHER REMARKS

A pilot study of an objective method for feature detection for MTB is proposed whereby the usual problems associated with visual interpretations of images are removed. Apart from detection, the system allows the medical practitioner to 'explore' the image and perform segmentation.

However, the robustness and the sensitivity of the method still need to be studied in the sense that a larger database (more patients) should be obtained and compared.

ACKNOWLEDGEMENT

We would like to acknowledge cooperation given by:
(i) Dr. I. Kuppusamy and Dr. Azwayati Abas from The Respiratory Unit, Kuala Lumpur General Hospital,and

(ii) Dr. Rosnah Hadis and Dr. Zaharah Musa from The Selayang Hospital.

This research was funded under short-term research grant from Research Management Centre, UTM.

REFERENCE

[1] Middlemiss, H, "Radiology of the future in developing countries", British Journal of Radiology, 55, pp. 698-699, 1982.

[2] Moores, B.M., Digital X-ray Imaging, IEE Proceedings, Vol. 134, part A, Number 2, Special Issues On Medical Imaging, 1987.

[3] Toman, K., Tuberculosis: Case Finding and Chemotherapy, W.H.O. Report, pp. 28, 1979.

[4] Noor, N. M., Rijal, O.M and Chang, Y.F., "Wavelet as features for Tuberculosis (MTB) using standard x-ray film images", IEEE proceedings of 6th International Conference on Signal Processing (ICSP02), Beijing, China, 2002.

[5] Daubechies, I., Ten Lectures on Wavelets Society For Industrial and Applied Mathematics, Philadelphia, 1992.

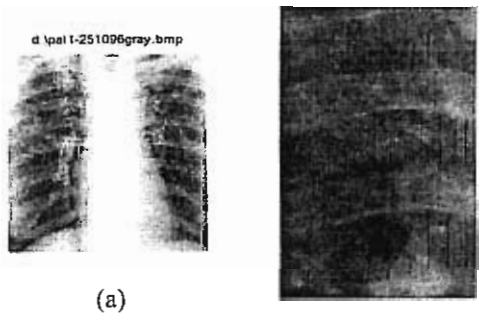
[6] Everitt, B. S., Cluster Analysis, Heinemann Educational Books Ltd, London, 1977.

[7] MINITAB Software for Windows, The MINITAB Inc.

[8] SPSS Software for Windows, The SPSS Inc.

[9] Andrews, D.F., "Plots of high dimensional data", Biometrics, 28, pp.125-36, 1972.

[10] Matlab software, The Language of Technical Computing, The Mathworks Inc.



(a) (b)

Figure 1: (a) A chest X-ray of a confirmed MTB patient; (b) A subset of (a) showing the line profiles taken between the area of 2nd, 3rd, 4th and 5th ribs.

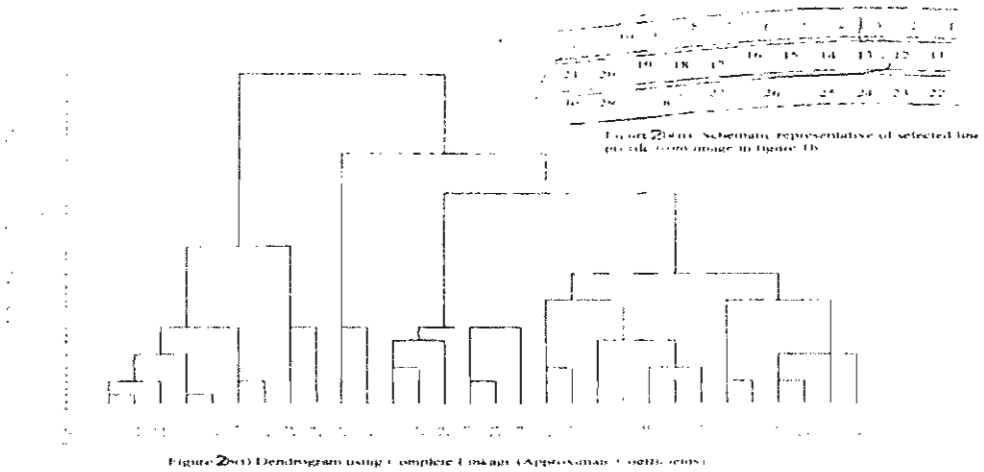


Figure 2

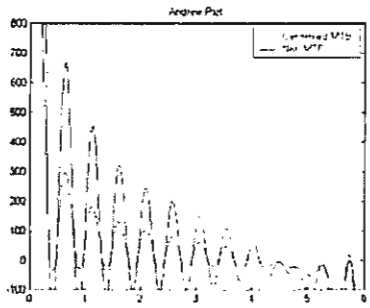


Figure 3

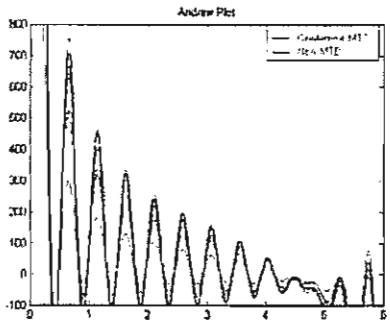


Figure 4

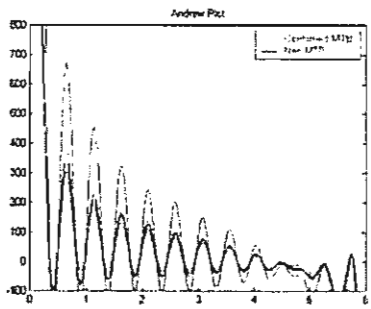


Figure 5

Table 1: Summarized grouping of line profiles using six hierarchical clustering method.

Clustering Method	Grouping of line profile
Complete Linkage	1, 2, 3, 11, 12, 22, 23, 24, 25
Centroid Method	1, 2, 3, 11, 12, 22, 23
Median Method	1, 2, 3, 4, 5, 6, 7, 11, 12, 13, 14, 15, 16, 22, 23
Ward's Method	1, 2, 3, 4, 11, 12, 22, 23, 24, 25
Between Group Average	1, 2, 3, 11, 12, 22, 23, 24, 25
Within Group Average	1, 2, 3, 11, 12, 22, 23, 24, 25

Table 2: Euclidean's Distance Matrix for X_A, X_B, X_C and X_D.

	X_A	X_B	X_C	X_D
X_A	0	0.000484730	0.005712382	0.020888985
X_B	0.000484730	0	0.003082583	0.015743008
X_C	0.005712382	0.003082583	0	0.005572539
X_D	0.020888985	0.015743008	0.005572539	0

Project Dissertation

GRAPHICAL METHODS FOR ANALYSING X-RAY IMAGES.

ONG EE LING
SES020672

SUPERVISOR
PROF. MADYA DR OMAR BIN MOHD RIJAL

A DISSERTATION SUBMITTED TO
THE INSTITUTE OF MATHEMATICAL SCIENCE, FACULTY OF SCIENCE, UNIVERSITY OF MALAYA, KUALA LUMPUR FOR THE COURSE SJES3483.
FEBRUARY 2005

Abstract

The x-ray film remains as an essential ingredient in the diagnostic process. The problem of interpretation of x-ray images require objective method to be developed. The image is 'broken down' into a set of line profiles and compressed using wavelets. The vectors of wavelet coefficient are then studied graphically. In particular, the Andrews curve and its properties will be investigated on x-ray images and in a simulation study and with the assistant of MATLAB

Acknowledgements

Firstly, I would like to express my utmost sincere thanks to my supervisor Dr Omar bin Mohd Rijal for his quality supervision, valuable guidance, support, insightful ideas, utmost patience and time during the term of my project.

As this project is funded under the fundamental research grant, Research Management Center in UTM, I would like to thank Pn Norliza Mohd Noor, whose experience in image processing for medical application had help me carry out this study and also her contribution in providing me a first course in MATLAB programming.

Thanks are also due to Dr Hamidah binti Shaban, a medical consultant for respiratory disease from Institute of Respiratory Disease, Kuala Lumpur Hospital for providing me a through understanding regarding disease detection and helps in visual interpretation of chest radiograph regarding MTB and lung cancer disease.

I would also like to express my deepest gratitude to my parents for without their strong support and understanding, this project may not be accomplished. Other than that, thanks are also due to Mr. Chang Yun Fah for providing helps from time to time.

Last but not least, I would also like to thank ISM staff and my fellow ISM friends who have given me advise, support and help throughout the term.

CONTENT

CHAPTER 1: INTRODUCTION

1.1 Introduction to Multivariate Data

In many area of research, the data to be analyzed and interpreted are essentially multivariate which are in the form of vectors of random variables. Typically, these vectors arise from taking measurement or observation on several different variables on a number of objects or persons. We denote the number of variables by p, and the number of objects or person by n. A typical multivariate data matrix will have the form

$$\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \qquad [1.1]$$

where the x_{jk} indicate the particular value of the k^{th} variable that is observed on the j^{th} individual. The data matrix can also be seen as n rows vectors, which we denote by $\mathbf{x}^T_1, \ldots, \mathbf{x}^T_n$, or as p column vectors, which we denote by $\mathbf{y}_1, \ldots, \mathbf{y}_p$. Thus

$$\mathbf{X}=\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}=\begin{bmatrix} y_1 & y_2 & \cdots & y_p \end{bmatrix} \qquad [1.2]$$

where $\mathbf{x}^T_1$ denote the transpose of $\mathbf{x}_1$. Sometimes $\mathbf{x}_j$ itself is referred to as an observation. The vectors may or may not come from the same probability distribution. Generally the variable are correlated and it may be quantitative (discrete or continuous) or qualitative (ordered or unordered categories). See Seber (1984).

Various statistical techniques both descriptive and analytical are designed to solve (real life) problems arise based on the multivariate data sought. Much of the information contained in the data can be assessed by calculating certain summary number known as descriptive statistics. See Johnson (1998).

The following are some basic statistics:

a) Sample mean. The sample mean of the *k*th variable is

$$\overline{X}_k = \frac{1}{n}\sum_{j=1}^{n} X_{jk} \qquad k = 1, 2\ldots, p \qquad [1.3]$$

and the *sample mean vector* (mean vector) is represented as

$$\bar{x} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_p \end{bmatrix}' \qquad [1.4]$$

b) Sample variance. The sample variance of the kth variable is

$$s_k^2 = s_{kk} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2 \qquad k = 1, 2, \ldots, p \qquad [1.5]$$

c) Sample covariance. A measure of the *linear association* between the variables i and k is given by the sample covariance

$$s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \qquad \text{for i, k} = 1, 2, \ldots, p \qquad [1.6]$$

The p x p matrix with elements given by [1.5] and [1.6] is called the *sample covariance matrix,* or simply the "covariance matrix". These covariance matrices are symmetric by nature.

$$\sum = S_n = \begin{bmatrix} s_{11} & s_{12} & . & . & s_{1p} \\ s_{21} & s_{22} & . & . & s_{2p} \\ . & . & . & & . \\ . & . & & . & . \\ s_{p1} & s_{p2} & . & . & s_{pp} \end{bmatrix} \qquad [1.7]$$

d) Sample correlation coefficient. The sample correlation coefficient of variables i and k is given by

$$r_{ik} = \frac{Sik}{\sqrt{Sii \times Skk}} \qquad \text{for i, k= 1,2,\ldots p.} \qquad [1.8]$$

See Bhattacharyya [1977]. This measure of the linear association between two variables does not depend on the units of measurement. These coefficients can be thought of as normalized sample covariance which lies between -1 and +1. Sample correlation matrix is also a form of symmetric matrix.

1.2 Multivariate Methods

Over the recent years, applications of multivariate methods have increased tremendously. The choices of the most appropriate methods depend on the type of data, the type of problem and the objective which are envisaged for the analysis. In this study, two important methods involving multivariate data are to be considered namely discrimination analysis and cluster analysis. In this section, we will also touch on the issue of data reduction or structural simplification and outliers detection.

1.2.1 Discrimination and Classification

Given an unknown observation $\mathbf{x}^*$ and given two population, say

$$\pi_1 \ : \ \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$$

and $$\pi_2 \ : \ \mathbf{x}_{n+1}, \mathbf{x}_{n+2}, \ldots, \mathbf{x}_{n+m}$$

we define discrimination as the problem of deciding whether $\mathbf{x}^*$ belong to π_1 or π_2. An assignment rule can be estimated from the sampled data and used to assign current observation (object). We would like our assignment rule to be optimal in some sense such as minimizing the number or cost of any error of misclassification that we might make on the average and also to consider *prior* probabilities. This is because the groups may overlap. See Johnson [1998] and Chatfield [1980]. These situations may be represented pictorially as in Figure1.1.

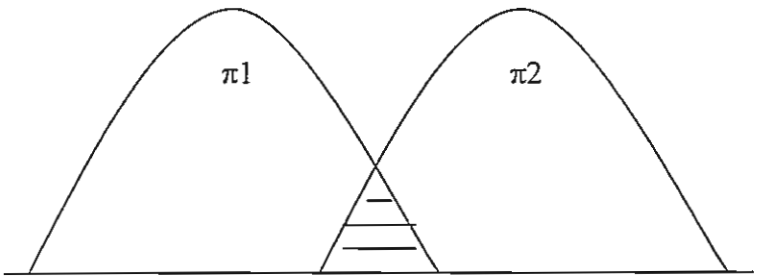


Figure 1.1: Distribution from two populations. The shaded area represent the conditional probabilities of misclassifying individual from population j to population i., P[i|j].

In the case of normal distribution, for the ith population, $\mathbf{X} \sim N_p(\mu_i, \sum)$, i=1,2

$$f_i(\mathbf{x}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\} \qquad [1.9]$$

we can derive a simplified version of allocation rule based on the following result:

Result 1.1:

Let the population π_1 and π_2 be described by multivariate normal densities of the form [1.9]. Then the allocation rule that minimizes the expected cost of misclassification is as follows:

Allocate $\boldsymbol{x}^*$ to π_1 if

$$(\mu_1 - \mu_2)^T \Sigma^{-1} x^* - \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 + \mu_2) \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right] \qquad [1.10]$$

where p_i is the prior probability of π_i i=1,2 and c(i|j), $i \neq j$ is the cost of misclassification. Allocate $\boldsymbol{x}^*$ to π_2 otherwise.

The cost are zero for correct classification, c(1|2) when an observation from π_2 is incorrectly classified as π_1 and c(2|1) when an observation from π_1 is incorrectly classified as π_2. Let p_1 be the *prior* probability of π_1 and p_2 the *prior* probability of π_2, p_1 *and* p_2 should be taken in a way so that $p_1 + p_2 = 1$. Further information can be obtained from Johnson [1998].

In cases when the probability density function, $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, cost of misclassification and prior probability are not available, we are unable to derive its allocation rule. So, alternative allocation rules or method should be used instead. However, the basic ideas of discrimination still remain the same.

1.2.2 Clustering.

Clustering (see Anderberg [1973]; Gnanadesikan [1977]) is a general scientific process of searching for pattern in data and then construct laws that explain the pattern. Given observation $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$. We define clustering as the problem of establishing groups or subsets of the observations and allocate a set of individuals to a set of mutually exclusive, exhaustive groups such that individuals within a group are similar to one another while individual in different group are dissimilar. For example, we may try to select two groups of students (say, mathematically gifted or not mathematically interested) from a given school based on their mathematic marks.

In order to carry out cluster analysis, we need to measure the similarity (or dissimilarity (distances)) of every pair of individual from data matrix using appropriate way. Even though a number of measures or *metrics* have been defined, Euclidean distance will be used here as it is one of the most common measures of distance, doesn't involve any computation of covariance and can be defined for any value of variables. See Mardia[(1979] and Johnson [1998]. Euclidean (straight line) distance between two p-dimensional observations is given as the following:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1-y_1)^2+\ldots+(x_p-y_p)^2} = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})} \qquad [1.11]$$

Once distances have been determined, clustering proceeds by applying a particular algorithm to these values. Agglomerative methods begin initially with as many clusters as the numbers of objects. The most similar objects are first grouped, and these initial groups are merged according to their similarities. Eventually all subgroups are fused into a single cluster. A more detail explanation regarding the procedures of clustering can be obtained from Seber [1984] and Johnson [1998]. The following are the hierarchical clustering methods used in this study.

a) **Single linkage (SL).** If C_1 and C_2 are two clusters, then the distance between them is defined to be the smallest dissimilarity between a member of C_1 and a member of C_2 namely,

$$d_{(c1)(c2)} = \min[d_{rs} : r \in c_1, s \in c_2] \qquad [1.12]$$

where r denotes "object r".

b) **Complete linkage (CL).** This methods is defined in terms of the largest dissimilarity between a member of C_1 and a member of C_2 namely,

$$d_{(c1)(c2)} = \max[d_{rs} : r \in c_1, s \in c_2] \qquad [1.13]$$

c) **Average linkage (AL).** The distance between C_1 and C_2 is defined to be the average of the n_1n_2 dissimilarities between all pairs namely,

$$d_{(c1)(c2)} = \frac{1}{n_1 n_2} \sum_{r \in c1} \sum_{s \in c2} d_{rs} \qquad [1.14]$$

d) **Centroid method (CM).** The distance between two clusters is defined to be the distance between the cluster centroids. If

$$\bar{x}_i = \sum_{i \in c} \frac{x_i}{n_i} \qquad [1.15]$$

is the centroid of n_1 members of C_1 and $\mathbf{x_2}$ is similarly defined for C_2, then

$$d_{(c1)(c2)} = P(\bar{x}_1, \bar{x}_2) \qquad [1.16]$$

where P is a proximity measure such as squared Euclidean distance or other dissimilarity.

e) **Ward Method (WM).** This method uses the incremental sum of squares; that is, the increase in the total within-group sum of squares as a result of joining groups C_1 and C_2. It is given by

$$d_{(c1)(c2)} = n_{c1} n_{c2} d^2_{c1c2} / (n_{c1} + n_{c2}) \qquad [1.17]$$

where d^2_{c1c2} is the distance between cluster C_1 and C_2 defined in the Centroid method.

According to Chatfield [1980], because some methods work well on certain types of data and not on others, it is sometimes suggested that several different clustering method are implemented to see if the results of grouping are roughly consistent. The results usually displayed graphically as a ***dendrogram*** (***tree diagram***) and will reveal the same grouping only when groups are spherically shaped and well separated.

1.2.3 Dimensionality and Outliers

In section 1.1, we pointed out that the data matrix can be regarded as n row vectors objects or as p column vectors of variables. Often in multivariate data, many variables are included when taking measurements on people or objects. This is done to avoid overlooking any variables that may have future relevance. Unfortunately, when the dimension p is large, a data set may not only be very costly to obtain but it may also be unmanageable, difficult to study and will be misleading when display graphically. See Seber [1984]. In case like this dimension reduction techniques (be it a multivariate technique or graphical methods) is necessary to reduce the dimensionality of p with the aim to exclude any unimportant or irrelevant variables in interpreting and summarizing data.

Besides dimensionality, another immediate concern is to determine whether an unknown observation **x** is different from a clusters of observations. This is a problem of determining an 'outlier'. In other word, the value in a data set establishes a norm and any value that are quite deviant are labeled as outliers. Outlier exerts a much stronger influence on summary statistics, confidence interval or test results. Outliers can be detected visually in a lower dimension diagram (e.g. univariate scatter plot) by looking for observations that are far from the others. However when p is large, the number of scatter plots p(p-1)/2 may prevent viewing them all. In case involving higher dimension, a large value of

$$\left(\frac{n-1}{n}\right)\left(x_j-\bar{x}_{-j}\right)^T S_{-j}^{-1}\left(x_j-\bar{x}_{-j}\right)\sim\frac{p}{n-p-1}F_{p,n-p-1} \qquad [1.18]$$

(compared to the critical value of F(p,n-p-1:1-α)) might suggest an unusual observation, even though it can not be detected visually. Once an outlier is identified, we have to decide whether to remove the erroneous outliers or to remain it as natural abnormalities.

1.3 Multivariate Graphical Representation

If appropriate assumption, e.g. the multivariate data is normally distributed, methods in section 1.1 can be used without any doubt. However in practice, the probability distribution of the multivariate vectors is unknown causing the investigation to depend solely on graphical methods. In this study, we will transform the numerical vector x to a graphical form. In other words, given say n person x_1, x_2,... x_n, after transformation we will have n graphs i.e. each individual will be represented by a graph. Our task thus is to investigate the suitability of the graph as a mean for solving the problem of dimension reduction, discrimination, clustering and outlier detection.

1.4 Brief Literature Review

The great success of graphical methodology is based on their simplicity and transparency. Graphical representation enables data to be explored thoroughly, to look for patterns and relationships, to confirm or disapprove the expected and to discover new phenomena. In multivariate, high dimensional data can be viewed as slices of various two-dimensional and three-dimensional perspectives using new graphical technique software. Examples of the choices of two and three dimensional graphical representation are the scatterplot, 3D scatter plot, boxes (Hartigan [1975]), stars (Welsch [1976]), k-sided polygons (siegel et. Al[1971]), glyphs and metroglyphs (Anderson [1960]), profiles (Bertin 1976]) Chernoff faces (Chernoff [1973]), profiles and Andrews curve (Andrews [1972]), weather wanes (Bruntz et al.[1974]) and constellations (Wakimoto and Taguri [1978]). These graphical representation may somehow enable us to carry out the multivariate describe earlier in their own way.

However, when many variables are involved, some problem may be encountered:

a) Data examination is likely to lead to confusion if the number of variables is greater than about ten as the number of plots to be examined becomes larger. (see Everitt [1978]).
b) Plots will be misleading since any structure present in the original p-dimensional space of the data is not necessarily reflected by that present in plot of pairs of variables. Examples of data illustrating this fact are given by Cattell and Coulter [1966], and by Nathenson [1971].
c) Some graphical methods are easily affected when variables are interchanged. See Chernoff and Rizvi [1975] and Fienberg [1979] for related experiments.

A number of studies have been conducted to determine which of the graphical representation are best able to bring to the fore differences between the observation (see, for example, Wang [1978] pp.123-141). It appears that Andrews' curve normally fare the best, Chernoff faces second best and profile the worst, with not much to choose from between stars, glyphs and boxes.

1.5 Introduction to Andrews Curve

Andrews [1972] have suggested a graphical method for displaying set of p-dimensional observations $\mathbf{x}^T = [x_1, x_2, \ldots\ldots\ x_p\ ]$ without a substantial loss of information. Each of the $\mathbf{x}^T$, defines the following function

$$f_x(t) = x_1/\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \ldots\ldots\ldots \quad [1.19]$$

This plot are then plotted the range $-\pi < t < \pi$. In a simplier form, $f_x(t)$ can also be defined as following

$$f_x(t) = \mathbf{x}^T\mathbf{a} \quad [1.20]$$

where $\mathbf{a}^T = \mathbf{a}(t)$ is chosen based on the fact that

$$\int_{-\pi}^{\pi} a_j(t)a_i(t)dt \begin{cases} 1 & if \quad i = j \\ 0 & if \quad i \neq j \end{cases} \quad [1.21]$$

especially when $\mathbf{a}(t)$ is chosen as $[\ 1/\sqrt{2}, \sin t, \cos t, \sin 2t, \cos 2t, \ldots]$.

When drawn across the plot, a p-dimensional observation will appear as an individual sinusoidal curve.

1.6 Properties of Andrews Curve.

a) This function preserves *Euclidean distances.*

The Euclidean distance d between two observation $\mathbf{x}^T = (x_1,\ldots,x_p)$ and $\mathbf{y}^T = (y_1,\ldots,y_p)$ is directly proportional to the Euclidean distance D between the two corresponding functions [1.19]. Euclidean distance between $\mathbf{x}^T$ and $\mathbf{y}^T$ is defined as

$$d^2 = \sum_{i=1}^{p} (x_i - y_i)^2 \qquad [1.22]$$

D^2 on the other hand composed of the sum of squares of all possible differences $(x_i - y_i)$ and make the usual association between summation for discrete variables and integration for continuous ones. Thus at the value t_0 the difference between the two function is $f\mathbf{x}(t_0) - f\mathbf{y}(t_0)$ and t_0 can take any value between $-\pi$ and π.

$$\begin{aligned}
D^2 &= \int_{-\pi}^{\pi} [f_x(t) - f_y(t)]^2 \, dt \qquad [1.23]\\
&= \int_{-\pi}^{\pi} [\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\mathbf{y}]^2 \, dt\\
&= \int_{-\pi}^{\pi} [\mathbf{a}^T(\mathbf{x} - \mathbf{y})]^2 \, dt\\
&= \int_{-\pi}^{\pi} [(\mathbf{a}^T\mathbf{v})(\mathbf{a}^T\mathbf{v})] \, dt \text{ where } \mathbf{v}=(\mathbf{x}-\mathbf{y})\\
&= \int_{-\pi}^{\pi} [(a_1v_1 + \ldots + a_kv_k)(a_1v_1 + \ldots + a_kv_k) \, dt\\
&= \int_{-\pi}^{\pi} a_1^2v_1^2 \, dt + \int_{-\pi}^{\pi} a_2^2v_2^2 \, dt + \ldots + \int_{-\pi}^{\pi} a_2^2v_2^2 \, dt +\\
&\quad \int_{-\pi}^{\pi} \sum\sum a_ia_j \, v_iv_j \, dt
\end{aligned}$$

based on orthogonal property of a defined in [1.21]

$$\begin{aligned}
&= 1v_1^2 + 1v_2^2 + \ldots + 1v_k^2 + 0\\
&= v_1^2 + v_2^2 + \ldots + v_k^2\\
&= \mathbf{v}^T\mathbf{v}\\
&= (\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j) \qquad [1.24]
\end{aligned}$$

b) **This function preserves *means.***

This implies that the function of the mean vector of the observed vectors $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n$ is the point wise mean of the function $f_{x1}(t), f_{x2}(t), \ldots, f_{xn}(t)$. We have

$$f_{\bar{x}}(t) = \frac{1}{n}\sum_{i=1}^{p} f_{xi}(t) \qquad [1.25]$$

so that the curve representing the mean looks like an "average" curve.

c) This function preserves the *variances.*

If the variables in the data matrix are uncorrelated with common variance σ^2 then the function value at t, $f_x(t)$, has variance

$$\text{Var}[f_x(t)] = \sigma^2(\tfrac{1}{2} + \sin^2 t + \cos^2 t + \sin^2 2t + \cos^2 2t + \ldots). \qquad [1.26]$$

If p is odd, this variance reduces to the constant $\frac{1}{2}\sigma^2 p$, while if p is even the variance lies between $\frac{1}{2}\sigma^2(p-1)$ and $1/2\,\sigma^2(p+1)$. In either case the relative dependence of σ^2_f on t is either very slight or non existent, so that the variability of the plotted function is almost constant across the graph.

d) The representation yields *one-dimensional projections.*

For a particular value of $t = t_0$, the function value $f_x(t_0)$ is proportional to length of the projection of the vector **x** on the vector

$$f_1(t_0) = \{1/\sqrt{2}, \sin t_0, \cos t_0, \sin 2t_0, \cos 2t_0, \ldots\} \qquad [1.27]$$

since $f_x(t_0) = \{\mathbf{x}^T f_1(t_0)/[f_1^T(t_0) f_1(t_0)]\}.[f_1^T(t_0) f_1(t_0)]$. This projection onto a one-dimensional space may show up clustering or any data peculiarities that occur in this subspace and which may be otherwise obscured by other dimensions.

e) This function preserves *linear relationships.*

If **y** lies on the line joining **x** and **z**, then $f_y(t)$ lies between $f_x(t)$ and $f_z(t)$ for all t.

1.7 Discussion on Andrews Curves

The above mention properties give rise to the fact that Andrews' curves can be useful in clustering observation points in homogeneous groups or to compare individual function with the mean function. See Andrew [1971]. For all values of t, if a set of Andrews curves remain close enough to form a band, then their corresponding points are located close together in the Euclidean metric. This band will represent a cluster of data points. Andrews curve is also capable of representing a p-dimensional vector in a two dimensional graphical display without any lost of information. Outliers may be identified visually from the plot of the functions. Any curves that stand out as quite different from any group will be considered as outliers.

In addition, Andrews [1971] also suggested significance tests for testing whether an individual observation vector **Xi** differs significantly from the hypothetical population mean μ_0 under the assumption that the p are independent normal variables and construct a confident interval for μ_0.

$$z = \frac{[f_x(t) - f_u(t)]}{\{\text{var}[f_x(t)]\}^{1/2}} \qquad [1.28]$$

However, note that this test is exact if the value of t is chosen *a priori*.

CHAPTER 2: DIGITAL IMAGE PROCESSING

2.1 Introduction to Digital Image

According to Gonzales [1992], the term digital image refers to a two-dimensional light intensity function $f(x, y)$, where x and y denotes spatial coordinates and the value of f at any point (x, y) is proportional to the brightness (or gray level) of the image at that point that has been discretized both in spatial coordinates and brightness. A digital image can be considered a matrix whose row and column indices identify a point in the image and the corresponding matrix element value is called a picture element or 'pixel'. Each pixel consists of a number range from zero to 4096 for a 12 bit image. These numbers are the gray level intensity value at that point where zero represents brightness and 4096 represent darkness.

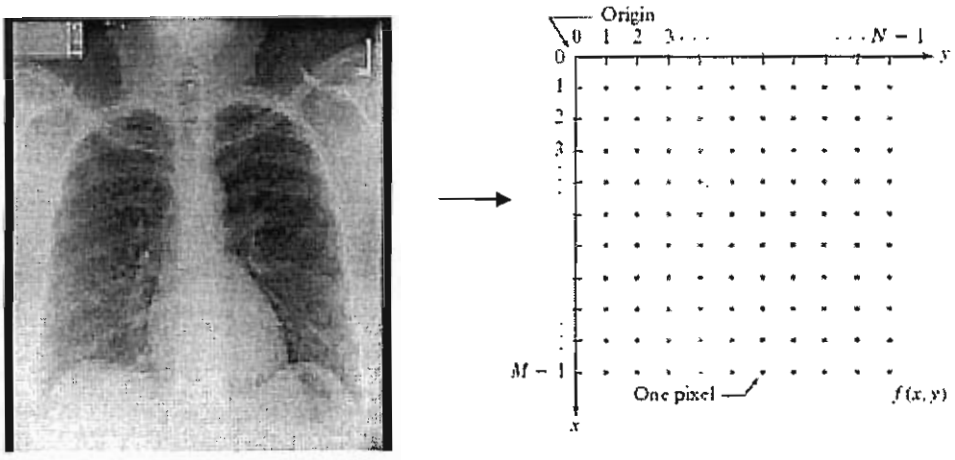


Figure 2.1: A diagram showing a digital x-ray image.

$$f(x,y)=\begin{bmatrix} f(0,0) & f(0,1) & \dots & f(0,M-1) \\ f(1,0) & f(1,1) & \dots & f(1,M-1) \\ \vdots & \vdots & \dots & \vdots \\ f(N-1,0) & f(N-1,1) & \dots & f(N-1,M-1) \end{bmatrix} \qquad [2.1]$$

Figure 2.2: Two-dimensional light intensity function of a digital image.

2.2 Gray Level Histogram.

The number of pixels in an image that have a particular gray level can be shown using the gray-level histogram. The gray-level histogram is one of the simplest and most useful tools in imaging as it summarizes the gray level content of an image. The y axis represents the gray level and the x axis represents the frequency of occurrence (number of pixels). However, at times, the characters of the original image don't seem to stand out as a desired peak clearly, indicating potential difficulties in interpreting the histogram. In order to get a better visualization, transformation is needed.

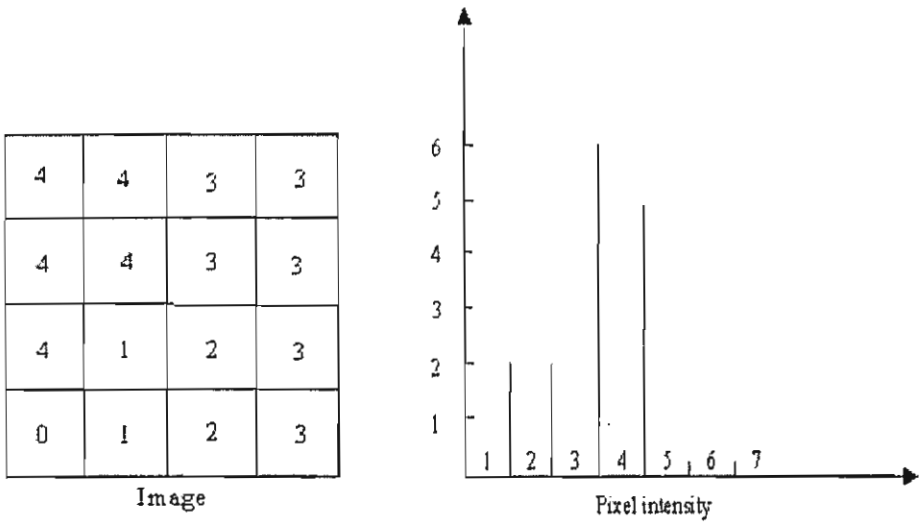


Figure 2.3: Gray level histogram

2.3 Fundamental of Digital image processing.

Digital image processing is a series of process which aim to produces a modified version of an image. The organization of the process is summarized in Figure 2.4 to provide a brief overview.

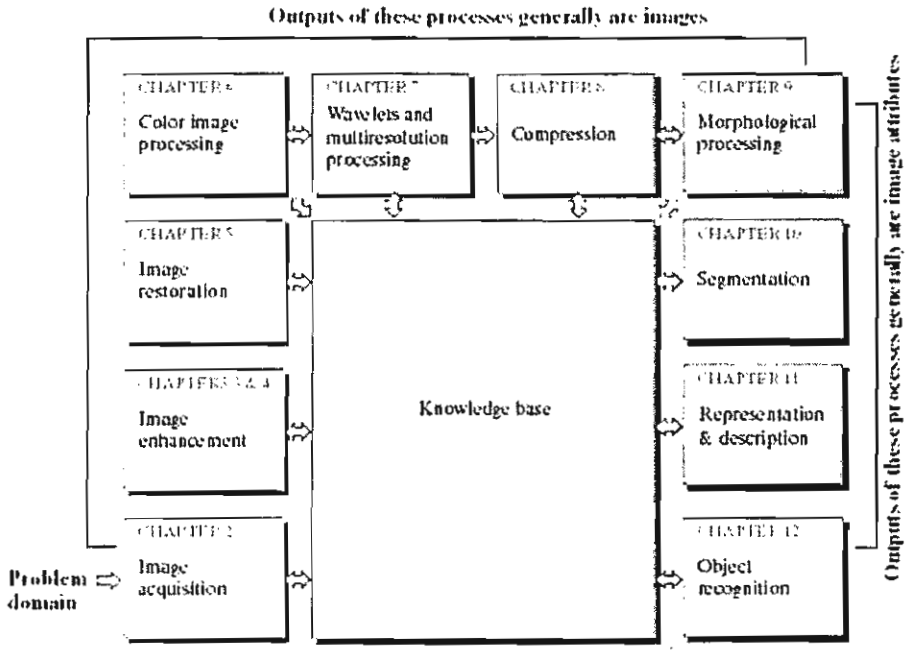


Figure 2.4: A summary of the organization of the general digital image processing. (Adapted from Gonzalez [1992]).

Image acquisition could be as simple as being given an image that is already in digital form. Generally, an imaging sensor and the capability to digitize the signal produced by the sensor are required in this process. Although many steps are involved in the organization of digital image processing, however, at times, not all steps are used. Two important aspects to be considered in this study are described briefly as the following section.

2.3.1 Enhancement and Filtering

When a picture is converted from one form to another, the quality of the output picture may be lower than that of the input. Therefore a method called enhancement is needed to enhance low quality pictures by extracting only the important information from the image or by increasing the visibility of one portion, aspect or component of an image though generally at the expense of others whose visibility is diminished (contrast enhancing). However, the number of pixels is not reduced. On the other hand, filtering is used to remove unwanted parts of image by removing periodic noise introduced by devices that stored and transmit images. It functions by removing selected frequencies and orientations by reducing the magnitude values for those terms. See Russ [1994]. In this study, the x-rays we obtained are already the enhanced version and the enhancement process has been carried out by a radiographer before hand.

2.3.2 Segmentation

Here, we consider an approach with the aim to extract hidden information in a picture implemented by digital image processing techniques. In segmentation, the output (raw data pixels) constitutes either boundry of a region or all the points in the region itself. This output is then converted to a form suitable for computer processing depending on the main focus of image analysis. In this study we will be using regional representation since our focus is on the internal properties, such as texture. It is a method for describing the data so that features of interest are highlighted. In some cases, segmentation may also be approached as a rather special clustering problem in which points in n-dimensional space with similar properties are grouped together.

2.4 Brief Literature Review.

Now that high density primary and secondary memory technology and powerful computers are a reality, the range of applications of digital imaging processing is growing extremely rapid. Many papers of digital image processing appear in many different journals. Successful applications to date include industrial application, space exploration, medical diagnostic, scientific analysis, remote sensing, telecommunication and etc. However, in all branches, the methods used in the pursuit of knowledge are very similar.

In the field of medical, medical radiology has developed imaging techniques to observe the inside of human body. These techniques include magnetic resonance imaging (MRI) and computer tomography (CT). These techniques provide detailed images of living tissues and are used for detecting tissue deformities such as cancer and injuries. By converting an image into digital form, it is possible to remove noise element from x-ray images, enhance their contrast and remove the blurring effects and perform segmentation. This form of representation makes it easier for physician to accurately measure the extent of tumors and other significant features. See for example Gao [2002], Tweed [2002] and Hiranos [2002] for the usage of segmentation in medical images and Cheng [2003]] for the example of usage of digital imaging in diagnosis a disease. Awcock [1995] has noted that the major motivating factor in the field of medical imaging is to eliminate the necessity for invasive surgery or treatment as far as possible, since this always involves some trauma to the patient as well as an inevitable element of risk.

CHAPTER 3: DETECTION OF MYCOBACTERIUM TUBERCULOSIS

3.1 Motivation.

Despite the existence of advanced technology such as ultrasound and MRI, detection of well known lung diseases like tuberculosis and lung cancer still depends on the use of standard x-ray films. This is a direct consequence of cost considerations. The main problem with interpreting x-ray film is the quality of image which in turn leads to difficulties in visual interpretation.

3.2 Primary Detection of Mycobacterium Tuberculosis (MTB).

On the x-ray film, detection of MTB is visually done by looking for white spots or "snow flakes". Serious cases of MTB will be indicated by existence of "cavities". This approach of detection clearly involves considerable "subjective visual interpretation". Our approach to achieve objectivity is done by using Digital Technology i.e. digital x-ray images. MATLAB will be used to analyze the "enhanced" digital images to identify the abnormal tissue in the lung.

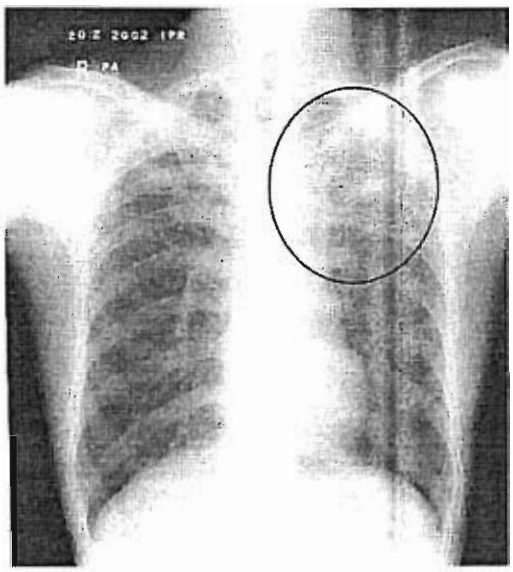


Figure 3.1: An x-ray film of a confirmed MTB patient. Red circle indicate MTB infected region.

3.3 Introduction to Line profile

A given subset of the image may further be seen as a collection of straight lines. Each line is defined as a line profile. A typical line profile may be seen as a "signal" (or wave), and a collection of line profiles may be useful to represent a given subset of the image. Here, a line profile can be regarded as a form of feature extraction. In the study presented here, we focus on obtaining just a simple rectangular framing to delimitate left and right lung field independently. 30 line profiles are then obtained between the ribs of the lung from the x-ray image of a confirmed MTB patient by a MTB medical expert as shown in the figure 3.2. This is done with the assistance of MATLAB.

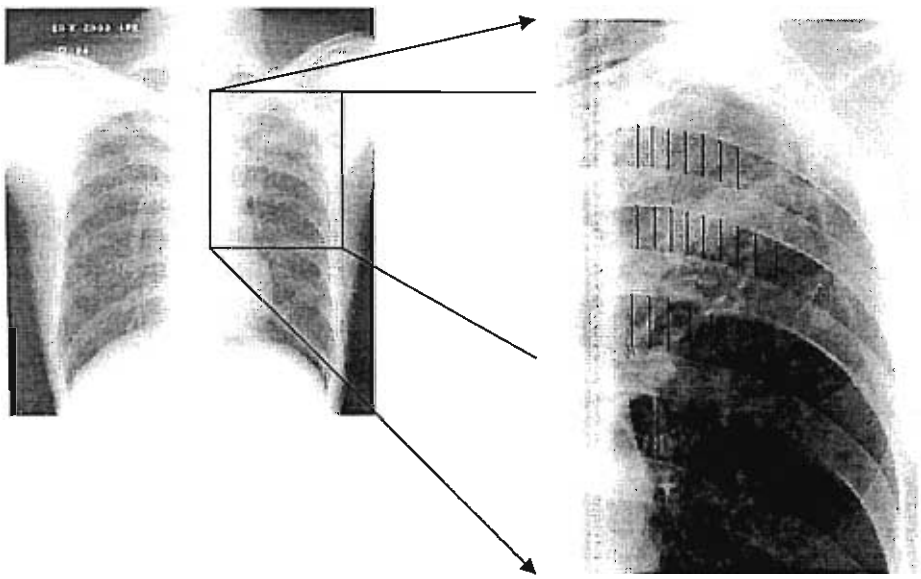


Figure 3.2(a) Line profiles taken on a lung x-ray of patient A.

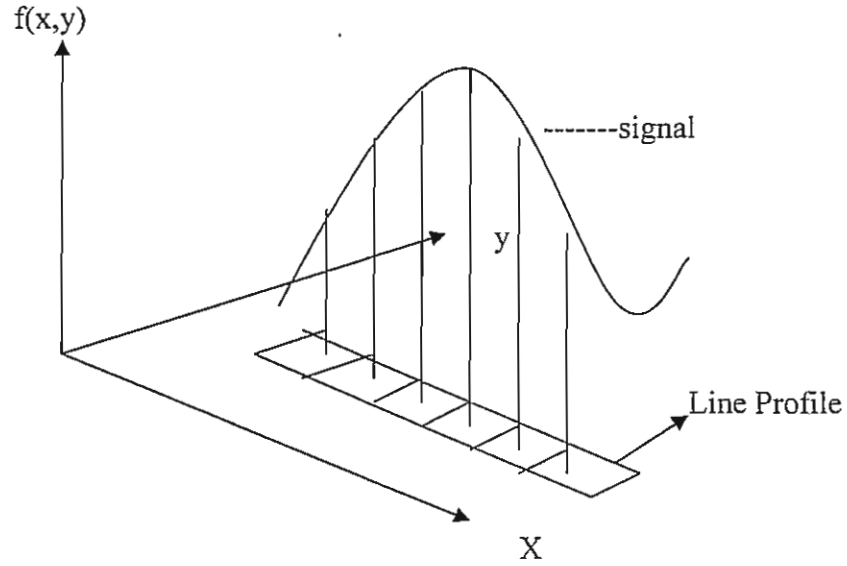


Figure 3.2(b) A Line profile: A two-dimensional light intensity function f(x, y), where x and y denotes spatial coordinates.

Each line profile consists of a set of p pixels with its corresponding gray level intensity value f(x,y) as illustrated in figure3.2(b). In another word, we can regard each line profile as a vector, say $\mathbf{u}^T = \mathbf{f} = [f(x_1), f(x_2), \ldots, f(x_p)]$. Wavelet analysis will then be performed on each line profile.

3.4 Wavelets transformation.

Wavelets analysis breaks up a signal into scaled and translated version of the wavelets-special function called mother wavelet. In this section, we introduce one of the mother wavelet of the Daubechies family wavelets, the db4 which is the fourth order of the Daubechies family. See Daubechies [1992]. In wavelets transform, we start with a small scale of the mother wavelets and continuously translate it along the function. If a section of the original signal is covered by wavelets, we calculate it as wavelet coefficients, Cw. This coefficient explains the quality of the match between the original signal and the mother wavelets. A high Cw represents a great similarity. Calculating Cw at every possible scale is a tedious work, therefore, discrete wavelet transform is used to make our analysis

more efficient and just as accurate by choosing only a subset of scales and position based on power of two.

An efficient way to implement this scheme using filters was developed by Mallat [1989]. Here, we will discuss this practical filtering algorithm which yield a fast wavelet transform briefly. For many signal, the low frequency content is the most important part as it gives the signal its identity. In wavelets analysis, the approximations are the high scale, low frequency components of the signal whereas the details are the low scale, high frequency components.

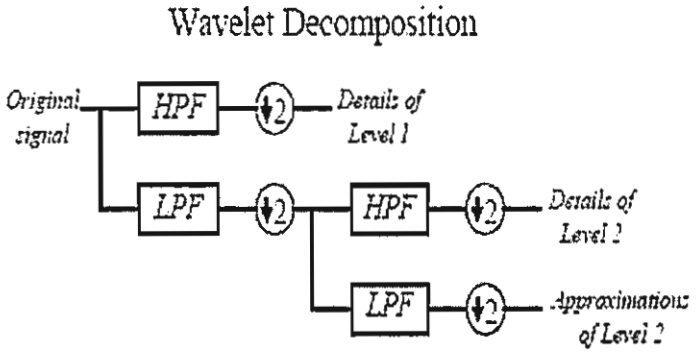


Figure 3.3: Wavelet transform at two-level decomposition

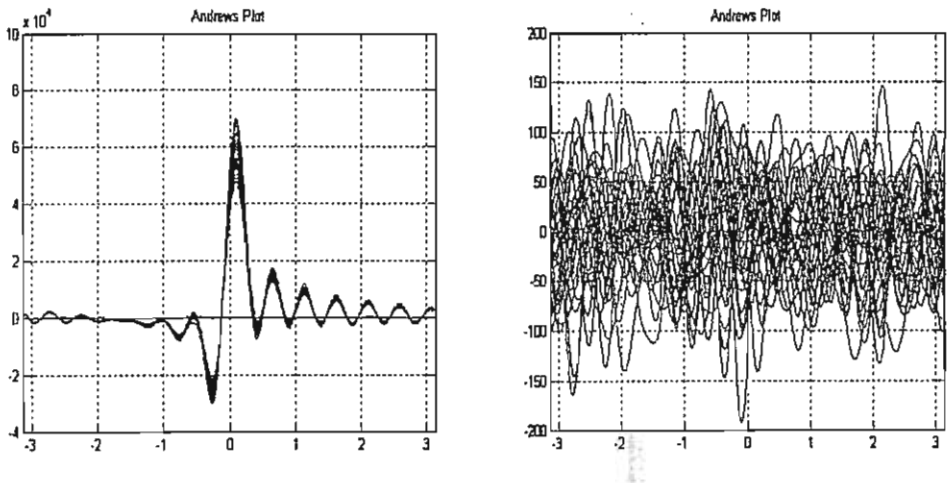
In the filtering process, the original signal passes through two filters and emerges as two signals A and D. These signals are useful but we would end up with twice, (2n) as much as the data we started with (n). By looking at the computation, we may keep only one point out of two in each of the two 2n length samples to get the complete information. This is the downsampling process (shown by the arrow). We eventually obtain two sequences called approximate coefficients cA and detail coefficients cD as illustrated in figure 3.3. These are the DWT coefficients.

The cD are small and consist mainly of a high frequency noise, while the cA contain lesser noise than does the original signal.

This decomposition process can be iterated indefinitely with successive approximation being decomposed in turn so that the signal can be broken down into many lower resolution components.

3.5 Andrews curve of Wavelets Coefficients

In this study, we will only use approximation and detail coefficients from the first level of decomposition. Having this matrix of coefficients, we will represent vectors of each line profiles of wavelets coefficients graphically using Andrews curve. Vectors of each line profiles $\mathbf{u}^T = \mathbf{f} = [\ f(x_1), f(x_2), \ldots, f(x_p)\ ]$ are transformed into $f(t) = \mathbf{a}^T\mathbf{u}$ and then plotted on the same axes over the range $-\pi \leq t \leq \pi$. Since the value of t is in the range of $-\pi$ to π and with increment of 0.01 each time, we have to repeat the computation of $f(t) = \mathbf{a}^T\mathbf{x}$ 629 times each for every observation. Eventually we will obtain a matrix of 629x30. We plot all the 629 values of f(t) of one observation value against its corresponding t value and by joining these points, we obtain a sinusoidal curve. Therefore by doing the same for all 30 line profiles, we have a set of 30 Andrews curves drawn across the plot. Andrews curves for approximate and details coefficients are illustrated in Figure 3.4.



a) Approximation coefficients b) Details coefficients

Figure 3.4: Thirty Andrews curves of approximation and details coefficients over the range $-\pi \leq t \leq \pi$.

From the above result, it can be clearly seen that Andrews curves of detail coefficients portray a rather messy display with out any common pattern and a detail study have to be carried out if we wishes to exert information from it. On the other hand, a closer examination on the Andrews curve of approximation coefficients immediately reveals some interesting information that warrants further studies i.e. at certain value of t, natural grouping somehow does occur. Therefore, from this section onward, we will use only the approximation coefficients. Both approximation and details coefficients are attached in Appendix A.

3.6 Clustering of Line Profile

Through cluster analysis, we hope that we will obtain some information about the similarity between each line profile and some natural grouping among them. The first step now is to calculate the Euclidean distance between each line profiles using equation (1.11) and form a matrix of similarity distance, D. The Euclidean Matrix of approximation coefficients are attached in Appendix A. After having done this, clustering methods such as Single Linkage, Complete Linkage, Average Linkage, Centroid Method and Ward Method are applied on the Euclidean matrix to cluster the line profiles. The outcomes of each of the clustering methods are display graphically in a separate dendrogram as illustrated in figure 3.5.

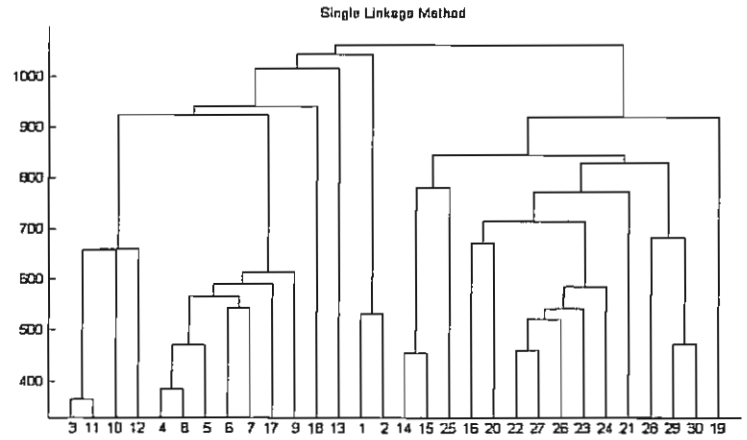


Figure 3.5 (a) Dendrogram of Single Linkage Method.

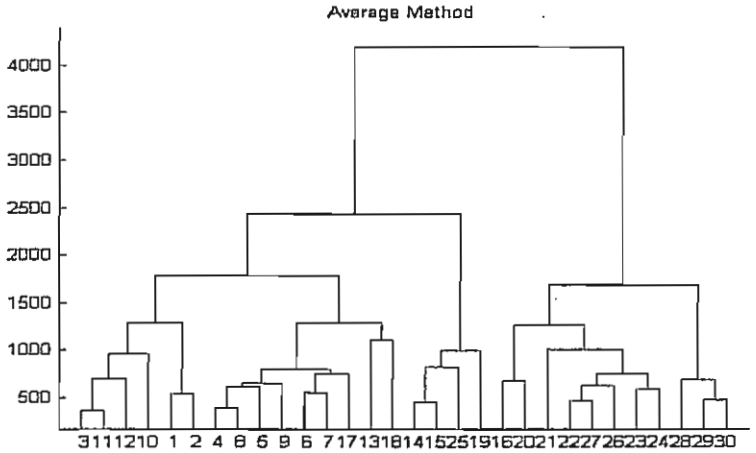


Figure 3.5 (b) Dendrogram of Average Linkage Method

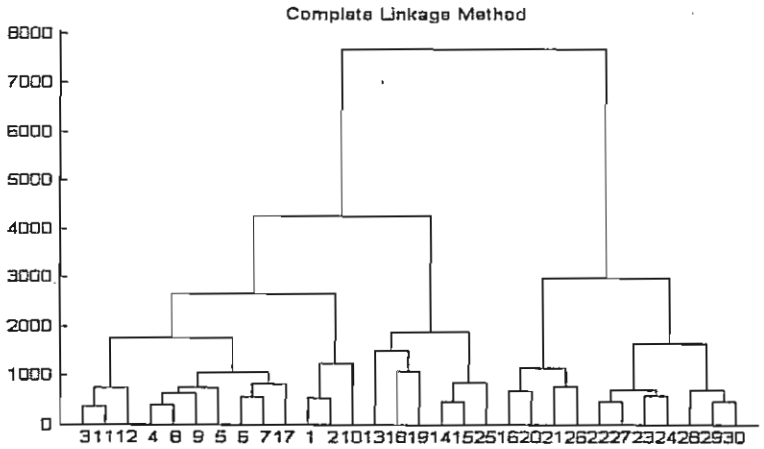


Figure 3.5 (c) Dendrogram of Complete Linkage Method.

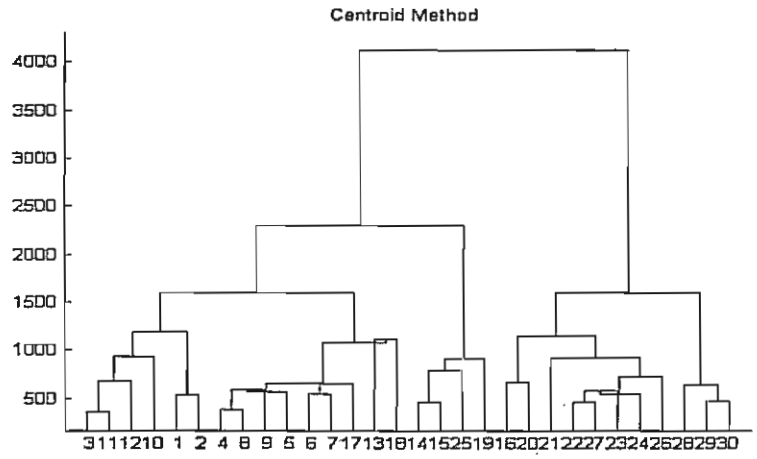


Figure 3.5 (d) Dendrogram of Centroid Method

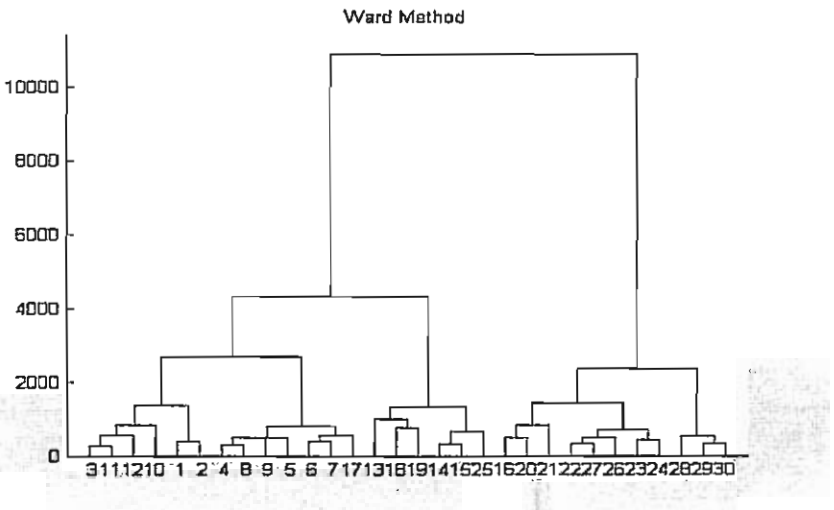


Figure 3.5 (e) Dendrogram of Ward Method

From the above results, when we fix the threshold value as two, we can see that each of the clustering method namely Average Linkage, Complete Linkage, Single Linkage, Centroid Method and Ward Method group rather consistant sets of line profiles under each of two clusters formed namely A and B. At times, this might not happen because different clustering algorithms are involved and perhaps the groups are not spherically shaped and well separated. In order to get the best representation of the line profiles in A and B, the average vectors in A and B of every clustering method $\overline{x_i}$ i=1,...,10 is used. Therefore, all average vectors of the first clusters obtained by different clustering methods are used to get the average of average vector, $\overline{\overline{x_1}}$. The same is done for the second cluster to obtain $\overline{\overline{x_2}}$. These average vectors are then used to plot Andrews curve and the outcomes are illustrated in figure 3.6.

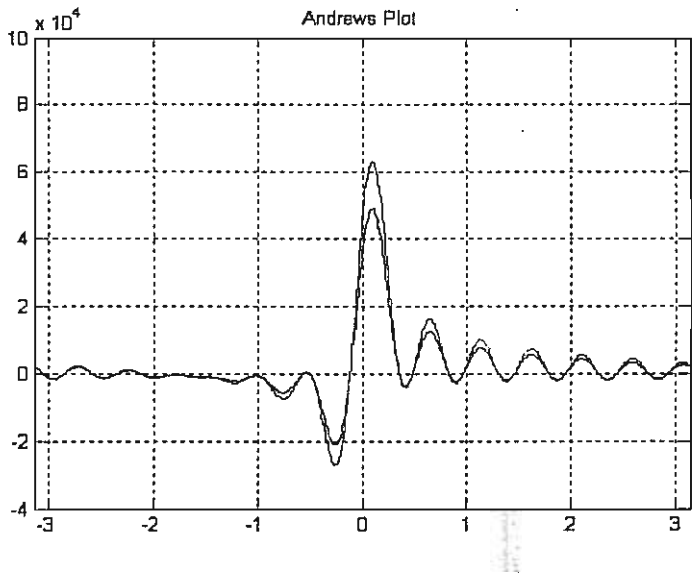


Figure 3.6: Andrews plot of average MTB line profiles and probable MTB line profiles.

From figure 3.6, we can now clearly see that at certain value of t, both average Andrews curve are well separated. For better visualization of the region in a lung infected by MTB, the Andrews curve of higher value of wavelet coefficient will be highlighted in red and the other in blue. The red lines are defined as the MTB line profiles and the latter a non MTB line profiles. On the x-ray, however, the most numbers of line profiles of higher value of wavelet coefficient formed under all clustering method will be used instead of using the average number of line profiles. This is done so as to reduce the risk of misclassifying an infected area as not infected.

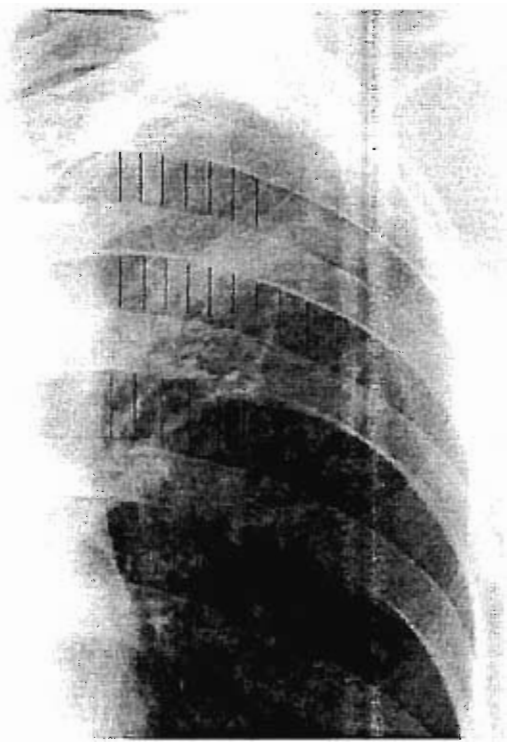


Figure 3.7: Two clusters of line profiles marked red and blue respectively on the lung x ray of patient A. The former represent MTB region and latter represent a non MTB region.

3.7 Discussion

In this study, we demonstrate an objective method of detection of Mycobacterium Tuberculosis (MTB) and address some aspect of analyzing the content of an image based on line profiles obtained from a person's chest x-rays. In this study, line profiles are used as method of feature extraction. Line profiles obtained from a subset of chest x-ray segmented are then transformed into Andrews curve for analysis. Multivariate clustering methods has been proposed in search of natural clusters within them. The results illustrated using dendogram shows that two groups of Andrews curve can be formed. These two groups resembled the MTB and non MTB region of the lung. A collection of line profiles from the same cluster provide a rough estimate of boundary of the infected and non infected area.

Detection of MTB using line profiles may somehow provides a more accurate conclusion as the analysis is based on the values of intensity level rather than individual visual perception. This is due to the fact that individual visual interpretation may sometimes be subjective as a very experience medical practitioner may give a more accurate diagnosis result than a newly trained.

In some cases, analysis using line profiles may be affected by the poor quality of x-rays images obtained. However, with the existence of digital imaging technology, techniques such as enhancement and filtering can be used to enhance the quality of the x rays. These 'enhanced' x-rays used may enable us to obtain a more satisfactory MTB detection result.

However, the procedures demonstrated here is merely a preliminary detection of the possible MTB infection. Clinical trial/ test (e.g. tuberculin skin test, sputum test, temperature, weight loss, cough and etc.) is jointly needed to verify this claim. Besides that, a basic knowledge of the chest x-rays radiology is essential as it does influence the line profiles acquisition procedures. For example, an area of nodus/ nerves may be mistaken for a MTB infected region.

CHAPTER 4: DISCRIMINATION BETWEEN MTB AND LUNG CANCER.

4.1 An Introduction to Lung Cancer

Lung cancer is the most common cause of cancer death in the world for both men and women. A lung cancer occurs when cells in the lung start to replicate uncontrollably forming growth called tumors. These tumors are malignant, meaning that they invade and destroy surrounding healthy cells and tissue. Studies have shown that a history of interstitial lung disease or MTB also increases the risk of getting lung cancer and lung cancer sometimes resembles MTB on lung x-ray. See Rubin [2001], Marcus [2000], Adjei [1999] and Bunn [2000].

4.2 Detection of Lung Cancer

A standard chest x ray can reveal an abnormal mass or nodule in the lung and a CT scan may show very small lesions and whether cancer has spread to other areas. But as with all type of cancer, lung cancer can be definitely diagnosed by looking at x-ray. Abnormal mass of cancer appears lighter in color on x-ray films than does normal, healthy lung tissue and MTB infected lungs. Therefore it is of interest to compare such a characteristic (the level of intensity values on x-ray) between MTB and LC as it may yield interesting results.

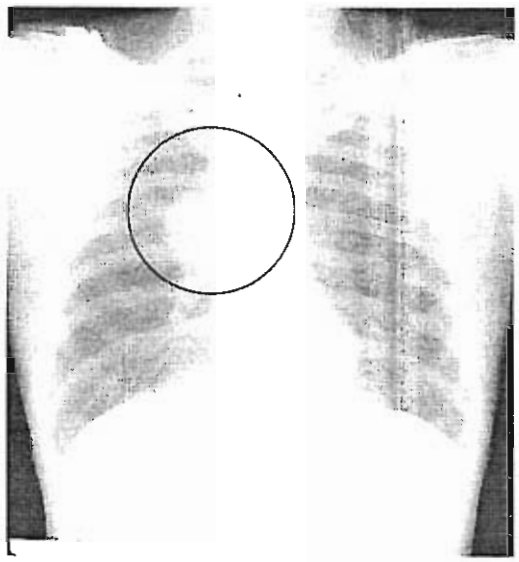


Figure 4.1: An x-ray of a Lung cancer patient. Red circle indicate a lung cancer region.

4.3 Comparison of Andrews Curve between MTB, LC and Healthy Lung.

X-rays from 11 confirm MTB patients, 9 lung cancer patients and 4 healthy people are used. The same procedures describe in chapter three will be used here except that no clustering methods within a chest x-ray of a lung cancer and MTB patient are involved. 30 line profiles are obtained from the LC/MTB/disease free region of each x-ray image with advice from a medical expert as illustrated. For every patient, vectors of the mean all 30 line profiles are then transformed to an Andrews curve. The results of 11 TB patients (in blue) and 9 lung cancer patients (in red) and 4 healthy people (in black) are as the following.

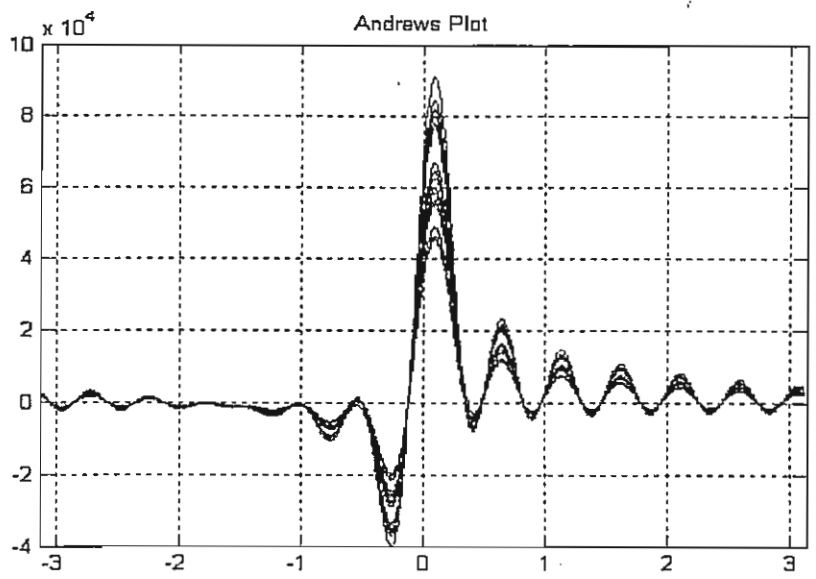


Figure 4.2(a): Andrews curve in the range of -pi <t< pi.

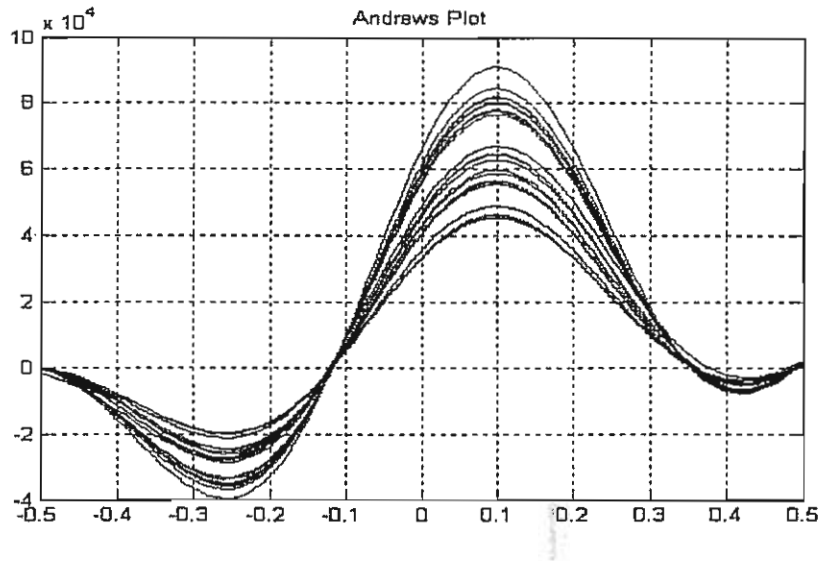


Figure 4.2(b): Andrews curve in the range of $-0.5 < t < 0.5$

From the above results, we can see that 11 TB patients (in blue) and 9 lung cancer patients (in red) and 4 healthy people (in black) form 3 distinct groups of clusters. Both figure 4.1 also shows that at certain values of t, namely t_i, three distinct groups of Andrews curves were observed. These values of t_i are worth considering for further study.

4.4 Selection of t_i

Values of t_i are obtained in a way such that for every value of t >0 with increment of 0.01 unit each time, we calculate the Euclidean Distance Difference (EDD) between two furthest curve. The t value with the largest value of EDD at every interval of 0.5 unit of t will be taken as t_i i=1, 2,... k. We define EDD as the difference between the maximum value of f(t) and the minimum value of f(t) of all 24 curve at t point i.e.

$$\text{EDD at } t = f_{max}(t) - f_{min}(t) \qquad [4.1]$$

The value of t_i and its corresponding f(ti) value for all 24 Andrews curves where i=1,2,...,6 are listed in Table 4.1.

t_i		t_1	t_2	t_3	t_4	t_5	t_6
Value of t_i		**0.10**	**0.66**	**1.16**	**1.64**	**2.12**	**2.60**
Value of $f(t_i)$	H1	45327	11483	7220.5	5295.3	4102.3	3228.3
	H2	48797	12335	7664.6	5696.9	4445.3	3427
	H3	46224	12211	7350.2	5473.6	4117.5	3251.8
	H4	46042	11981	7499.9	5398.2	4221.3	3263.2
	TB1	55370	13815	8771.4	6581.5	5122.9	4012.4
	TB2	58503	14836	9303.3	6859.6	5251.5	4081.7
	TB3	64116	16449	10226	7455.1	5811.8	4512.1
	TB4	59649	15459	9570.2	7071.1	5490.3	4258.3
	TB5	62579	15574	9816.6	7066.1	5645.3	4317.9
	TB6	58509	15128	9350.4	6765.7	5097.5	3962.6
	TB 7	66763	15922	10344	7615.3	5921.9	4499.6
	TB8	64529	16363	10217	7601.8	5991.8	4679.4
	TB9	56132	14644	9027.6	6664	5127.3	4019.7
	TB10	56357	14833	9133.2	6694.8	5279.9	4047.3
	TB11	77916	19840	12355	9195.8	7079.2	5556.1
	LC1	81861	21050	13182	9779.2	7624.4	5966.4
	LC2	80081	20448	12797	9424.3	7356.3	5762.2
	LC3	81507	21057	12969	9538.2	7399.1	5741.2
	LC4	76411	19561	12140	8965.4	6916.1	5419.3
	LC5	84470	21572	13399	9886.4	7647.3	5975.8
	LC6	90854	23160	14456	10690	8289.3	6503.7
	LC7	81385	20689	12888	9561.2	7461.1	5848.7
	LC8	79708	20300	12720	9350	7260	5752.3
	LC9	77352	19752	12333	9134.4	7108.3	5576.8

Table 4.1: Value of t_i and its $f(t_i)$ for all 23 Andrews curves.

4.5 Discrimination and Misclassification

At a t point, all 24 values of f(t) is obtained and median value of each cluster is calculated. This is done so that future classification can be carried out by comparing the value of the three median and the value of $f_w(t)$ of an unknown new curve, say W, at selected value of t. For a set of Andrews curve at selected value of t where $f_1(t) < f_2(t) < ... < f_n(t)$, the median is defined as

$$\text{Median} = \frac{1}{2}[f_k(t) + f_{k+1}(t)] \quad \text{if n=2k} \qquad [4.2]$$

$$\text{Median} = f_k(t) \quad \text{if n=2k-1} \qquad [4.3]$$

The classification rule is defined as the following:

Rule	Description
A	if $f_w(t)$ value is closer to median of lung cancer's group, we say W belongs to the lung cancer's group
B	if $f_w(t)$ value is closer to median of tuberculosis' group, we say W belongs to the tuberculosis' group
C	we say W belong to the healthy groups

Table 4.2: Classification Rule

To assess the performance of this classification method, we estimate the misclassification probabilities. Since our sample size is quite small we use the one-in-one-out method on the samples to estimate the misclassification probability. Each time, out of the 24 Andrews curves, we will take out one curve, say W and leave it aside. The remaining 23 curves will then be used as the control sets. At the selected value of ti, where i=1,2, …,6, the median of $f(t_i)$ for each group (lung cancer, tuberculosis and healthy) formed by the remaining curves will then be calculated. These values will be used as the reference values for classifying the earlier left out curve. However, since W is a known curve, therefore, at any selected value of ti, if W falls in any groups other than the group it is supposed to belong to, then we say misclassification occurs. The results of misclassification are shown in table 4.3.

Sample	W	t1	t2	t3	t4	t5	t6
1	H1	/	/	/	/	/	/
2	H2	/	/	/	/	/	/
3	H3	/	/	/	/	/	/
4	H5	/	/	/	/	/	/
5	TB1	/	/	/	/	/	/
6	TB2	/	/	/	/	/	/
7	TB3	/	/	/	/	/	/
8	TB4	/	/	/	/	/	/
9	TB5	/	/	/	/	/	/
10	TB6	/	/	/	/	/	/
11	TB7	/	/	/	/	/	/
12	TB8	/	/	/	/	/	/
13	TB9	/	/	/	/	/	/
14	TB10	/	/	/	/	/	/
15	TB11	X	X	X	X	X	X
16	LC1	/	/	/	/	/	/
17	LC2	/	/	/	/	/	/
18	LC3	/	/	/	/	/	/
19	LC4	/	/	/	/	/	/
20	LC5	/	/	/	/	/	/
21	LC6	/	/	/	/	/	/
22	LC7	/	/	/	/	/	/
23	LC8	/	/	/	/	/	/
24	LC9	/	/	/	/	/	/

X= misclassification / = correct classification

Table 4.3: Results of classification of a known Andrews curve.

From the above results, we obtained a 95.8% of correct classification (which is equivalent to 4.167 percent of misclassification) based on small sample size. a high percentage of correct classification was obtained due to the fact that all these curves are from known population. 4.167% of misclassification is due to the reason that TB11 patient has developed cavities (serious stage of MTB).

4.6 The Median Graph

For a better illustration, a median plot can be established by just plotting the median point, $f_{med}(t)$ of the control group against certain selected value of t. Below are the value of median of all three groups computed from all 24 Andrews curves and the plot is illustrated in Figure 4.2. By using the same classification rule defined earlier in Table 4.2, we can perform discrimination procedure and classify a new Andrews curve. Note that always the median of LC>median of TB>median of H.

Group	t1	t2	t3	t4	t5	t6
H	46133	12096	7425	5435.9	4169.4	3257.5
TB	59079	15294	9460.3	6962.9	5385.1	4170
LC	81385	20689	12888	9538.2	7399.1	5762.2

H=healthy lung TB=tuberculosis LC=lung cancer

Table 4.4: Reference value of median

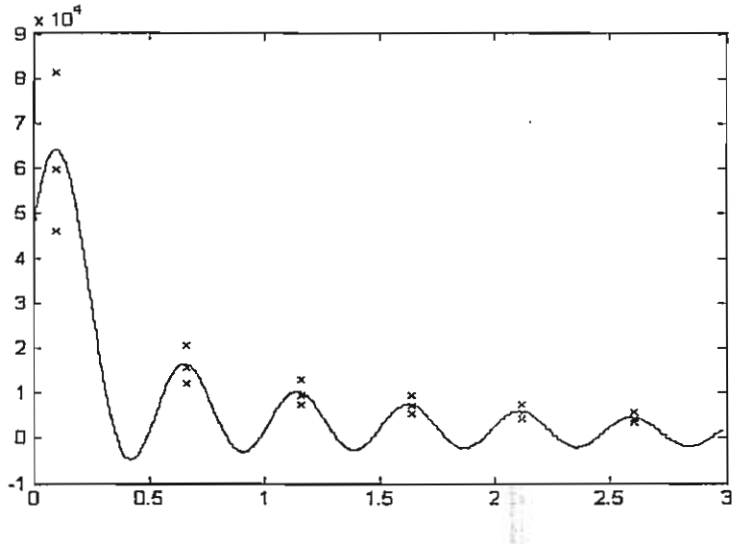


Figure 4.3: Plots of median point of all three groups i.e. LC, MTB, H at t1, …, t6 and an Andrews curve of a MTB patient.

4.7 Discussion

With a little modification to the earlier define procedures in detection of MTB, our results shows that Andrews curve of LC, MTB and H can be differentiated from one another at certain selected value of t. This happen due to the fact that a region of lung tumor has a lower density level than of a region of healthy lung and region of MTB falls between them. At certain value of t, observation using Andrews curve are able to show clusters of the three groups. In order to discriminate these groups, median of each group is used as the reference point as it is not much affected by the existence of possible outliers compared to mean. Classification of an Andrews curve is then based on the 'nearest Euclidean Distance' rule. A simple testing is carried out to test the misclassification probability and the results obtain are quite satisfactory i.e. only 4.167 % of misclassification occurs. A median plot can be constructed to illustrate the classification and discrimination rule graphically.

CHAPTER 5: A SIMULATION STUDY OF ANDREWS CURVES

5.1 Introduction.

The results from empirical studies in chapter 3 show that Andrews curve are capable of differentiating the 'disease' line profiles from those of 'disease-free' line profiles. When line profiles are from known probability distribution, a detail study of the behavior of Andrews curve can be carried out. Therefore in this chapter, vectors from known multivariate normal distribution are generated and then converted to its corresponding Andrews curves to study in particular the potential of using groups of curves to perform discrimination. Generating these vectors can be regarded as having large samples of line profiles.

Here, in simulation, two situations are considered;

a) generating a set of Andrews curves from a given normal population with
   i) fixed mean and varying variance;
   ii) varying means and fixed variance;
   iii) varying variables

with the aim to have a preliminary look at the behaviors of the curves.

b) generating two sets of Andrews curves from two known normal population with
   i) fixed mean and varying variance;
   ii) varying means and fixed variance;

with the aim to study the potential of using groups of curves to perform discrimination.

5.2 Simulation for One Normal Population

We are interested in generating a set of normally distributed multivariate data. For a given observation from $Np(\mu^*,\sum)$, we can transform it to an simpler form i.e. observation from $Np(\mu,D)$ where $D= diag(d_1,d_2,\ldots, d_p)$. Therefore, in the following section, n vectors (which is equivalent to n lines profiles) will be generated from $Np(\mu,D)$ for a selected value of μ and D. The procedure is outlined in Figure 5.1. These $\mathbf{x}_1, \mathbf{x}_2\ldots, \mathbf{x}_n$ are then transformed to $f_j(t)= \mathbf{a}^T\mathbf{x}_j$ j=1, 2,..., n. Based on multivariate property in Appendix B, $f_j(t)= \mathbf{a}^T\mathbf{x}_j$ has the distribution of $Np(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T D\mathbf{a})$. A plot of [t, f(t)] is then drawn to form n Andrews curves corresponding to $\mathbf{x}_1, \mathbf{x}_2\ldots, \mathbf{x}_n$. These procedures are shown in Figure 5.2. The Andrews curves of the generated data are illustrated in Figure 5.3 and 5.4.

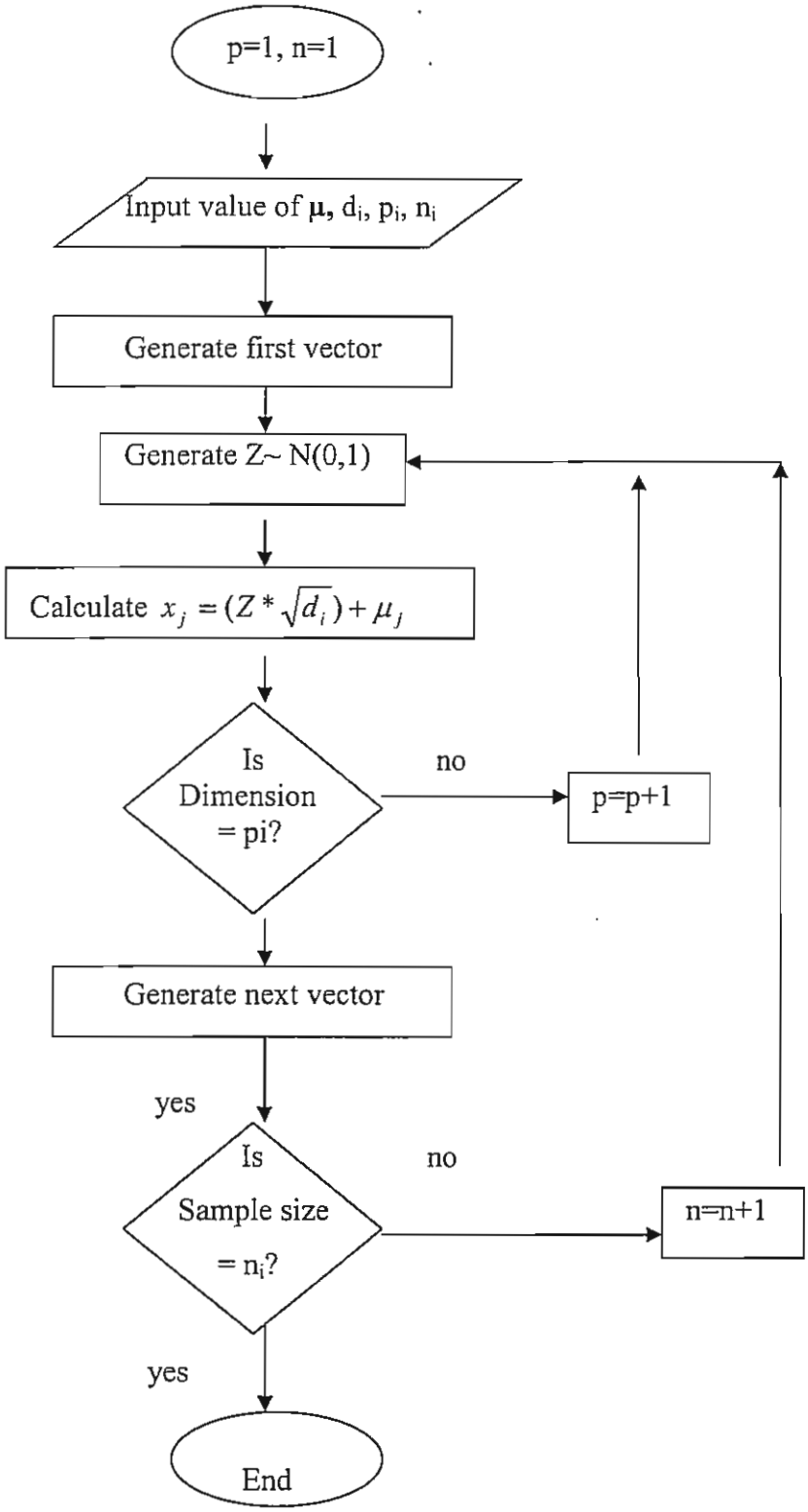


Figure 5.1 Flowchart showing generation of $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n$ from $N_p(\boldsymbol{\mu}, \mathbf{D})$ where $\boldsymbol{\mu}^T = (\mu_1, \mu_2, \ldots, \mu_2)$ and $D = \text{diag}(d_1, d_2, \ldots, d_p)$.

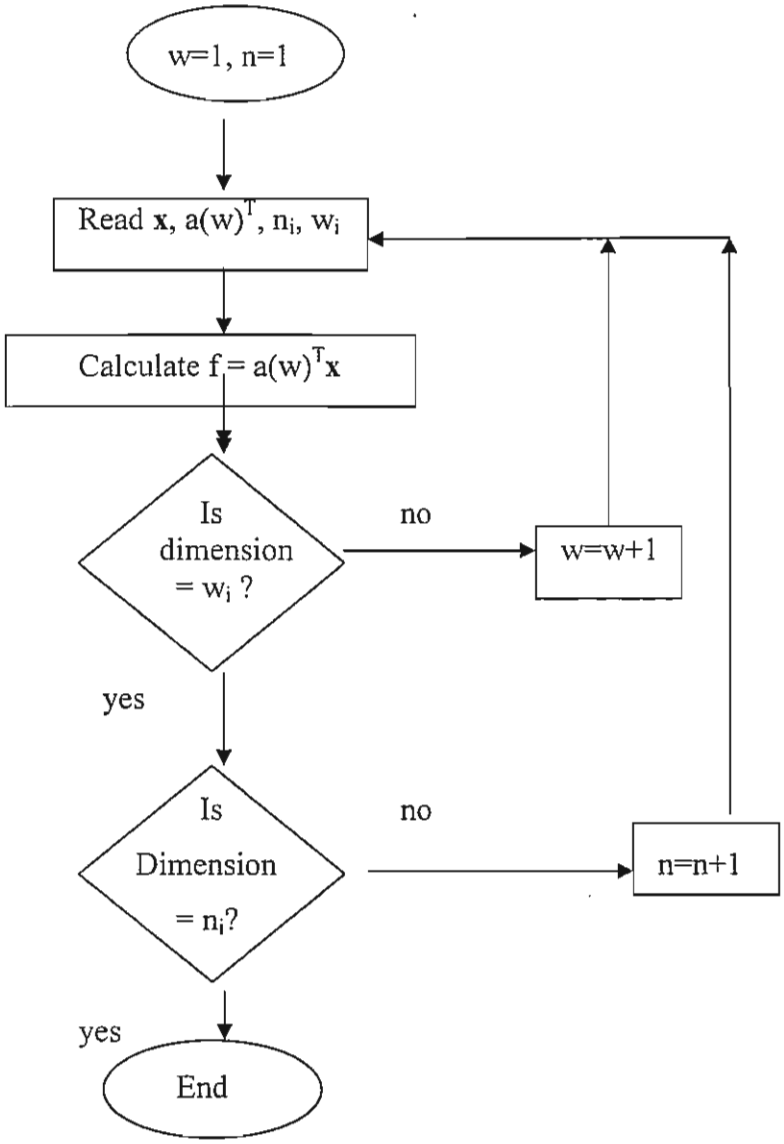
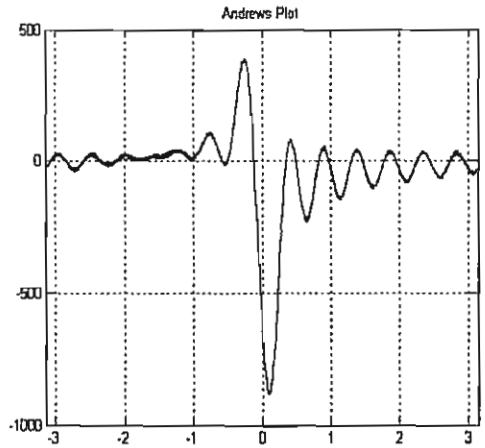


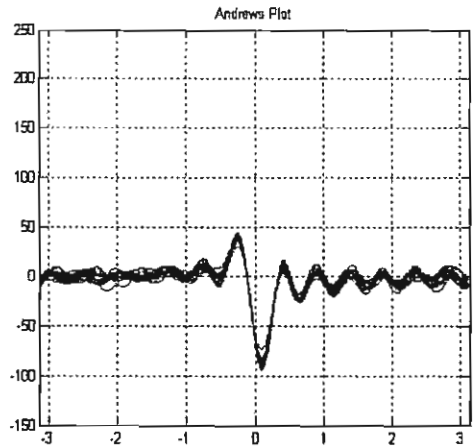
Figure 5.2: Flow chart showing transformation of **x** to $[\mathbf{a}(w)]^T\mathbf{x}$.

$$\mathbf{a}(w)=\left[\frac{1}{\sqrt{2}}, \sin \pi w, kos\pi w, \sin 2\pi w, kos2\pi w....\right]$$

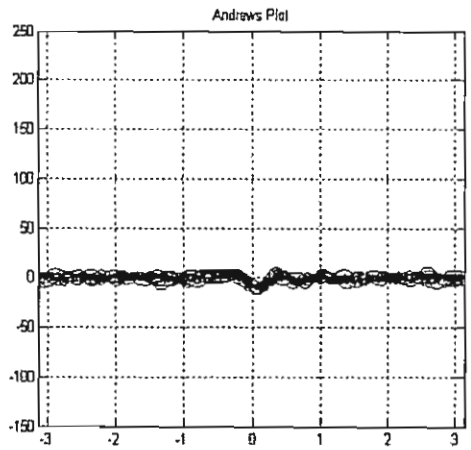
5.2.1 Varying mean with fixed Variance



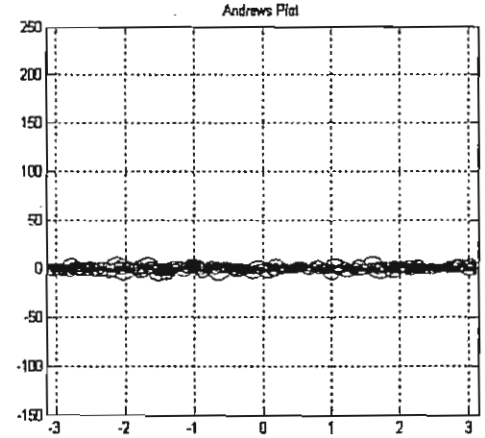
a) n = 40, c = -50, σ^2 =1, p=26



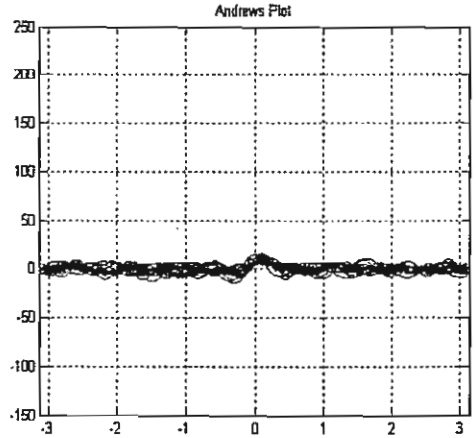
b)n = 40, c = -5, σ^2 =1, p=26



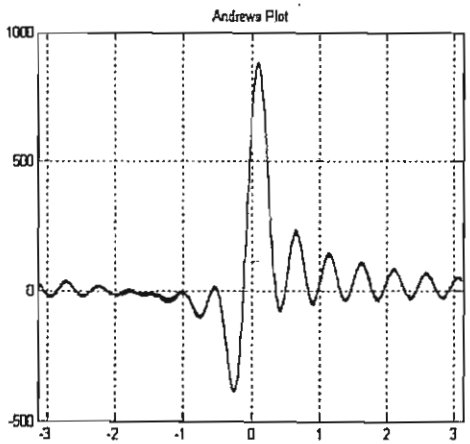
c) n = 40, c = -0.5, σ^2 =1, p=26



d)n = 40, c = 0, σ^2 =1, p=26



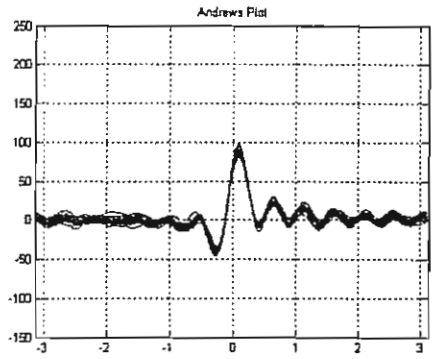
e) n = 40, c = 0.5, σ^2 =1, p=26



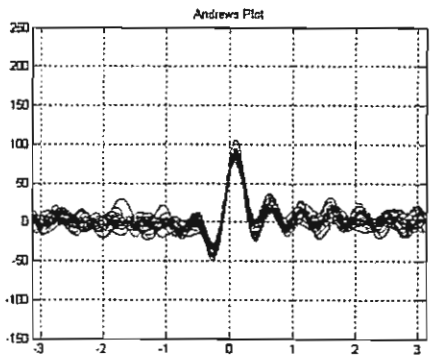
f) n = 40, c = 50, σ^2 =1, p=26

Figure 5.3: Andrews curve for sample of size n, variable p, from N(μ, D), where μ=c(1,1,…,1) and D=diag(σ^2, σ^2,…, σ^2).

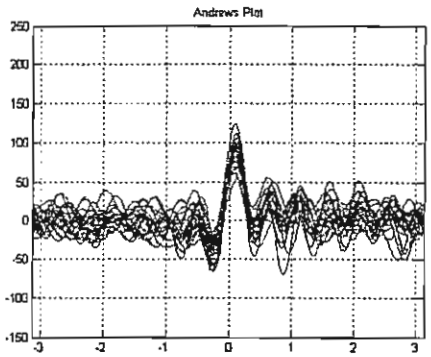
5.2.2 Varying Variance and Fixed Mean



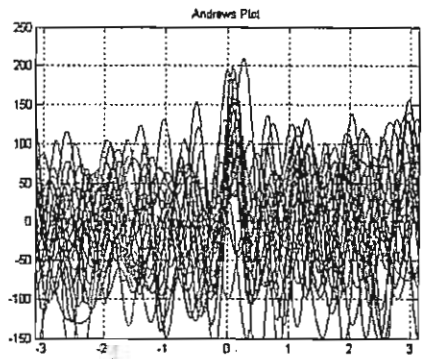
a) n = 40, c = 5, σ=1, p=26



b) n = 40, c = 5, σ =2, p=26



c) n = 40, c = 5, σ =4, p=26



d) n = 40, c = 5, σ =9, p=26

Figure 5.4: Andrews curve for sample of size n with p variables from N($\boldsymbol{\mu}$, D), where $\boldsymbol{\mu}$=c(1,1,...,1) and D=diag(σ^2, σ^2,..., σ^2).

5.2.3 Discussion

When vector of p variables generated from the same population, all n curves tend to behave in the same way and overlaps with each another group. As the value of mean moves from zero, the curve exhibits a clearer sinusoid curve and has more peaks. When c takes a negative value, the highest peak will have negative value and when c takes a positive value, the highest peak will be positive as well. As the value of variance increase with fixed mean, the degree of thickness of the curve will increase. The reason for this behavior is explained in section 1.6 (property c). Apparently, not much information can be obtain when value of variance increases. This can be seen in figure 5.4.

5.3 Simulation for Two Normal Populations

In this section, we are interested in generating two sets of n_i, i=1,2 vectors coming from different normal distribution each i.e. πi, $xi \sim Np(\mu_i, \sum_i)$, i=1,2. However, when generating $xi \sim Np(\mu_i, \sum_i)$, i=1,2 many parameters are involved and analyzing it would be rather difficult. By using the theorem below, the numbers of parameters involved are reduced as **x** is transformed into **x***, where now $\mathbf{x}^* \sim Np(\boldsymbol{\mu}, D)$ or $\mathbf{x}^* \sim Np(\mathbf{0}, \mathbf{I})$.

Theorem 1:

We have for population πi, $x \sim Np(\mu_i, \sum_i)$, i=1,2. We apply a linear transformation $\boldsymbol{x} \rightarrow A\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{x}^*$ *and convert the distribution into canonical form as the following:*

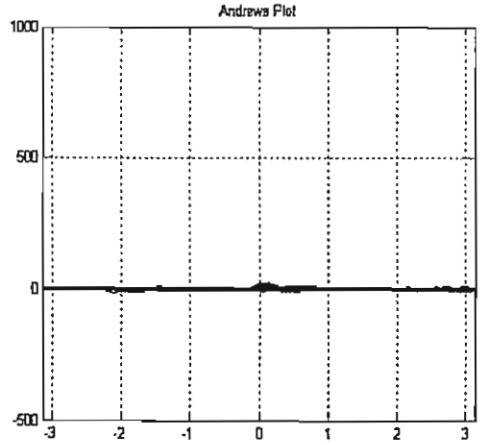
$$\pi_1 : \boldsymbol{x}^* \sim Np(\boldsymbol{\mu}, D)$$

$$\pi_2 : \boldsymbol{x}^* \sim Np(\boldsymbol{0}, \boldsymbol{I})$$

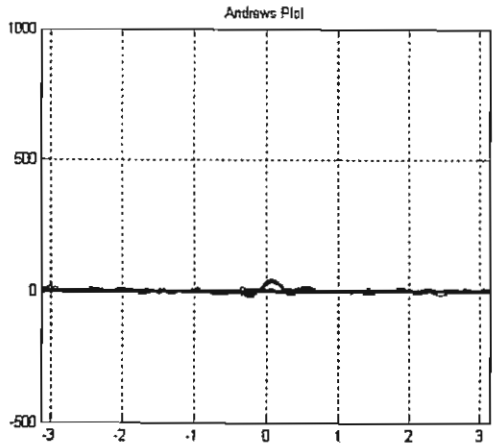
where $\boldsymbol{0} = (0, 0, \ldots, 0)^T$ *and* $\boldsymbol{\mu}$ *is a p variate vector.* See Appendix B for proof.

The procedures for generating sets of $\boldsymbol{x}^*$is the same as illustrated in Figure 5.1 and 5.2 and the results are illustrated in Figure 5.5, 5.6 and 5.7.

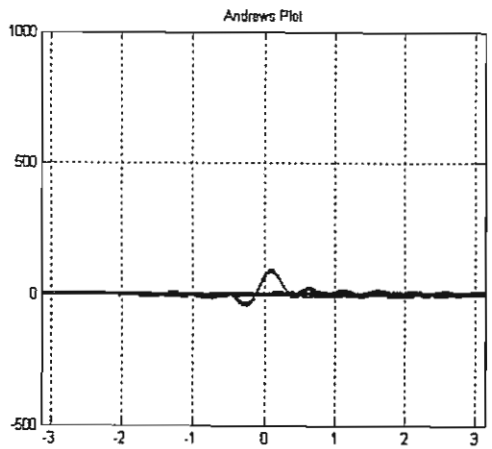
5.3.1Varying mean and fixed variance.



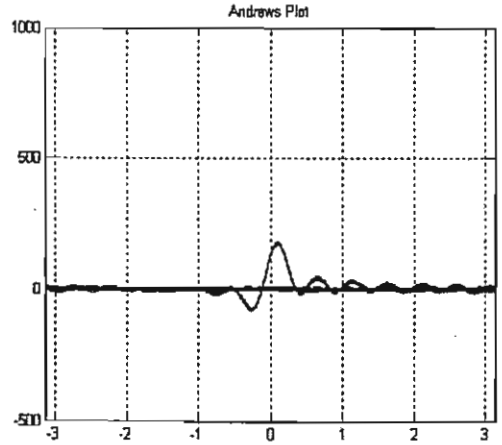
a) n = 40, c = 1, σ^2 =1, p=26



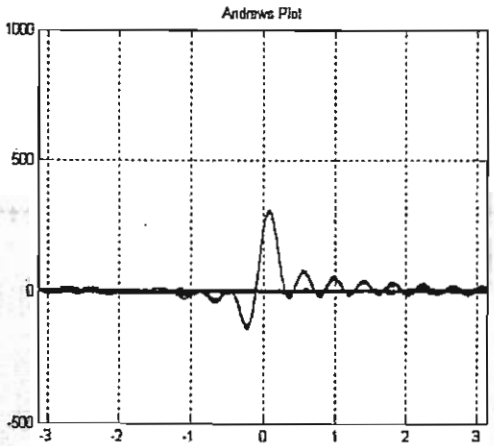
b) n = 40, c = 2, σ^2 =1, p=26



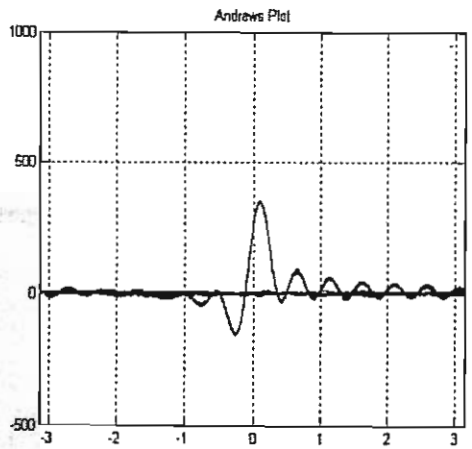
c) n = 40, c = 5, σ^2 =1, p=26



d) n = 40, c = 10, σ^2 =1, p=26



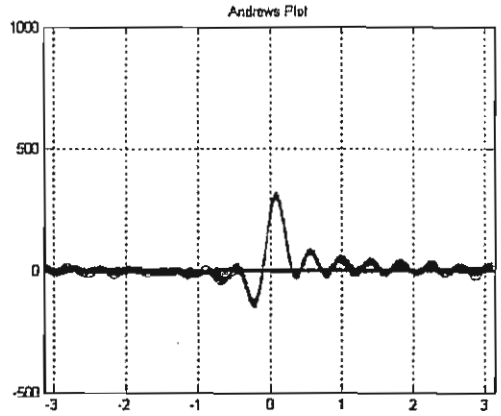
e) n = 40, c = 15, σ^2 =1, p=26



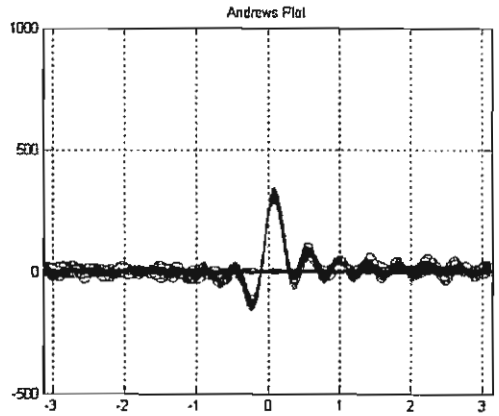
f) n = 40, c = 20, σ^2 =1, p=26

Figure 5.5: Andrews curve for (a) sample of size n with p variables from N(μ, D), where μ=c(1,1,…,1), D=diag(σ^2, σ^2,…, σ^2) and (b) sample of size n with p variables from N (I,0).

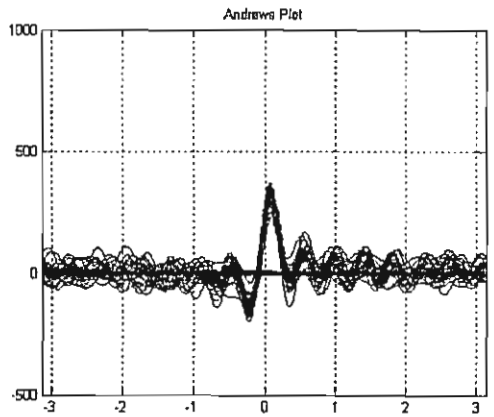
5.3.2 Varying variance and fixed mean



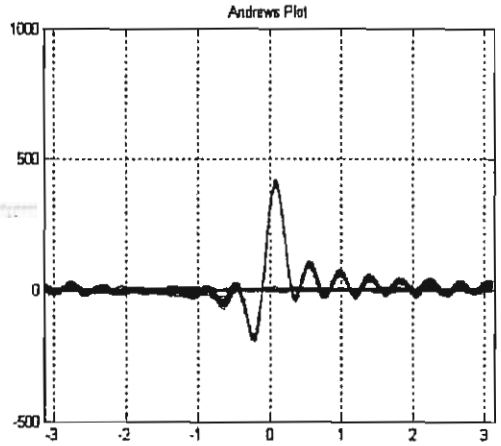
a) n = 40, c = 15, σ =2, p=26



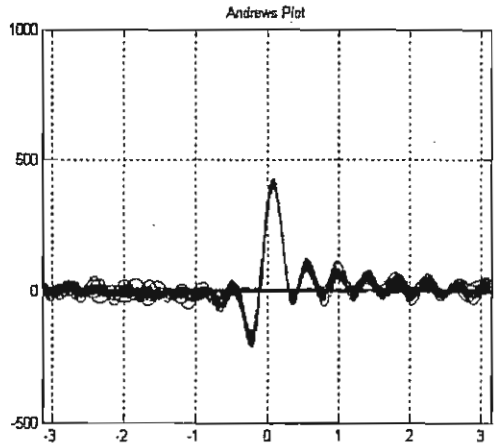
b) n = 40, c = 15, σ =4, p=26



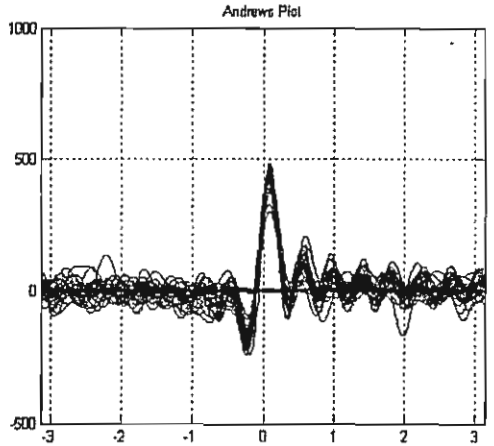
c) n = 40, c = 15, σ =9, p=26



d) n = 40, c = 20, σ =2, p=26

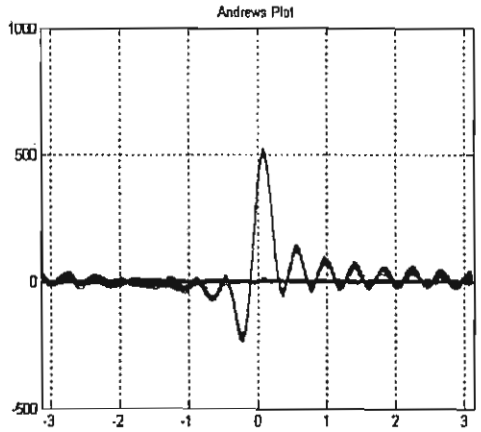


e) n = 40, c = 20, σ =4, p=26

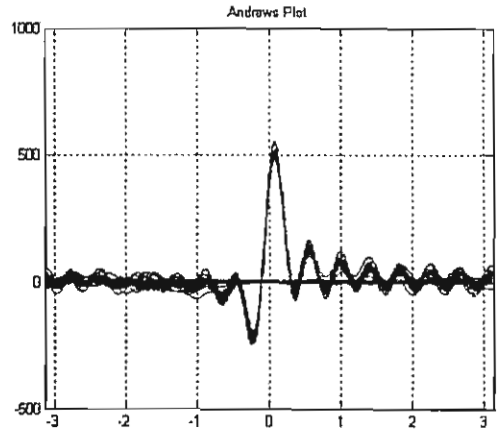


f) n = 40, c = 20, σ =9, p=26

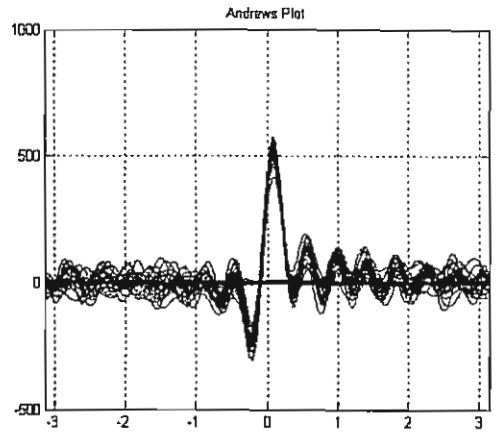
Figure 5.6(a): Andrews curve for (a) sample of size n with p variables from N(μ, D), where μ=c(1,1,…,1), D=diag(σ^2, σ^2,…, σ^2) and (b) sample of size n with p variables from N (I,0).



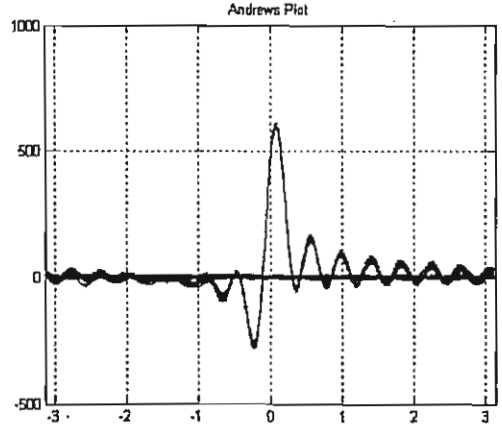
a) n = 40, c = 25, σ =2, p=26



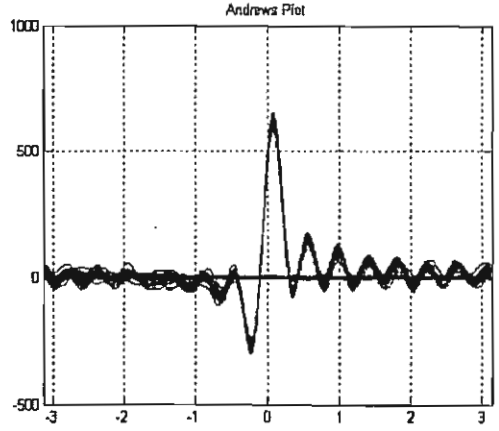
b)n = 40, c = 25, σ =4, p=26



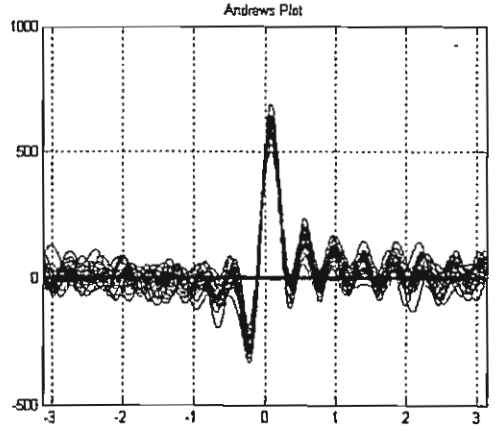
c) n = 40, c =25, σ =9, p=26



d) n = 40, c = 30, σ =2, p=26



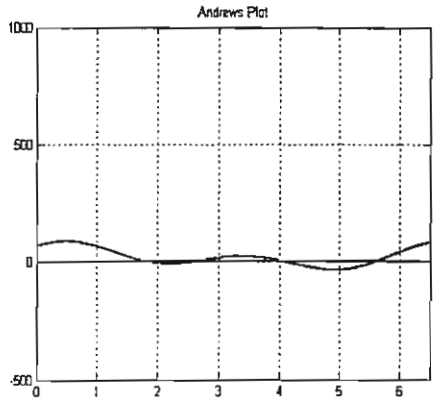
e)n = 40, c = 30, σ =4, p=26



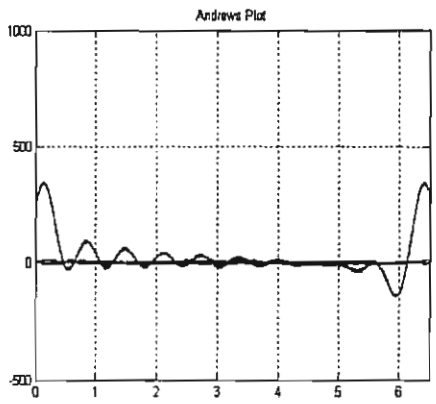
f) n = 40, c = 30, σ^2 =9, p=26

Figure 5.6(b): Andrews curve for (a) sample of size n with p variables from N($\boldsymbol{\mu}$, D), where $\boldsymbol{\mu}$=c(1,1,…,1), D=diag(σ^2, σ^2,…, σ^2) and (b) sample of size n with p variables from N (I,0).

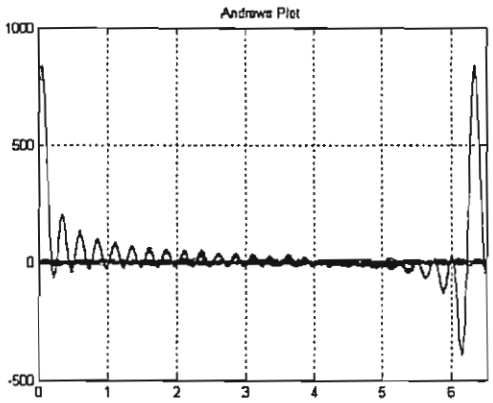
5.3.3 Varying variables.



a) n = 40, c = 25, σ =1, p=1



b)n = 40, c = 25, σ=1, p=20



c) n = 40, c = 25, σ =1, p=50

Figure5.7: Andrews curve for (a) sample of size n with p variables from N(μ, D), where μ=c(1,1,...,1), D=diag(σ^2, σ^2,..., σ^2) and (b) sample of size n with p variables from N (I,0).

5.3.4 Discussion

When vectors of variable are generated from two different populations, two distinct bands of curves are observed. As the difference between mean of both populations increases, clusters of curves become more well-separated. With a fixed variance, the higher the value of mean, the higher the number of "peak" will be observed as illustrated in figure 5.5. On the other hand, with a fixed mean, value of variance higher than 4 will be rendered useless as the curves become messy and a band of curve will be superior to the other unless then value of mean is more than 30 as illustrated in figure 5.6. Number of variable involve will

influence the number of peak formed as illustrated in figure 5.7. In the following section, we will fix c as 25, σ^2 as 1 and p as 26 for further study of discrimination.

5.4 Study of distribution properties

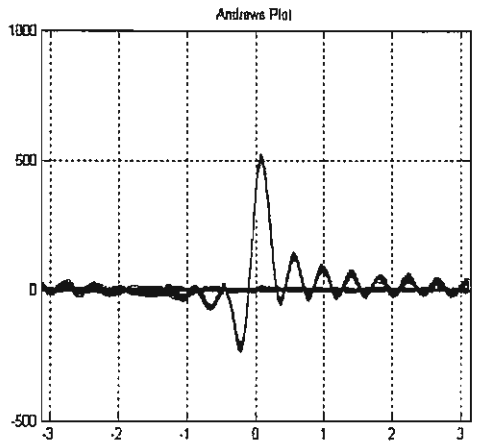


Figure 5.8: Andrews curve for sample of size 30 with 26 variables each from (a) N($\boldsymbol{\mu}$, D), where $\boldsymbol{\mu}$=c(1,…,1), D=diag(σ^2,…, σ^2), c = 25, σ^2 =2 and (b) N (I,0).

A closer look at Figure 5.8 shows that only at certain value of t, both groups tend to form distinct clusters. At other values of t, they tend to overlap with each another making it impossible for us to distinguish them. We are interested in obtaining the value of t where groups are distinct so that some statistical inference can be performed on all its 60 corresponding value of f(t) and define a new way of classification so that when given an unknown curve or a set of unknown curves we are able to classify it into a group correctly by considering value mean of f(t), $f_{mean}(t)$ of a sets of curves with known distribution. The following section describe a procedure on obtaining the values of ti, i =1, .., k.

5.4.1 Selection of value of t for each sample.

At certain value of t where both cluster of curves are distinct, a set of Andrews curves from the same population will group together closely enough to form a 'range' as illustrated in Figure 5.9.

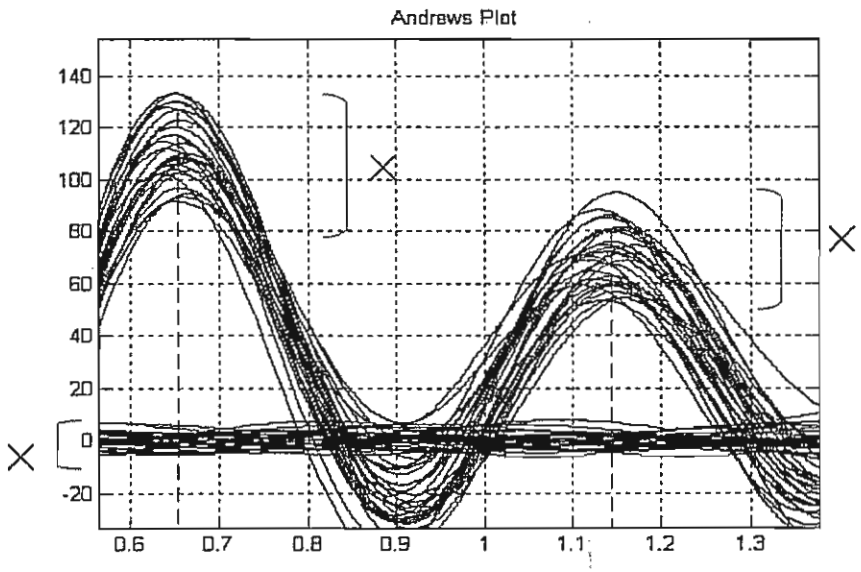


Figure 5.9 Range [marked as x] formed by a sets of 30 f(t) values at a selected value of t.

The steps of selecting the value of t are as described in previous section. These steps are summarized in figure 5.10. All values of f(t) at the selected value of t then used for the construction of an classification interval (confident interval).

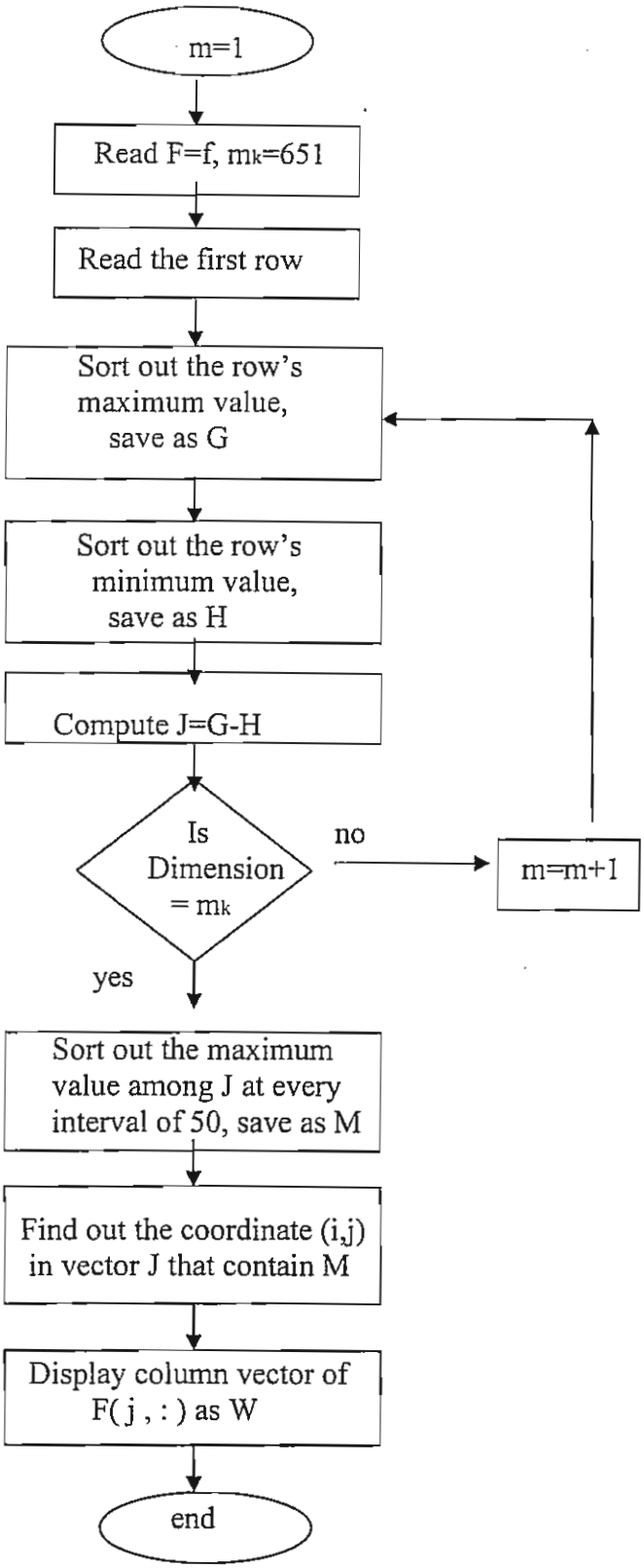


Figure5.10: Steps in obtaining the value of t. (f is a matrix of 651x60 formed using the procedure in figure 5.2).

5.4.2 Hypothesis Testing

Hypothesis testing was implemented on the 60 values at every value of t obtained in section 5.4 with the aim to test whether there is a significant difference between the means of both groups for every chosen value of t_i i=1,..., 12 i.e.

$$H_0 : \mu_x - \mu_y = 0 \text{ against } H_1 : \mu_x - \mu_y \neq 0$$

Since the probability distribution is known, the test statistics of z-test is used as follows:

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \qquad [5.1]$$

The null hypothesis will be rejected in favor of alternative hypothesis if the observed value of $|Z| \geq z_{\alpha/2}$ at an α significance level. With hypothesis testing, all we know is that the hypothesized value is a reasonable value for estimating the population parameter but then we don't know what it is likely to be. Therefore a 90% confident interval (90%CI) for $\mu_1 - \mu_2 = 0$ was constructed so that we can be 90% confident that the parameter resides within a certain range. The confidence level for $\mu_x - \mu_y$ can be defined as

$$\left[\quad (\bar{x} - \bar{y}) \pm Z_{\alpha/2} \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}} \quad \right] \qquad [5.2]$$

A large value contain in the confidence interval implies that two groups of sample are well separated whereas confidence interval containing zero value implies that both groups have the probability of overlapping which may later lead to the risk of misclassification. Therefore, these values of t will certainly be omitted for further study. When we are confident enough to say that both groups are well separated, a separate 90% confident interval for each of both groups are constructed at every chosen values of t. These confidence intervals may be used as a standard interval for future classification.

5.4.3 Confidence Interval of Each Cluster.

Given a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution N (μ, σ^2), we want to consider the closeness of $\overline{X}$ (the unbiased estimator of μ) to the unknown mean μ. For the probability 1-α, we can find a number $z_{\alpha/2}$ from statistical table so that

$$P\left[z_{\alpha/2} \leq \frac{(\overline{X} - \mu)}{(S/\sqrt{n})} \leq z_{\alpha/2} \right] = 1 - \alpha \qquad [5.3]$$

Therefore the random interval which includes the unknown mean μ with probability 1-α is

$$\left[\bar{x} - z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right), \bar{x} + z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \right] \qquad [5.4]$$

The results obtained are presented in table 5.1.

5.4.4 Discrimination Rule.

By constructing the confidence intervals at selected values of t we are hoping to derive a discrimination rule such that when a new unknown curve is drawn across the plot, two conclusions can be made:

a) If the f(t) value from a curve falls in the one of the earlier define range, then we conclude that the unknown curve belongs to a certain existing group.

b) If it doesn't fall in either one of the range, then we conclude that the unknown curve doesn't belong to any of the existing group.

The higher the number of time a curve falls within the range of an existing group at selected different value of t, the higher the probability we say it belongs to a certain population.

5.4.5 Result on Sampling Properties

	t_0	Ho (Null hypothesis)	Alternative hypothesis	Hypothesis Conclusion at 90%	90% Confidence interval for $\mu_1 - \mu_2 = 0$	90% Confident interval for $\mu_1 = C$	
						Sample one	Sample two
t_1	0.11	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Reject Ho	437.9487 441.0749	-2.0069 0.3491	437.6555 439.7102
t_2	0.66	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Reject Ho	111.3443 114.2355	-1.6086 0.3318	111.0798 113.2232
t_3	1.16	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Reject Ho	68.3419 71.2775	-1.0301 0.5765	68.3544 70.8113
t_4	1.63	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Reject Ho	49.4634 52.4422	-1.0371 0.9067	49.7591 52.0162
t_5	2.12	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Reject Ho	37.6161 40.6509	-0.4565 1.6976	38.6852 40.8229
t_6	2.60	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Reject Ho	28.9447 32.1772	-1.0626 1.3070	29.5838 31.7824
t_7	***3.08***	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Reject Ho	***21.4406 24.5758***	***-0.6752 1.4315***	***22.2255 24.5473***
t_8	***3.54***	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Reject Ho	***14.1639 17.0193***	***-1.3766 0.2237***	***13.8327 16.1976***
t_9	***4.35***	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Reject Ho	***-8.0139 -4.9884***	***-1.4108 0.5478***	***-8.0856 -5.7796***
t_{10}	***5.00***	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Reject Ho	***-19.6825 -16.6423***	***-0.3719 1.6656***	***-18.6437 -16.3874***
t_{11}	5.50	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Reject Ho	-50.8130 -47.8177	-0.9129 1.0712	-50.3582 -48.1143

Table 5.1: Results for hypothesis testing and confidence interval for a sample size of 30 obtained from N(0,1) and N(25,1) each at selected value of t.

5.5 Discussion

The simulation study in this paper was carried out to investigate the potential of using groups of curves to perform discrimination. Two sets of well separated normal distributed data with known probability distribution function were generated. Investigation on the consistency of the pattern of the curve was carried out using different combination of parameter on the data. Result show that the curves exhibit a rather consistent pattern throughout the study and suggest possibility of discrimination at certain value of t. However, discrimination is hard to be implemented when variance is large and it seems that in general the value of t will rely on the number of variable used as it affects the location of the widest separation (peaks) that determine the value of t. When the values of t are fixed, confidence interval can be formed and maybe be used as a standard classification range for future. However, we are still uncertain of how good is this discrimination rule when it is applied on the real data set. Thus, further studies will have to be carried out to verify this claim but for sure not for the time being in this project.

CHAPTER 6: CONCLUSION

6.1 Concluding Remarks

A positive reaction to a tuberculin test indicates the presence of MTB antibodies, but it cannot diagnose an active infection. Skin test and sputum test also produces many false negative results especially among AIDS patients and others who have weakened immune system. Therefore, chest x rays and other imaging studies is needed to support the above mention test. A great motivation in the work presented here is to produce a more objective method for analyzing x-rays images compared to the typical and rather subjective visual detection.

Exploratory Data Analysis (EDA) using graphical representation moves us in another direction i.e. from simplification of complex multivariate data to reflection of its inherent multidimensional nature. Statistical techniques are designed to be the best when stringent assumptions apply. However, these classical techniques can behave badly when the practical situation departs from the ideal described by such assumption. The techniques of EDA i.e. Andrews curve used here help us to cope with a set of data in a fairly informal way, guiding us toward a relatively easy and quick structure that can be used to improve visual interpretation. Besides that, it is capable to provide us with an extensive method for a detail study of a set of data and it emphasis on flexible probing of data before comparing them to any probabilistic model or implementing any relevant statistical methods.

While detection of MTB visually can be considered as being subjective, the Andrews curve together with wavelet transformation enables us to organize arrays of gray level intensity values graphically in a way that directs our attention to various unanticipated features of the data. Implementation of multivariate clustering analysis on the approximation wavelets coefficients show positive results of some natural clustering within them. The finding illustrates that Andrews curve allows the explication of variables to identify relevant MTB and non MTB region of the lung. In some cases, the regions may be defined as the

first stage of infection and secondary stage of infection. This definition relies heavily on the line profiles selection strategy and also on the patients' health profiles.

An extension to the context of discriminating two well known lung diseases i.e. MTB and LC based on the above procedures was presented. In this real case study, Andrews curve seem to be able to classify and discriminate both diseases. A crude discrimination rule was designed and its misclassification probability was then estimated by performing the 'one-in-one-out' procedure since the size of sampled data obtained is rather small. In all tested cases, our method exhibits a rather satisfactory performance with 95.8% of correct classification.

Finally, a simulation study on the potential of using groups of curves to perform discrimination under the assumption of normality was carried out. Sets of normal distributed simulated data with a larger sample size were generated to compare the effects of different parameter combinations on the data. It was found that Andrews curve is not much affected by varying value of mean however it can be misleading when value of variance increases. When two populations from well separated known probability distribution were generated, they are easily to compare because their properties can be summarized by their means and variances. These approaches suggest further studies of sampling distribution and classification based on confidence interval constructed at selected value of t.

6.2 Limitation and Further Studies.

It is worth noting that the methodology presented in this work do have some limitation. However, this is not a limitation of our methodology proposed but rather limitation of the situation (eg. time, cost and other constrains). The main limitation faced here concerns with the acquisition of numbers of x-rays for a more thorough and comprehensive study. As the clustering results rely greatly on the selection of line profiles, a sound knowledge of the anatomy of chest and x-rays radiology is essential to ensure an accurate result is produced. Here, in the

preliminary study of the x-rays, a medical expert on respiratory disease was consulted to provide a better understanding of the disease and to obtain health's profiles of the confirmed patients. Last but not least, the limitation is due to the availability of a digitizer. The procedures presented here can only be applicable on digital images. However, not many hospitals are equipped with a digitizer at the time being.

Further application and evaluation of graphical methods still need to be carried out in response to the current debates that preoccupy the medical sciences and digital image/signal processing. In this particular study for example, the inability of Andrews curve in discriminating a MTB patient with cavities and a lung cancer patient also warrants further study.

While the methods describe here has been tailored to the particular application at hand, the methods could be applied with suitable modification to design future trials having similar goals.

APPENDIX A

Table A.1: Approximation coefficients of 30 line profiles with 26 pixels obtained from a lung x-ray image of patient A.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4170.7	4109.4	3958.8	3647.2	3723.6	3595.1	3536.7	3622.1	3616	3786.6	3895.8	3959.2	3793.2	3418.7	3545.2
2	4212.7	4012.2	3993.3	3680.3	3711.6	3569.8	3511.4	3648.7	3590.7	3776.2	3938.8	3947.3	3841.1	3564.9	3533.9
3	4250.9	4118.1	4052.7	3733.8	3711.6	3618	3520	3674.5	3711.3	3779.8	3921.3	4050.6	3829.6	3549.1	3526.6
4	4232.2	4026.4	4015.4	3700.8	3706	3581	3509.2	3664.7	3629.8	3778.4	3938.8	3976.3	3844.2	3588.8	3525.6
5	4188	4073.6	3984.1	3669.4	3727.8	3585.9	3527.5	3627.3	3595.5	3773.8	3917.8	3963.1	3817.1	3476.6	3551.8
6	4161.3	4121.5	3917.9	3612.4	3701.2	3598.8	3541.5	3634.1	3653.5	3812.4	3874.5	3934.9	3769	3388.5	3514.9
7	4134.2	4165.6	3960.9	3635.7	3614.8	3589.9	3525.9	3649.6	3756.1	3856.5	3907.1	3966.7	3721.1	3399.9	3418.4
8	4091.3	4201.7	3941.9	3636.8	3620.3	3579.2	3488.6	3576.8	3785.8	3889.5	3913.8	3919.1	3696.8	3350.3	3398
9	4017	4193.7	3877	3609.8	3629.5	3586.9	3485.2	3608.4	3832.9	3915.2	3963.6	3862.4	3700.3	3240.3	3410.3
10	4021.7	4152.4	3934	3619.5	3618.6	3567.2	3470.3	3613.8	3822.2	3951.1	3983.3	3897.2	3668.5	3243.6	3318
11	4017.8	4123.2	3862.7	3595.2	3649.2	3531.1	3480.9	3605.4	3796.2	3917.1	3927	3824	3617	3267.3	3253.8
12	3953.7	4097.7	3850.4	3568.7	3710.1	3514.6	3489.6	3642.1	3747.5	3907.5	3915.1	3794.9	3555.9	3207.5	3215
13	3921.3	4058.9	3788.3	3581	3708	3478	3484.7	3672	3720.5	3917.6	3905.1	3802	3433.3	3150.8	3192.7
14	3956.8	4005.9	3792.8	3581.4	3642.1	3505.7	3499.6	3638.1	3759.9	3953.1	3869.2	3748.2	3452.9	3136.5	3125.9
15	3995.3	3970.5	3768.9	3584.4	3649.1	3506.2	3513.6	3603.9	3777.4	3943.6	3846	3714.2	3445.2	3126.5	3068.6
16	3969.8	3932.6	3781	3621.8	3632.7	3502.2	3519.6	3604.5	3797.5	3994	3842	3673.1	3450.6	3172.6	3047.1
17	3966.2	3940.3	3760.4	3588.6	3597	3454	3558.6	3655.1	3750.8	3964.8	3838.4	3603.5	3235.6	3151.7	3044.2
18	3952.2	3923.8	3741.7	3643.2	3597.2	3403.3	3546.8	3694.5	3687.5	3974.9	3859.7	3640.1	3249.2	3229.5	3026.5
19	3956.4	3939.6	3724.3	3662.6	3592	3378.5	3513.1	3745.7	3655.5	3963.3	3787	3605.5	3256	3243.6	3061.4
20	3969.4	3924.2	3752.3	3609.5	3534.6	3428.9	3563.1	3747.3	3637.2	3956	3812.1	3631.3	3310.4	3174.7	3115.3
21	3979.5	3897.5	3721.2	3591.7	3537	3416.2	3602.9	3734.1	3595.1	3952.5	3741.7	3625.4	3234.9	3111.7	3070.7
22	3954.6	3952.8	3709.4	3629.5	3512.1	3415.5	3619.7	3712.5	3578.6	3937.9	3725.9	3567	3177.3	3055.2	3090.5
23	3957	4006.2	3681.3	3630.1	3466.9	3520.8	3677.7	3695.9	3626.7	3860.1	3671	3428.5	3248.7	3051.5	3100.2
24	3983.9	3952	3657	3653.6	3488.9	3533.6	3717.3	3744.3	3696	3763.6	3615.6	3401.1	3224	3052.8	3154.6
25	4048.3	3929.8	3644.7	3669	3481.7	3604.4	3737	3810	3711.1	3720.4	3566.2	3352.5	3228.8	3096.3	3112.2
26	4069.6	3925.1	3646.8	3666.6	3477	3630.3	3736.7	3824.5	3707.3	3713.9	3552.6	3333.7	3240.8	3116.1	3097.2

Table A.1: Approximation coefficients of 30 line profiles with 26 pixels obtained from a lung x-ray image of patient A.(CONT)

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	3244.8	3788.7	3528.6	3404.1	3061.5	3070.8	2878.6	2853.8	2708	3306.3	2961.3	2741.1	2720.7	2573.9	2513.6
2	3279.6	3798.3	3614.8	3626	3070.8	3183.4	2885.1	2827.8	2718.6	3278.1	3006.3	2794.5	2655.4	2612.8	2479.2
3	3345.5	3829.1	3743.4	3745.6	3088.9	3334.2	2938.5	2884.4	2794.5	3370.4	3011.3	2794.3	2650.3	2614.3	2504.6
4	3306.3	3803.5	3672.4	3708.1	3073.1	3255.2	2905.1	2845.5	2746	3305.7	3011.5	2805.5	2640.4	2617.8	2482
5	3264.3	3809.9	3570	3495.6	3081.2	3116.8	2880.9	2832.5	2713.1	3285.7	2995	2761.4	2697.7	2601	2496.4
6	3218.8	3736	3489.9	3346.1	3011.8	3042.5	2876.2	2889.4	2701.8	3343.9	2906.3	2731.4	2725	2529.7	2529.6
7	3221.3	3709.7	3427.1	3192.4	3026	2921.1	2845.4	2859.1	2735.7	3326.5	2893.4	2785.9	2687.5	2539.9	2539.1
8	3184.6	3655.4	3359.7	3069.8	3162.3	2910.9	2890.9	2873.3	2776.5	3215.7	2968.1	2811.1	2778.4	2552	2514.4
9	3143.3	3601.5	3362.7	3058.9	3055.9	2890.4	2894.9	2871	2819.3	3239.4	2905.1	2738.7	2763.9	2517	2524.8
10	3161.6	3571.4	3355.2	3188.7	2987.6	2794.8	2898.8	2843.8	2790.2	3144.6	2887.6	2819.3	2725.4	2527.3	2469.7
11	3166	3569	3356.2	3333.2	2980.2	2765.7	2863.3	2847.7	2793.3	3123.9	2930.7	2814	2621.5	2543.9	2455.6
12	3108.9	3553.5	3385.7	3351.8	3088.7	2751.2	2867.5	2807.2	2745.8	3042.6	2814.5	2891.7	2527.6	2647.3	2466.8
13	3072.2	3461.2	3447	3273	3093.5	2662.3	2883.4	2833.8	2715.5	3027.9	2835.1	2953.3	2478.6	2750.5	2524.5
14	3037.1	3425.7	3527.9	3228.5	3062	2686.7	2866.2	2864	2715.7	3088.8	2953.1	2987.3	2452.9	2715.8	2527.3
15	2986.3	3359.6	3647	3300.5	3022.7	2788.9	2820.1	2853.7	2790.8	3103.3	2926.2	3002.4	2489.6	2634.9	2518.5
16	2976.8	3408.8	3610.1	3233.4	2985.9	2829.3	2841.1	2717.2	2850.1	3074.2	2979.8	2940.6	2503.8	2611.6	2518.4
17	2962	3490.5	3551.8	3245.9	2904.8	2855.5	2829.3	2679.6	2863.3	3047.5	3127.6	2869	2543.3	2586.3	2567.7
18	2989.7	3484.4	3550.9	3220.1	2892.7	2851	2870.1	2596.1	2842.1	3110.4	3085.7	2867	2581.9	2592.9	2643.7
19	3022.6	3479.8	3524.3	3246.4	2882.7	2833.4	2808.5	2617.7	2802.4	3250.1	2967.3	2826	2731.2	2647.5	2637.4
20	3000.4	3452.4	3470	3327.3	2902.6	2859.5	2822.1	2626	2696.6	3296.4	2867.2	2777.4	2658.7	2657.1	2657.2
21	3015.9	3472.8	3373.6	3339.9	2937	2850.5	2797.1	2680.8	2682.1	3255.3	2805.8	2783.5	2603.5	2626.6	2613.6
22	3044.8	3654.3	3338.2	3337	2951.3	2870.2	2802.9	2673.3	2667.3	3246.5	2798.5	2798	2568.4	2581.9	2597.5
23	2989.9	3690.8	3312.5	3368	2930.7	2934.2	2826.6	2706.4	2614.8	3291.4	2827.7	2847.8	2592.5	2540	2572.6
24	2963.7	3726.6	3249.1	3340	2955	2883.6	2792.4	2652.7	2630.8	3245.7	2825.4	2832.6	2565.8	2525.1	2548.4
25	2958.4	3658	3175.1	3413	2943.4	2787.4	2774.7	2679.5	2634.5	3276.5	2786	2829.2	2516.7	2572.2	2574.6
26	2959.4	3633.1	3157.9	3437	2934	2760.1	2776	2689.8	2634.7	3293.4	2776.3	2832.9	2503.6	2585.8	2587.6

Table A.2: Euclidean Distance Matrix of 30 line profiles obtained from patient A's x-ray image.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	533.06	1168.3	2136.8	2225.5	2658.3	2594.6	2008.5	1886.5	1225.4	1244.5	1759	2895.4	4056.9	4082.8
2	533.06	0	1108.4	2092.6	2144.7	2592.2	2550.4	1969	1738.6	1044.7	1083.7	1660.2	2832.1	4021.2	4028.7
3	1168.3	1108.4	0	1159.9	1135.4	1625	1689.1	1182	966.6	869.97	364.04	658.69	1779.4	2953.7	2981.6
4	2136.8	2092.6	1159.9	0	470.17	632.27	567.15	384.3	614.79	1377.7	1232	1178.7	1343.1	2064.2	2137.4
5	2225.5	2144.7	1135.4	470.17	0	664.61	847.69	743.7	700.62	1454.2	1163.8	941.53	1076.3	1947.2	2011.6
6	2658.3	2592.2	1625	632.27	664.61	0	543.52	910.55	1011	1932.5	1712.2	1460	1018.1	1570.5	1597.1
7	2594.6	2550.4	1689.1	567.15	847.69	543.52	0	647.37	995.94	1796	1750.3	1689.3	1519.2	1874.6	1943.4
8	2008.5	1969	1182	384.3	743.7	910.55	647.37	0	621.01	1214.4	1224	1354.8	1686.8	2356.6	2433.2
9	1886.5	1738.6	966.6	614.79	700.62	1011	995.94	621.01	0	1007.5	923.97	1102	1604.3	2487	2543.9
10	1225.4	1044.7	869.97	1377.7	1454.2	1932.5	1796	1214.4	1007.5	0	658.15	1338.6	2374.4	3352.5	3424.2
11	1244.5	1083.7	364.04	1232	1163.8	1712.2	1750.3	1224	923.97	658.15	0	742.71	1901	3039.8	3084.2
12	1759	1660.2	658.69	1178.7	941.53	1460	1689.3	1354.8	1102	1338.6	742.71	0	1300.9	2549	2564.1
13	2895.4	2832.1	1779.4	1343.1	1076.3	1018.1	1519.2	1686.8	1604.3	2374.4	1901	1300.9	0	1412.9	1359.2
14	4056.9	4021.2	2953.7	2064.2	1947.2	1570.5	1874.6	2356.6	2487	3352.5	3039.8	2549	1412.9	0	455.15
15	4082.8	4028.7	2981.6	2137.4	2011.6	1597.1	1943.4	2433.2	2543.9	3424.2	3084.2	2564.1	1359.2	455.15	0
16	4818.1	4751.9	3710.2	2764.2	2664.4	2231.9	2476.4	3025.4	3159.4	4048.1	3777.5	3302.9	2148.2	844.86	870.99
17	2256	2246.7	1288	591.8	747.52	678.87	801.23	840.32	1052.3	1746.5	1470.7	1231.1	1139.9	1929.4	1922.2
18	3092.9	3058.6	2010.8	1165.8	991.3	941.64	1159.8	1456.8	1539.7	2290.4	2037.5	1676.4	1107.4	1241.8	1444.4
19	3644.5	3660.4	2632.8	1649.5	1639.4	1238.7	1381.3	1874.2	2145.7	2967.3	2739.9	2375.6	1491.3	921.25	1083.3
20	5323.5	5241.1	4225.5	3230.8	3143.7	2691.8	2885.1	3464.2	3593.6	4494.1	4271.6	3839.4	2711.5	1440.2	1466.2
21	5832.9	5803.2	4752.4	3779.2	3718.4	3268.3	3451.6	4028.1	4209.9	5084.5	4834.1	4360.3	3190.7	1848.7	1885
22	6094.8	6015.4	4998.2	3987.7	3913.5	3461.9	3624.6	4210	4353.8	5244.4	5037.4	4609.3	3474.8	2154.6	2196.1
23	6494.1	6409.2	5389.9	4409.5	4315.3	3860.1	4055.4	4641	4761.2	5666	5435.7	4976.9	3804.9	2532.4	2527.5
24	6688.1	6605.2	5589.5	4578	4501	4061.6	4214.7	4797.1	4926.5	5812.4	5617.1	5194.8	4067.9	2745.1	2808.8
25	4245.6	4203.6	3201.7	2163.2	2171.1	1663.8	1803.7	2382.2	2606.5	3478.5	3284.4	2897.3	1876.6	780.94	847.07
26	5770.7	5703.9	4676.4	3671.6	3593.6	3167.1	3328.3	3904	4040.2	4918.5	4710.8	4291.7	3184.9	1842.2	1950.9
27	6224.8	6142.2	5143.5	4105.6	4040.5	3591.1	3718.8	4308.2	4451.8	5337.5	5168.6	4778.9	3684.5	2381.9	2456.1
28	7302.4	7224	6206.3	5205.1	5139.9	4679.8	4838.2	5426.4	5578.6	6464.1	6252.7	5801.6	4636.3	3320.1	3328.3
29	7407.6	7333.5	6319	5293.3	5222.8	4776.7	4909	5495.8	5658.8	6537.5	6352.2	5927.9	4792.9	3462.5	3507
30	7692.2	7620.1	6614.4	5572.3	5524.3	5060.6	5172.7	5768	5943.6	6819.8	6650.4	6236.3	5104.5	3760.9	3800.8

Table A.2: Euclidean Distance Matrix of 30 line profiles obtained from patient A's x-ray image. (CONT)

16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
4818.1	2256	3092.9	3644.5	5323.5	5832.9	6094.8	6494.1	6688.1	4245.6	5770.7	6224.8	7302.4	7407.6	7692.2
4751.9	2246.7	3058.6	3660.4	5241.1	5803.2	6015.4	6409.2	6605.2	4203.6	5703.9	6142.2	7224	7333.5	7620.1
3710.2	1288	2010.8	2632.8	4225.5	4752.4	4998.2	5389.9	5589.5	3201.7	4676.4	5143.5	6206.3	6319	6614.4
2764.2	591.8	1165.8	1649.5	3230.8	3779.2	3987.7	4409.5	4578	2163.2	3671.6	4105.6	5205.1	5293.3	5572.3
2664.4	747.52	991.3	1639.4	3143.7	3718.4	3913.5	4315.3	4501	2171.1	3593.6	4040.5	5139.9	5222.8	5524.3
2231.9	678.87	941.64	1238.7	2691.8	3268.3	3461.9	3860.1	4061.6	1663.8	3167.1	3591.1	4679.8	4776.7	5060.6
2476.4	801.23	1159.8	1381.3	2885.1	3451.6	3624.6	4055.4	4214.7	1803.7	3328.3	3718.8	4838.2	4909	5172.7
3025.4	840.32	1456.8	1874.2	3464.2	4028.1	4210	4641	4797.1	2382.2	3904	4308.2	5426.4	5495.8	5768
3159.4	1052.3	1539.7	2145.7	3593.6	4209.9	4353.8	4761.2	4926.5	2606.5	4040.2	4451.8	5578.6	5658.8	5943.6
4048.1	1746.5	2290.4	2967.3	4494.1	5084.5	5244.4	5666	5812.4	3478.5	4918.5	5337.5	6464.1	6537.5	6819.8
3777.5	1470.7	2037.5	2739.9	4271.6	4834.1	5037.4	5435.7	5617.1	3284.4	4710.8	5168.6	6252.7	6352.2	6650.4
3302.9	1231.1	1676.4	2375.6	3839.4	4360.3	4609.3	4976.9	5194.8	2897.3	4291.7	4778.9	5801.6	5927.9	6236.3
2148.2	1139.9	1107.4	1491.3	2711.5	3190.7	3474.8	3804.9	4067.9	1876.6	3184.9	3684.5	4636.3	4792.9	5104.5
844.86	1929.4	1241.8	921.25	1440.2	1848.7	2154.6	2532.4	2745.1	780.94	1842.2	2381.9	3320.1	3462.5	3760.9
870.99	1922.2	1444.4	1083.3	1466.2	1885	2196.1	2527.5	2808.8	847.07	1950.9	2456.1	3328.3	3507	3800.8
0	2640.4	1940	1444.8	670.01	1149.3	1356.1	1720.6	1969.8	847.61	1135.7	1621.8	2529.4	2673.4	2973.8
2640.4	0	1293.3	1539.8	3157.9	3633.1	3918.5	4318.4	4530.6	2064	3621.7	4081.4	5107.5	5236.5	5509.8
1940	1293.3	0	1064.6	2410.6	2908.7	3139.7	3551.9	3703.2	1535.2	2779.6	3264.5	4362.1	4426.7	4732.6
1444.8	1539.8	1064.6	0	1916.4	2314.2	2620.4	3029.9	3225	977.8	2338.5	2763.1	3826.6	3880.4	4177.6
670.01	3157.9	2410.6	1916.4	0	946.12	821.54	1201	1441.7	1258.1	713.53	1017.5	2055.1	2122.1	2437.7
1149.3	3633.1	2908.7	2314.2	946.12	0	800.94	1062.9	1241.7	1706.7	772.08	1147.9	1643.4	1832.8	2080.4
1356.1	3918.5	3139.7	2620.4	821.54	800.94	0	541.23	659.9	1939	521.9	460.1	1280.9	1348.1	1645.3
1720.6	4318.4	3551.9	3029.9	1201	1062.9	541.23	0	585.14	2358.2	965.19	698.19	993.77	1075.6	1398
1969.8	4530.6	3703.2	3225	1441.7	1241.7	659.9	585.14	0	2548.3	968.36	661.03	829.29	849.48	1117.9
847.61	2064	1535.2	977.8	1258.1	1706.7	1939	2358.2	2548.3	0	1694	2122.3	3097.3	3229.5	3480.7
1135.7	3621.7	2779.6	2338.5	713.53	772.08	521.9	965.19	968.36	1694	0	726.23	1650.9	1723.6	2005.3
1621.8	4081.4	3264.5	2763.1	1017.5	1147.9	460.1	698.19	661.03	2122.3	726.23	0	1365.4	1239.7	1557.4
2529.4	5107.5	4362.1	3826.6	2055.1	1643.4	1280.9	993.77	829.29	3097.3	1650.9	1365.4	0	682.61	691.61
2673.4	5236.5	4426.7	3880.4	2122.1	1832.8	1348.1	1075.6	849.48	3229.5	1723.6	1239.7	682.61	0	471.18
2973.8	5509.8	4732.6	4177.6	2437.7	2080.4	1645.3	1398	1117.9	3480.7	2005.3	1557.4	691.61	471.18	0

Appendix B

B1: Multivariate Normal Distribution Property I:

The following are true for a random vector **x** having a multivariate normal distribution:

I. Linear combination of components of **x** are normally distributed;
II. All subsets of the components of **x** have a (multivariate) normal distribution;
III. Zero covariance implies that the corresponding components are independently distributed;
IV. The conditional distributions of the components are (multivariate) normal.

B2: Multivariate Normal Distribution Property II:

These statements are then reproduced mathematically in the results that follow which are also the properties of multivariate normal distribution.
For all multivariate normal random vector x,

I. If $x \sim Np(\mu, \sum)$ of rank p, so that $\sum^{-1}$ exists then $(x-\mu)^T\sum^{-1}(x-\mu) \sim X^2(p)$;
II. If $x \sim Np(\mu, \sum)$, $a^Tx = a_1x_1 + a_2x_2 + \ldots + a_px_p$, then $a^Tx \sim Np(a^T\mu, a^T\sum a)$. Also if $a^Tx \sim Np(a^T\mu, a^T\sum a)$ for very a then $x \sim Np(\mu, \sum)$;
III. If $x \sim Np(\mu, \sum)$, $A \in M_{pq}$ then $A^Tx+b \sim Nq(A^T\mu+b, A^T\sum A)$. Also x+d, where d is a vector of constants $\sim Np(\mu+d, \sum)$;
IV. All subsets of x are normally distributed. If we respectively partitioned x, its mean vector μ and its covariance matrix $\sum$ as

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \sum = \begin{bmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{bmatrix}$$

then $x_1 \sim Np(\mu_1, \sum_{11})$,

- If x_1 and x_2 are independent, then $\text{cov}(x_1, x_2) = 0$.
- If $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N_{q1+q2}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{bmatrix}\right)$ then x_1 and x_2 are independent if and only if $\sum_{12} = 0$.
- If x_1 and x_2 are independent and are distributed as $Nq_1(\mu_1, \sum_{11})$, $Nq_1(\mu_2, \sum_{22})$ respectively then $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N_{q1+q2}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sum_{11} & 0 \\ 0 & \sum_{22} \end{bmatrix}\right)$

B3: Canonical Transformation

We require $A\Sigma_2 A^T \rightarrow I$ and $A\Sigma_1 A^T \rightarrow D$.
Since Σ_2 is symmetric and non singular, we can write it in spectral form:

$\Sigma_2 = Q_2 A_2 Q_2^T$, where Q_2 is the matrix of eigenvectors and
A_2 is the matrix of eigenvalue.

Let $P = A_2^{-1/2} Q_2 T$, then $P\Sigma_2 P^T = I$ with a P non singular.

Let $C = P\Sigma_1 P^T$, clearly C is symmetric and has it's own spectral decomposition, say

$C = QcAcQc^T$ or $Qc^T CQc = Ac$.

Finally, by letting $A = Qc^T P$
Thus, $A\Sigma_2 A^T = Qc^T P\Sigma_2 P^T Qc = Qc^T IQc = I$
And $A\Sigma_1 A^T = Qc^T P\Sigma_1 P^T Qc = Qc^T CQc = Ac$
The mean in the **x*** space can be obtained as follows:

$\boldsymbol{\mu}_1^* = \mathbf{A}\boldsymbol{\mu}_1 + \mathbf{b} = \boldsymbol{\mu}$
$\boldsymbol{\mu}_2^* = \mathbf{A}\boldsymbol{\mu}_2 + \mathbf{b} = 0$
Hence $\mathbf{b} = -\mathbf{A}\boldsymbol{\mu}_2$
And $\boldsymbol{\mu} = \mathbf{A}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

Appendix C: Matlab Programming

C1: CLUSTERING

```
disp('Single');
D=pdist(c1A','euclid');% Take the matrix transpose since we want to cluster the row vector
S = squareform(D);
Z1 = linkage(D,'single'); % calculate dendrogram value using Single Linkage method
dendrogram(Z1),title('Single Linkage Method');

disp('Complete');
Z2=linkage(D,'complete');  % calculate dendrogram value using Complete Linkage method
figure;
dendrogram(Z2),title('Complete Linkage Method');

disp('Average');
Z3=linkage(D,'average');  % calculate dendrogram value using Average Linkage method
figure;
dendrogram(Z3), title('Average Method');

disp('Centroid');
Z4=linkage(D,'centroid');  % calculate dendrogram value using Centroid method
figure;
dendrogram(Z4),title('Centroid Method');

disp('Ward');
Z5=linkage(D,'ward');  % calculate dendrogram value using Ward method
figure;
dendrogram(Z5),title('Ward Method');

T1=cluster(Z1,2);
C1=find(T1==1);
D1=find(T1==2);
```

C2: SELECTION OF T VALUE AND COMPUTATION OF EDD.

```
%PROGRAM TO SELECT THE VALUE OF T BY CALCULATING THE E.D.D. VALUE.AT
EVERY INTERVAL OF 0.5 UNIT OF T.

FF=[F']%matrix of 651x60
fsmax = max(FF(6:14,:))          %max(FF') treats the columns of A as vectors, returning a row
vector
fsmin = min(FF(15:23,:))         %containing the maximum element from each column.
diff=fsmin-fsmax                 %matrix of 1x651

g=1, h=50;                       %gh to dine the range of t to be used in calculation
fboth=[fsmax(:,g:h) ;fsmin(:,g:h) ;diff(:,g:h)] %matrix 3x651
maxdif1=max(diff(:,g:h))
[i,j,v] = find(fboth==maxdif1)
k1=g-1+j
rowj1=F(k1,:)'                   % g+jth column and all 10 row,

g2=51, h2=100;                   %gh to dine the range of t to be used in calculation
fboth2=[fsmax(:,g2:h2) ;fsmin(:,g2:h2) ;diff(:,g2:h2)] %matrix 3x651
maxdif2=max(diff(:,g2:h2))
```

```
[i2,j2,v2] = find(fboth2==maxdif2)
k2=g2-1+j2
rowj2=F(k2,:)'                          % g+jth row and all 60 column,

g3=101, h3=150;                         %gh to dine the range of t to be used in calculation
fboth3=[fsmax(:,g3:h3) ;fsmin(:,g3:h3) ;diff(:,g3:h3)] %matrix 3x651
maxdif3=max(diff(:,g3:h3))
[i3,j3,v3] = find(fboth3==maxdif3)
k3=g3-1+j3
rowj3=F(k3,:)' % g+jth row and all 60 column,

g4=151, h4=200;                         %gh to dine the range of t to be used in calculation
fboth4=[fsmax(:,g4:h4) ;fsmin(:,g4:h4) ;diff(:,g4:h4)] %matrix 3x651
maxdif4=max(diff(:,g4:h4))
[i4,j4,v4] = find(fboth4==maxdif4)
k4=g4-1+j4
rowj4=F(k4,:)' % g+jth row and all 60 column,

g5=201, h5=250;                         %gh to dine the range of t to be used in calculation
fboth5=[fsmax(:,g5:h5) ;fsmin(:,g5:h5) ;diff(:,g5:h5)] %matrix 3x651
maxdif5=max(diff(:,g5:h5))
[i5,j5,v5] = find(fboth5==maxdif5)
k5=g5-1+j5
rowj5=F(k5,:)' % g+jth row and all 60 column,

g6=251, h6=300;                         %gh to dine the range of t to be used in calculation
fboth6=[fsmax(:,g6:h6) ;fsmin(:,g6:h6) ;diff(:,g6:h6)] %matrix 3x651
maxdif6=max(diff(:,g6:h6))
[i6,j6,v6] = find(fboth6==maxdif6)
k6=g6-1+j6
rowj6=F(k6,:)' % g+jth row and all 60 column,

alldif=[maxdif1 maxdif2 maxdif3 maxdif4 maxdif5 maxdif6]
kk=[k1 k2 k3 k4 k5 k6]
ROWS=[rowj1 rowj2 rowj3 rowj4 rowj5 rowj6
```

C3: COMPUTATION OF MEDIAN AND MINIMUM DISTANCE

```
ROWS=[rowj1 rowj2 rowj3 rowj4 rowj5 rowj6];

medH=median([ ROWS(1,:); ROWS(3,:); ROWS(4,:); ROWS(2,:)])
medTB=median([ ROWS(5,:); ROWS(6,:); ROWS(7,:); ROWS(8,:); ROWS(9,:); ROWS(10,:);
ROWS(11,:); ROWS(12,:); ROWS(13,:); ROWS(14,:);])
medLC=median([ ROWS(15,:);    ROWS(16,:); ROWS(17,:); ROWS(18,:); ROWS(19,:);
ROWS(20,:); ROWS(21,:); ROWS(22,:); ROWS(23,:); ])

medALL=[medH; medTB; medLC]
dist1=sqrt((ROWS(23,:)-medH).^2)
dist2=sqrt((ROWS(23,:)-medTB).^2)
dist3=sqrt((ROWS(23,:)-medLC).^2)
dist=[dist1; dist2; dist3;]
```

```
dist1=sqrt(dist1*dist1);
dist2=sqrt(dist2*dist2);
dist3=sqrt(dist3*dist3);
mindif=min([dist1 dist2 dist3])
```

C4: GENERATION OF NORMAL DISTRIBUTED DATA

```
%this program is to generate normal random numbers from N(miu,sigma)

k=30;                    %to define k variables
p=40;                    % to define p observations
R =[p,k];                % Preallocate matrix
R2=[p,k];
%FIRST POPULATION----------------------------------------------
for n = 1:p;
  for m = 1:k;
    R(n,m) = normrnd(0,1,1,1);
  end
end

%SECOND POPULATION----------------------------------------------
for n = 1:p;
  for m = 1:k;
    R2(n,m) =normrnd(30,2,1,1);
  end
end
%--------------------------------------------------------------
%transformation into andrews curve.
a=1;
for t=-pi:0.01:pi;    %coefficient values for andrew's plot
   A=[1/sqrt(2); sin(t); cos(t); sin(2*t); cos(2*t); sin(3*t); cos(3*t); sin(4*t); cos(4*t);
sin(5*t); cos(5*t); sin(6*t); cos(6*t); sin(7*t); cos(7*t);
     sin(8*t); cos(8*t); sin(9*t); cos(9*t); sin(10*t); cos(10*t); sin(11*t); cos(11*t);
sin(12*t); cos(12*t);
     sin(13*t); cos(13*t); sin(14*t); cos(14*t); sin(15*t); cos(15*t); sin(16*t); cos(16*t);
sin(17*t); cos(17*t);
     sin(18*t); cos(18*t); sin(19*t); cos(19*t); sin(20*t); cos(20*t); sin(21*t); cos(21*t);
sin(22*t); cos(22*t);
     sin(23*t); cos(23*t); sin(24*t); cos(24*t); sin(25*t)];

  % choose a subset of coefficients vector A that has size k
  A1=A(1:k,:);
  H=R';
  H2=R2';
 % calculate andrews curve values
  F(a,:)=A1'*[H(:,1) H(:,2) H(:,3) H(:,4) H(:,5) H(:,6) H(:,7) H(:,8) H(:,9) H(:,10)
H(:,11) H(:,12) H(:,13) H(:,14) H(:,15) H(:,16) H(:,17) H(:,18) H(:,19) H(:,20) H(:,21)
H(:,22) H(:,23) H(:,24) H(:,25) H(:,26) H(:,27) H(:,28) H(:,29) H(:,30) H(:,31) H(:,32)
H(:,33) H(:,34) H(:,35) H(:,36) H(:,37) H(:,38) H(:,39) H(:,40)];
```

```
 F2(a,:)=A1'*[H2(:,1) H2(:,2) H2(:,3) H2(:,4) H2(:,5) H2(:,6) H2(:,7) H2(:,8) H2(:,9)
H2(:,10) H2(:,11) H2(:,12) H2(:,13) H2(:,14) H2(:,15) H2(:,16) H2(:,17) H2(:,18)
H2(:,19) H2(:,20) H2(:,21) H2(:,22) H2(:,23) H2(:,24) H2(:,25) H2(:,26) H2(:,27)
H2(:,28) H2(:,29) H2(:,30) H2(:,31) H2(:,32) H2(:,33) H2(:,34) H2(:,35) H(:,36) H(:,37)
H(:,38) H(:,39) H(:,40)];

   % define andrews curve x-axis
   T1(a)=t;
   a=a+1;
end
% plotting andrews plot and label its legend

plot(T1,F(:,1),'b',T1,F(:,2),'b',T1,F(:,3),'b',T1,F(:,4),'b',T1,F(:,5),'b',T1,F(:,6),'b',T1,F(:,7),'
b',T1,F(:,8),'b',T1,F(:,9),'b',T1,F(:,10),'b')
hold on;
plot(T1,F(:,11),'b',T1,F(:,12),'b',T1,F(:,13),'b',T1,F(:,14),'b',T1,F(:,15),'b')%,T1,F(:,16),'b',
T1,F(:,17),'b',T1,F(:,18),'b',T1,F(:,19),'b',T1,F(:,20),'b')
hold on;
plot(T1,F(:,21),'b',T1,F(:,22),'b',T1,F(:,23),'b',T1,F(:,24),'b',T1,F(:,25),'b',T1,F(:,26),'b',T1
,F(:,27),'b',T1,F(:,28),'b',T1,F(:,29),'b',T1,F(:,30),'b'),
hold on;
plot(T1,F2(:,1),'k',T1,F2(:,2),'k',T1,F2(:,3),'k',T1,F2(:,4),'k',T1,F2(:,5),'k',T1,F2(:,6),'k',T1
,F2(:,7),'k',T1,F2(:,8),'k',T1,F2(:,9),'k',T1,F2(:,10),'k')
hold on;
plot(T1,F2(:,11),'k',T1,F2(:,12),'k',T1,F2(:,13),'k',T1,F2(:,14),'k',T1,F2(:,15),'k',T1,F2(:,1
6),'k',T1,F2(:,17),'k',T1,F2(:,18),'k',T1,F2(:,19),'k',T1,F2(:,20),'k')
hold on;
plot(T1,F2(:,21),'k',T1,F2(:,22),'k',T1,F2(:,23),'k',T1,F2(:,24),'k',T1,F2(:,25),'k',T1,F2(:,2
6),'k',T1,F2(:,27),'k',T1,F2(:,28),'k',T1,F2(:,29),'k',T1,F2(:,30),'k'),

title('Andrews Plot'),axis([0 0.5 -500 1000]);
grid on;
figure;
plot(T1,F(:,1),'b',T1,F(:,2),'b',T1,F(:,3),'b',T1,F(:,4),'b',T1,F(:,5),'b',T1,F(:,6),'b',T1,F(:,7),'
b',T1,F(:,8),'b',T1,F(:,9),'b',T1,F(:,10),'b')
hold on;
plot(T1,F(:,11),'b',T1,F(:,12),'b',T1,F(:,13),'b',T1,F(:,14),'b',T1,F(:,15),'b',T1,F(:,16),'b',T1
,F(:,17),'b',T1,F(:,18),'b',T1,F(:,19),'b',T1,F(:,20),'b')
hold on;
plot(T1,F(:,21),'b',T1,F(:,22),'b',T1,F(:,23),'b',T1,F(:,24),'b',T1,F(:,25),'b',T1,F(:,26),'b',T1
,F(:,27),'b',T1,F(:,28),'b',T1,F(:,29),'b',T1,F(:,30),'b'),
hold on;
plot(T1,F2(:,1),'k',T1,F2(:,2),'k',T1,F2(:,3),'k',T1,F2(:,4),'k',T1,F2(:,5),'k',T1,F2(:,6),'k',T1
,F2(:,7),'k',T1,F2(:,8),'k',T1,F2(:,9),'k',T1,F2(:,10),'k')
hold on;
plot(T1,F2(:,11),'k',T1,F2(:,12),'k',T1,F2(:,13),'k',T1,F2(:,14),'k',T1,F2(:,15),'k',T1,F2(:,1
6),'k',T1,F2(:,17),'k',T1,F2(:,18),'k',T1,F2(:,19),'k',T1,F2(:,20),'k')
hold on;
```

```
plot(T1,F2(:,21),'k',T1,F2(:,22),'k',T1,F2(:,23),'k',T1,F2(:,24),'k',T1,F2(:,25),'k',T1,F2(:,2
6),'k',T1,F2(:,27),'k',T1,F2(:,28),'k',T1,F2(:,29),'k',T1,F2(:,30),'k'),

title('Andrews Plot'),axis([-pi pi -500 1000]);
grid on;
```

REFERENCES

Adjei, A.A., Marks, R.S., bonner, J.A.(1999). Current guidelines for the management of small cell lung cancer. Mayo Cllinic Proceedings, 74(8),809-816.

Anderberg, M.R.(1973) *Cluster Analysis for Applications.* New York: Academic Press.

Andrews, D. F. (1972). Plot of high dimensional data. *Biometrics* 28: 125-136.

Awcock, G. J. and R. Thomas.(1995) *Applied Image Processing.* London: Macmillan.

Bertin, T. (1967). *Semiologie Graphique.* Gauthier-Villars: Paris.

Bhattacharyya, G. K., and R. A. Johnson. (1977). *Statistical Concepts and Methods.* New York: John Wiley.

Bruntz, S.M., Cleveland, W.S., Kleiner, B., and Warner, J.L. (1974). The dependence of ambient ozone on solar radiation, wind, temperature and mixing height. *Proc. Symp. Atmos. Diffus. Air pollution Am. Meterol. Soc., 125-128.*

Bunn, P.A. and Kelly, K.(2000). New combination in the Treatment of lung cancer: A time for optimism. Chest, 117(4) Supplement 1, 1385-1435.

Castleman. K. R.(1996).*Digital Image Processing.* New Jersey: Prentice Hall.

Cattell, R. B., and Coulter, M. A. (1966). Principles of behavioural taxonomy and the mathematical basis of the taxanome computer program. *Br. J. Math. Stat. Psychol.*, **19**, 237-269

Chambers, J.M., W.S. Cleverland, B. Kleiner and P.A. Tukey. (1983).*Graphical Methods For Data Analysis.* New Jersey: Bell.

Chatfield, C., and A. J. Collins.(1980). *Introduction to Multivariate Analysis.* London: Chapman and Hall.

Cheng, S,C, and Y.M. Huang [2003]A novel approach to diagnose diabetes based on the fractal characteristics of retinal images. *Information Technology in Biomedicine.***7**, 163-170

Chernoff, H. (1973). Using faces to represent points in K-dimensional space graphically. *J. Amer. Statist. Assoc.* **68,** 361-368.

Chernoff,, H., and Rivzi, H. m. (1975). Effect on classification error of random permutations of features in representing multivariate data by faces. *J. Am. Stat. Assoc.*, **70,** 548-554.

Daubechies, I.(1992).*Ten lectures on wavelets*. Vermont: Capital City Press.

Dillion, W. R. (1984). *Multivariate Analysis Methods and Applications* .New York: John Wiley.

Du Toit, S. H. C., A. G. W. Steyn, and R. H. Stumpf. (1986). *Graphical Exploratory Data Analysis*. New York: Springer-Verlag.

Everitt, B. S. (1978). *Graphical Techniques for Multivariate Data.* London: Heinemann.

Everitt, B.S., and G. Dunn.(2001). *Applied Multivariate Data Analysis.* London: Arnold.

Fienberg, S. E. (1979). Graphical methods in statistics. *Am. Stat.*, **33**, 165-178.

Gao, X.B , Ji, H.B., Shen A.D. and Wang, G.(2002). An interative segmentation method for medical images. *Signal Processing 2002 6th International Conference*, **1**,580-583.

Gnanadesikan, R.(1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: John Wiley.

Gonzalez, R. C., and R.E. Woods. (1992). *Digital Image Processing.* United States: Addison-Wesley.

Hartigan, J. A. (1975). Printer graphics for clustering. *J. of Statist. Computation and Simulation.* **4,** 187-213.

Hirano, S., Sun X.G.and Tsumoto, S.(2002). Dealing with multiple types of expert knowledge in medical image segmentation: a rough sets style approach. *Fuzzy System.* **2,** 884-889.

Hogg R.V. and Tanis, E.A. (2001)*Probability and Statistical Inference*. New Jersey: Prentice Hall.

Johnson, R. A., and D. W. Wichern. (1998). *Applied Mutivariate Statistical Analysis*. New Jersey: Prentice Hall.

Mallat, S. (1989), A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Pattern Anal. and Machine Intell.*, **11**, 674-693.

Mardia, K. V., J. T. Kent, and J. M. Bibby. (1979). *Multivariate Analysis.* London: Academic Press.

Rubin, P. and Williams, J.P.,(2001). *Clinical Oncology: A Multidisiplinary Approach for physician and students.* Pennsylvania: W.B. Saunders Company.

Russ, J. C. (1995). The image processing handbook. Florida: CRC.

Seber, G.A. F. (1984). *Multivariate Analysis.* New York: John Wiley.

Siegel, J. H., Goldwyn, R. M., and Friedman, H. P. (1971). Pattern and process of the evolution of human septic shock. *Surgery,* **70,** 232-245.

Tweed T and Miguet S. (2002). Automatic detection of region of interest in mammographies based on a combined analysis of texture and histogram. *Pattern Recognition,* **2**, 448-452.

Wakimoto, K., and Taguri, M. (1978). Consellation graphical method for representing multidimensional data. *Ann. Inst. Stat. Math.,* **30,** 97-104.

Walker, J.S. (1999). *A Primer on Wavelets and their Scientific Applications.* Florida: CRC Press.

Wang, P.C.C. (ed.) (1978). *Graphical Representation of Multivariate Data.* New York:Academic Press.

Welsch, R. E. (1976). Graphics for data Analysis. *Comput. Graphics,* **2,** 31-37.