THE UNIVERSITY OF CALGARY

Cell Shape Analysis, a Statistical Approach

by

Alberto Nettel-Aguirre

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICS AND STATISTICS

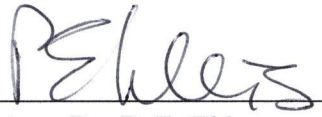
CALGARY, ALBERTA

July, 2005

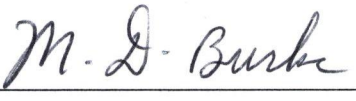
THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

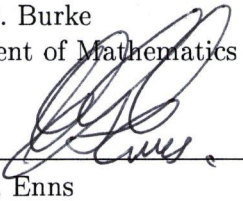
The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Cell Shape Analysis, a Statistical Approach" submitted by Alberto Nettel-Aguirre in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY.



Supervisor, Dr. P. F. Ehlers
Department of Mathematics and Statistics

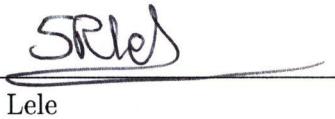


Dr. M. D. Burke
Department of Mathematics and Statistics



Dr. E. G. Enns
Department of Mathematics and Statistics

Dr. W. W. Zwirner
Department of Applied Psychology



Dr. S. Lele
University of Alberta

July 4, 2005

Date

Abstract

This work presents statistical techniques to aid in the diagnosis of cancerous tissue by studying the shape of nuclei of cells quantitatively. A brief summary of stochastic geometry approaches is discussed. Techniques from functional data analysis are applied to the X, Y coordinates of two-dimensional profiles of nuclei. An exploratory analysis of variability via functional principal components and functional linear discriminant analysis is presented. The profiles are described by their Fourier series expansions and characterised by the use of principal differential analysis. Confidence-like intervals for residual functions arising from such characterisations are provided.

Acknowledgements

Thanks to Dr Ehlers for all the discussions, the chats, for encouraging me into looking at all areas that were of interest to me in Statistics and his support for attending conferences (as did Ernest) and workshops in such areas; for his trust in academic and consulting matters, for being an inspiration and an example of what a Statistician is; for giving me the chance to share the challenges of consulting and for making *StatCaR* a place of fun and learning and a great environment for working.

To the Mathematics and Statistics department at U of C which became my home away from home. To Joanne for always looking out for us grad students and being our second mom. To Marguerite for trusting and encouraging me to TA for Stats as well as Amat and for always being so accommodating when it came to conferences and workshops. To Joanne L., Sue, Jan, Marc, Betty, Laurie, Maria, Shelley and all the staff that were always good friends and that kept the camaraderie feeling alive.

To Joan, Jim, Nancy and Kris for being a source of inspiration for teaching and for fun and interesting discussions. A special thanks to Nancy for her support in the final stages of this thesis.

Many thanks to Eva Vedel Jensen and Asger Hobolth from the University of Aarhus in Denmark for very kindly providing the nuclei profiles which they shared with me even though they were planning on using them for analysis of their own in the (at the time) future.

To Canada my adoptive home and country that is blind to ethnicity or social status and that has given me so much.

Most specially to my parents: Gloria and Vladimiro who brought me up to believe

in myself, to set goals and achieve them. To them for so inconspicuously planting the seed of love for learning and for not only encouraging but also nurturing the search for excellence. For the great education they provided, be it great schools or great home environment regardless of the efforts or sacrifices needed.

To my (older) brothers, Erick and Vladimiro, for setting up an example, for their support in "playing school" for so long and for letting me know they are as proud of me as I am of them.

To Hector, Tota, Gaby and Laura for allowing me to rip away from them what I consider to be my most precious, proud and loved achievement: keeping Luz for myself.

And, obviously... to Luz, for being the source of light throughout my many dark hours, frustrated moments and sad instances; for showing me the best years of my life, for her sweet loving care and support; for venturing into starting a new life in a new country by my side and for staying by my side ever since; for the infinitely often statistical discussions, for keeping our home alive and well in the months of writing this thesis and making my life so much simpler.

But most of all, for being who she is and being so by my side, with me,... for me.

Table of Contents

List of Tables

List of Figures

Introduction

"Statisticians get to play in everybody's backyard" paraphrases Tukey's quote about our science. It is true to some extent that Statistics lets us immerse ourselves into many different subjects. Given this appealing property of statistical thought, we tend to find ourselves involved in making decisions or calls about an issue at hand that is of interest to the community. The use of statistics in the medical sciences is thought of being concentrated in clinical trials and pharmaceutical discoveries. Few are the people that think of it in other aspects of medicine. One of the main applications of Statistics is that of classifying into groups, given information on characteristics of observations.

Another closely related application is that of learning and discovering relationships and/or features that allow us to characterise objects and therefore classify them into groups. The area of supervised learning or supervised clustering is an ever evolving area in statistics and it is an area that is easily related to medical applications.

Through my life, I have seen several dear and loved family members pass away to cancer and heard the remark "Had it been caught in time, there would have been a better chance of survival" and hence I have developed an interest in merging my training as a statistician and my desire to aid in the detection of cancer. "There has to be some use for statistics in cancer detection!!" has been my battle flag for these years, and luckily enough, there is a way I found exciting enough to immerse myself into it when it comes to cancer detection and the use of statistics.

The shape of cancer cells and mainly of their nuclei seemed to be an identifier

for such type of tissues. Statistical shape analysis is the connection of both of these worlds.

Although the dream and aim of the author of this thesis was to find a "holy grail" for cancer detection, it was soon recognised that, as most of the cases in statistics, all that could be found was a contribution for explaining the uncertainty and variability of shapes that could lead to indicate or propose a diagnostic of cancerous cells.

At the beginning of my doctoral program, topics in stereology and the study of two dimensional profiles of cells opened a door to step into a fascinating world of discoveries. By the time I was getting a-hold of and started understanding the various techniques that could be used in this area, sadly and happily enough, many results and discoveries had been reached by people like Hobolth, Jensen and Pedersen [14, 16]. Sadly because it narrowed the field of opportunity to do something significant and "new" for the field; happily as they had found some good parameterisations to aid in the diagnosis of cancer.

Rather than being disappointed and disheartened, the decision to "build on the shoulders of giants" was reached, a large amount of time was spent in analysing their approaches and in learning and analysing new tools that could be used to overcome some of the limitations that past approaches had encountered. This thesis deals with the use of a specific tool that was found to be useful in finding characterisations (though maybe not unique) of normal and malignant nuclei of cells, and its intent is that of setting an approach to cancer and non-cancer cell nuclei classification.

The most important gain in the approach used in this thesis is being able to relax the constraint on the shape of the nuclei to be analysed. Past work in this field had been aimed at analysis of star shaped nuclei, which excluded many actual

observations from being included in the analysis.

This thesis will seem to be more "algorithmic modelling" than "data modelling" to quote Breiman's words in his 2001 paper [3]. Yet it parts from the idea of letting the data talk and from intuitive ideas that the reader will hopefully find "natural" and interesting. This thesis presents the use of Functional Data Analysis (FDA) [29] in a rather new perspective. Such techniques had not been used on closed curves such as the cell profiles and this thesis focuses on such applications.

Although Functional Data Analysis' techniques are not deemed to be inferential per se, still some generalisations like Functional ANOVA have been discussed and used by the Ramsay and Silverman [30]. In recent conferences FDA has been receiving criticism over the validity of this type of extrapolations where we can not impose statistical assumptions on the functional error terms as directly as with the non-functional data.

Chapter 1 of this thesis presents preliminaries on the statistical and biological motivation for the research. It also presents an overview of the stereological approaches mentioned previously in this introduction and it ends with a brief discussion on FDA. Chapter 2 describes what all statisticians have to go through every time they have data, and that is data preprocessing for the proposed analyses to follow. Functional analysis using linear interpolation is discussed in Chapter 3 and the first attempts at finding characteristics for classification are attacked, namely Principal Components Analysis and Discriminant Analysis. A seemingly more complex approximation for the functional data is performed via basis functions in Chapter 5 and a curvature based classification approach is presented. In the latter part of Chapter 5 we take advantage of the functional form of the data and its basis function approximation for

an in-depth model-based analysis of the variability in the curves. Principal Differential Analysis is presented and used to gain more knowledge on the characteristics that are particular to each set of nuclei.

The main goal of this thesis is to give the reader a taste of the richness of Functional Data Analysis and to encourage the mixture of algorithmic and data-based modelling at the same time as we aim to find characteristics that could give a diagnostic for a set of nuclei profiles. We are not proposing these methods to be the one approach for objectively attempting to classify cancer and healthy cells, but to be another aid in the diagnosis and early detection.

Chapter 1

Preliminaries

Tissues and organs in the human body are generated by the process of mitosis, that is, a process of cell division. This process is most of the time regulated and in such cases regeneration of damaged tissues or organs is achieved. Cells stop their mitosis once the cells have become specialised and have taken on specific functions. On the other hand, if the process is not normal and therefore the mitosis unregulated, then cells divide either too slowly or too rapidly. When mitosis is too slow organs or tissues are not replaced in a timely fashion and problems occur. If the mitosis is too rapid and/or uncontrolled the result can be the generation of cancerous tissue.

Mitosis is a process that runs at the nuclei level and, as such, there is an interest in studying the nuclei of cells with the purpose of detecting cancerous cells. Several studies (e. g. [23, 16]) indicate that the morphology of the cell nucleus will tend to be different in a healthy cell from what it is in an unhealthy cell. For example, Popescu *et al* [26] mention that "apoptotic cells have smaller, condensed and intensely stained nuclei compared to normal cells". In this sense, it is expected that one could be able to find morphological characteristics proper of cancer cells. As a matter of fact, when cancer cells are in an advanced state, the pathologist is able to visually differentiate them from normal cells.

The two-dimensional profile of the nucleus of a cell can be considered a continuum of points. Norris *et al* [24] mention in their study of endometrial cancer that, even though pathologists are specifically trained, their ability to distinguish morphological

characteristics or parameters along this continuum is still a subjective matter which is often irreproducible. Hence, an objective quantitative reproducible method of analysing the morphology is needed.

The study of shapes involves the imaging process step to get a "drawing" or graph, and the quantitative study of descriptors that serve the purpose of characterising such shapes. It is in the characterisation step that we are interested.

The study of shapes has been visited with different approaches by many authors with different techniques; see Loncaric [21]. Grenander and Manbeck [13], in their potato experiment, attack the problem by approximating first a polygonal template to the shape and later using discretised versions of continuous templates. Others, like Chang *et al* [4] (mentioned by Loncaric [21]) and their points of higher curvature, like Lele [18, 19] or Dryden and Mardia [7], have used the existence of obvious landmarks when analysing shapes. The continuum of the cell membranes or of the nuclei profiles calls for a continuous approach where there might be no obvious landmarks. Much work has been done in this area by many authors, such as Grenander [12], Hobolth, Pedersen and Jensen [14, 15, 16].

When observing nuclei, it is difficult to distinguish, with the naked eye, specific features or landmarks that could define the shape. In this sense, Miller *et al* [23] described a model for representing spatial profiles with no obvious landmarks, as is the case of cell shapes.

Recently Hobolth, Pedersen and Jensen have described cell nuclei as a deformable template model, both as a stochastic model on the deformations from an ellipse [16] and with the residual process from a radius vector function defining a star-shaped body [14].

The basis of the first analysis is to model the profile as the possible deformations from a known shape or template. They start by fitting an ellipse to the profile via least squares [10]. Then, from a discretised set of points in the ellipse, the edge that delimits the nuclei is observable as the set of points in the profile that are orthogonal to the set of points in the ellipse. Let F be the profile defined by $F(t)$, let C be the template curve defined by $C(t)$ and let $X(t)$ be the "height" process that defines the deformation from the template. The value of the function $X(t)$ is negative if the profile is inside the template and positive if the profile is outside the template. Let $\omega(t)$ be the unit vector that denotes the normal direction at a given point in the template.

The nucleus profile is then seen as a function $F(t)$ for t in $[0,T]$. This is: $F(t) = C(t) + X(t)\omega(t), \quad 0 \leq t \leq T$

So, formally we can define :

$C = \{C(t) \in \mathcal{R}^2 : 0 \leq t \leq T\}$

$F = \{F(t) \in \mathcal{R}^2 : 0 \leq t \leq T\}$

The parameter T is, without loss of generality, normalised to 1.

The work of Hobolth, and Jensen [16] dealt with the challenge of modelling the process $X(t)$. This process was modelled as a stochastic process where, given the nature of the connections between points in the nuclei, the points can not be considered to be totally independent. Markov second order properties were imposed on the stationary cyclic stochastic process. The process was also considered to be Gaussian with mean zero. The class of Gaussian process was then defined by the parameterisation of the covariance function for the process.

Mainly, the inverse of the covariance matrix Σ of the stationary residual process

$X(t)$ is a circulant matrix and the order of the Markov property is constructed into it. A circulant matrix $M_{n\times n}(a_0, a_1, a_2)$ is defined by:

$$M_{ij} = \begin{cases} a_0 & i = j \\ a_1 & i = j-1, j+1 \mod n \\ a_2 & i = j-2, j+2 \mod n \\ 0 & \text{otherwise} \end{cases}$$

The matrix for the second order Markov property they use is such that the inverse Σ^{-1} is $M_{n\times n}(\alpha/n + 2\beta n + 6\gamma n^3, -\beta n - 4\gamma n^3, \gamma n^3)$. Given that the elements off the diagonal by 3 places are zeroes, and that the process is Gaussian, the Markov second order property is satisfied. α, β and γ satisfy a one-to-one relationship with the parameters of the covariance function.

Hypothesis tests were performed on β and γ to assess the order of the Markov process and it was found that the second order model with $\beta = 0$ was best. With this model they estimate α and γ and find that these parameters are useful in separating the malignant and benign classes of nuclei.

In 2002, the deformable template model was revisited by Hobolth *et al* [14] and now the shape was modelled with a radius-vector function and once again $(X(t))$ played the role of a Gaussian residual process or deformation process. The radius vector function is of the form

$$R(t) = r(t) + X(t), \quad t \in [0, 1]$$

where $r(t)$ is the radius vector function of the template and $X(t)$ is the residual process. The Fourier transform of such functions was used and the analysis of the

amplitude, angle and phase coefficients performed. They assume the residual process to be Gaussian, so the expansions for each of $r(t)$ and $X(t)$ are

$$r(t) = a_0 + \sqrt{2}\sum_{s=1}^{\infty} a_s \cos(2\pi st) + \sqrt{2}\sum_{s=1}^{\infty} b_s \sin(2\pi st)$$

and

$$X(t) = A_0 + \sqrt{2}\sum_{s=1}^{\infty} A_s \cos(2\pi st) + \sqrt{2}\sum_{s=1}^{\infty} B_s \sin(2\pi st)$$

where, because of the Gaussian assumption, A_0, $A_s, B_s, s \geq 1$ are all mutually independent and $A_s, B_s \sim N(0, \lambda_s)$. The variances λ_s are modelled with what they call "the p-order model" having

$$\lambda_s^{-1} = \alpha + \beta(s^{2p} - 3^{2p}), \quad s \geq 3$$

Here, α determines 'global' deviation from the template while β, p determine 'roughness' of the boundary.

Their findings were that on average the estimates of the global shape parameter α were significantly lower for the malignant sample, the estimates of local shape β are also significantly lower in the malignant sample, and the variance of $log\beta$ was significantly larger in the malignant sample.

The process $\{X(t)\}$ of deformations from a template has been presented and used to represent the continuous process of the continuous cell membrane that creates the shape or profile of such cell. In this sense it seems reasonable to consider the cell profile as a functional data source. In the methods of Hobolth and Jensen statistical assumptions are imposed on the residual process, in their 2000 work [16] directly on the covariance structure of the Markov process, and in 2002 [14] on the Fourier expansion coefficients of the process.

There are advantages in considering the profile itself and/or the residual process as a functional data source. The main idea in the analyses mentioned so far is that there exists some difference in the morphology of cancer and non-cancer cell nuclei. Based on this, it is not unreasonable to consider that there is a functional process that generates each type of profile. Hence, the approaches suggested in this thesis will deal with the analysis of profiles expressed as a bivariate functional data source. The tools used in analysing them comprise:

- The use of the whole profile as generated by the bivariate function $(X(t), Y(t))$ which we observe as x, y coordinates creating the continuous function as the result of linear interpolations of the discretised data and applying:
  - Principal Components Analysis directly on the observed fine grid,
  - Linear Discriminant Analysis directly on the observed fine grid,
- Analysing the correlation structure within each of the univariate functional processes $X(t)$ and $Y(t)$ that form the bivariate process, as well as the cross-covariance between them.
- Creating smooth functional data as approximated by basis functions expansion in order to:
  - Analyse the behaviour of their derivatives
  - Perform statistical analysis on the coefficients of such basis functions.
- The study of the cell profile as a univariate functional data source has been visited with the residual process $X(t)$ as the deformation from a template by other authors, most recently Hobolth and Jensen [16, 14].

The radius vector function of Hobolth, Pedersen and Jensen [14] is, in part and although they do not present it as such, a FDA approach. Given the restrictions on star-shaped bodies, they are creating their radius vector function as a function of an angle θ as is done in polar coordinates. Once this radius vector function is defined, it is approximated by Fourier series and the statistical assumptions are imposed on the amplitudes and angles of such series. This translates into imposing statistical assumptions on the coefficients of the Fourier approximation. In this work they part from the assumption of a stationary process given the fact that the template is not circular but elliptical. However, they perform a transformation on their "time" parameter, which they call a "time change" such that the process in the new "time" is stationary. This is similar to time-warping in FDA for registration. In short, they have done a thorough job in applying FDA-like techniques for their analysis and it is therefore not dealt with in this thesis.

The analysis they perform is constrained to star-shaped planar objects. A star-shaped planar object is defined as an object in which at least one interior point, say z, exists and has the characteristic that, for all points y in the boundary of the object, the entire segment $\overline{yz}$ belongs to the object.

Dealing with only this type of object sets constraints on the type of profiles we can study. Figure 1.1 shows some of the profiles that we have, which would violate the definition of a star-shaped object.

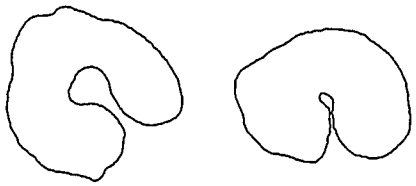


Figure 1.1: Non-star-shaped profiles

The star-shaped body constraint is relaxed when dealing with the bivariate functional data since the separation of X and Y functions enables the cell to follow any desired pattern. The parameter t that controls the bivariate function does not conflict with having 2 or 3 points lying at different distances from the center z for a given ray at a given angle. These discretely observed functional data can be approximated with different basis functions such as linear approximations, Fourier transforms, splines and wavelets to render them in continuous form. The derivatives of the functional data can be computed from these, now continuous, data. Analysis on the behaviour of the derivatives sheds light on possible discriminant features that may be hidden to the naked eye.

1.1 Functional data analysis overview

In recent years, the feasibility of obtaining, recording and storing successively measured data from continuous processes such as the position in three dimensional space of knees, ankles and hips of a person walking, have called for more efficient and interpretable ways to analyse such data [1]. Functional Data Analysis (FDA) recognises that the succession of measurements, say y_i, although measured at discrete intervals (e.g. times t_i), are realisations of continuous processes or functions ($x(t)$), and provides suitable analysis methods for such type of data. In order to "preserve" and render the continuous nature of the functions, the discretised data are interpolated or smoothed [28, 29, 8] with a choice of different procedures depending on the nature of the data (Fourier for cyclic sinusoidal type data, polynomials and splines for other types).

A more tangible way of thinking about FDA, mentioned in [28, 29], is to consider it as an extension from the multivariate scene, in a similar way in which Multivariate Data Analysis (MDA) would be an extension of univariate data analysis. The "extension step" involves the need for a more structured mathematical basis, but the "discrete-to-continuous" extension is intuitively appreciated as consistent with the nature of the data.

We can run into trouble in MDA if the number of observations we have is less than the dimension of the data. The continuous curve that represents the functional datum can be taken to be an infinite-dimensional discrete variable which would be difficult to approach with MDA. An advantage gained from the FDA approach is that the multicomponent vector that might be unsuitable for MDA can be analysed

as a curve with FDA techniques.

Data can be considered functional to accomplish data reduction. In some of the smoothing techniques a comparatively small number of parameters is needed to permit the analysis of functions as a whole. There is, in FDA, the possibility of using summary statistics analogous to the univariate and multivariate case in their functional form, such as mean,

$$\bar{x}(t) = N^{-1} \sum_{i=1}^{N} x_i(t)$$

variances,

$$\mathbf{var}_X(t) = (N-1)^{-1} \sum_{i=1}^{N} [x_i(t) - \bar{x}(t)]^2$$

covariances and correlations,

$$\mathbf{cov}_X(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} \{x_i(t_1) - \bar{x}(t_1)\} \{x_i(t_2) - \bar{x}(t_2)\}$$

$$\mathbf{corr}_X(t_1, t_2) = \frac{\mathbf{cov}_X(t_1, t_2)}{\sqrt{\mathbf{var}_X(t_1)\mathbf{var}_X(t_2)}}$$

permitting then the possibility of descriptive analysis and inference-based decisions.

Information between sampled or measured points becomes available through the smoothing process. FDA estimates the continuous function from the observed data and is then used to discretise the function to enable us to find the values of the function at the desired points for analysis. It is not necessary that the sampled points be equally spaced. If we are interested in more intense sampling over a certain interval, this can be done easily [11, 29]. As part of the models considered in this thesis, and for purposes of convenience, when sampling points from the profiles for analysis, evenly spaced points were used.

Figure 1.2 shows the discrete values of the X coordinate of one of the profiles used in this thesis as circles and its spline-smoothed version as the line.

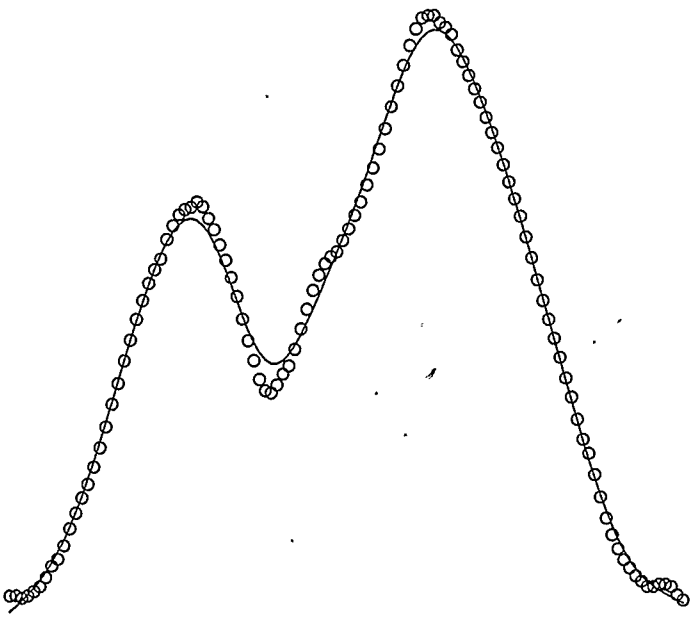


Figure 1.2: Discrete data and smoothed function

FDA is useful in analysing variation of functions across different study units or different types of study units. A considerable advantage resides in analyses done with FDA, in that functions, particularly smooth functions, can be analysed through their derivatives and such derivatives can have an interesting interpretation in the context of the study. In this thesis, the use of derivatives will enable a quantitative and

objective analysis of the profiles' behaviour regarding smoothness and overall shape.

The use of FDA techniques and the representation of data in functional form involves the use of some reference parameter. In most situations time is the parameter used. The word 'time' is used often and very loosely in FDA. In studies that deal with processes that specifically happen through time, such as angle measurements in a gait analysis [1], the word refers indeed to time. The functions can be thought of depending on a variable t that is assumed to be defined (as is time) continuously. According to this, the functions have a starting point "time" zero $t = 0$, and without loss of generality, the "time" can be normalised to the interval $[0, 1]$. In this sense, when we talk about equally time spaced points we are talking about having $n + 1$ points at times $t_i = i/n$ for $i = 0, 1, ..., n$ so that $t_0 = 0$ and $t_n = 1$.

As described, FDA does not necessarily deal with functions of time as such, and when it does, it is analysing features of variation and features that separate or define one type of function from another type. It does not deal with forecasting nor signal processing and hence is essentially different from Time Series Analysis.

In order to obtain the continuous underlying function from the discrete observations, FDA techniques approximate the functions by expansion of basis functions [11]. Part of the techniques and a consequence of this expansion is smoothing which has sometimes brought on the criticism that FDA is mainly a smoothing process; see the discussion in the commentaries to [28]. Although smoothing plays an important role in the 'discrete-to-continuous' extension, it is not "just smoothing"; in reality, the smoothing creates connections between discretely observed values and more importantly, the result of this smoothing is the estimation of the values of the function, hence the smoothing or interpolation has much more meaning in essence than it has

in, say, time series.

1.2 Functional Data Analysis Techniques

In this thesis, the profiles are considered to be continuous functions and hence functional data. The analysis for these data considers the profiles as the result of a bivariate process $Z(t) = (X(t), Y(t))$ and studies the relationship between the X and Y parts as well as the difference between these in malignant and benign profiles. We intend to determine out if the autocovariance of each the X, Y processes is different for cancer and non-cancer cells. We also study the cross-covariance, that is the influence that each component has on the other at given times. An exaggerated but illustrative example is to think of normal profiles as perfect circles and cancerous profiles as rather flattened ellipses; the mapping of the values for the covariances and cross-covariances would be different.

1.2.1 Principal Components Analysis

In MDA, ordinary PCA creates linear combinations of the X_j variables, where the first of such combinations is given by finding ξ_1 for

$$f_{i1} = \sum_{j=1}^{p} \xi_{j1} x_{ij}$$

such that $N^{-1} \sum_i f_{i1}^2$ is maximised subject to $\sum_j \xi_{j1}^2 = \|\xi_1\| = 1$.

Subsequently, second and up to p (number of variables) steps are carried out, computing, in the lth step ($l \leq p$), ξ_l and new f_{il} that again has the maximum mean square, subject not only to the norm of ξ_l being one but also having $l-1$ additional

constraints

$$\sum_j \xi_{jk}\xi_{kl} = 0, \quad k < l,$$

which amounts to having orthogonal components.

The functional counterpart of PCA becomes

$$f_i = \int \xi(s)x_i(s)ds$$

such that $N^{-1}\sum_i f_{i1}^2$ is maximised subject to $\int \xi_1^2(s)ds = 1$ for the first component and the analogous orthogonality constraints are imposed for ξ_l. Hence instead of having the loadings ξ_i being p-dimensional vectors, we have functions $\xi_i(t)$.

A first analysis for the bivariate functional form is to create the functional datum by considering linear interpolation of the observed x and y coordinates. With the linear interpolation the Principal Components Analysis (PCA) [17] is applied to the 2-vector forms and we expect to find a difference in the means of some of the PCA scores which might reflect distinguishing characteristics for the overall shape of the profiles. In this step, graphical interpretations of each principal component are performed by studying the mean shape of nuclei augmented by the principal component.

In PCA we expect to have only a few principal components to be important in setting malignant T-cell lymphoma nuclei and normal T-lymphocytes nuclei apart. This is expected because the profiles are comparable to elliptical templates. As a consequence, the first two components are expected to deal with the overall shape of the profiles - eccentricity and convexity - where most of the large-scale variability would exist. The subsequent components are expected to deal with more local differences in outlines for the profiles.

1.2.2 Linear Discriminant Analysis

Another method used is Linear Discriminant Analysis [17, 30]. This enables the classification of profiles into malignant T-cell lymphoma nuclei or normal T-lymphocytes nuclei profiles, from what seems to be a less geometric approach. However, the analytical form of the discriminant function takes into account the geometry. Having the linear discriminant function as:

$$\hat{\delta}_i = \int_0^1 \{X_i(t)\alpha_X(t)dt + Y_i(t)\alpha_Y(t)dt\}, \tag{1.1}$$

where $\alpha_X(t), \alpha_Y(t)$ are weight functions, it is seen that the X, Y functions will yield different $\hat{\delta}$ values depending on the geometrical body they delimit, even if linear combinations of the discrete $X(t)$ and $Y(t)$ were used - for example using the first principal components as our functions. Linear Discriminant Analysis is discussed further in Section 3.2.

1.2.3 Basis Functions Expansions

Linear interpolation from the discretised observations of $X(t), Y(t)$ is the interpolating method indicated in the previous sections. This method offers simplicity but presents the problem of not having smoothness. Information on derivatives of the continuous processes is sought and hence smoothing the functional data as expansions of smooth basis functions is needed to be able to get such derivatives [30, 28].

These types of expansions, apart from offering derivative information, can be useful in the possibility of analysing the coefficients of the basis functions themselves. Apart from the behaviour of the derivatives, we can use differential equations models

that would separate the intrinsic periodical structure and leave residual processes for analysis. This approach is named Principal Differential Analysis by Ramsay and Dalzell [28].

The most commonly used basis functions are splines and Fourier. Ramsay and Silverman [29], and Green and Silverman [11] have used both of these expansions and have developed tools for their use, as well as applied the penalised approach to smoothing. These tools are used in analyses performed in this thesis.

Chapter 2

Data and data preparation

A set of 100 nuclei profiles was obtained, provided by Hobolth and Jensen. The data comprise the profiles of 50 normal T-lymphocytes nuclei and 50 malignant T-cell lymphoma nuclei.

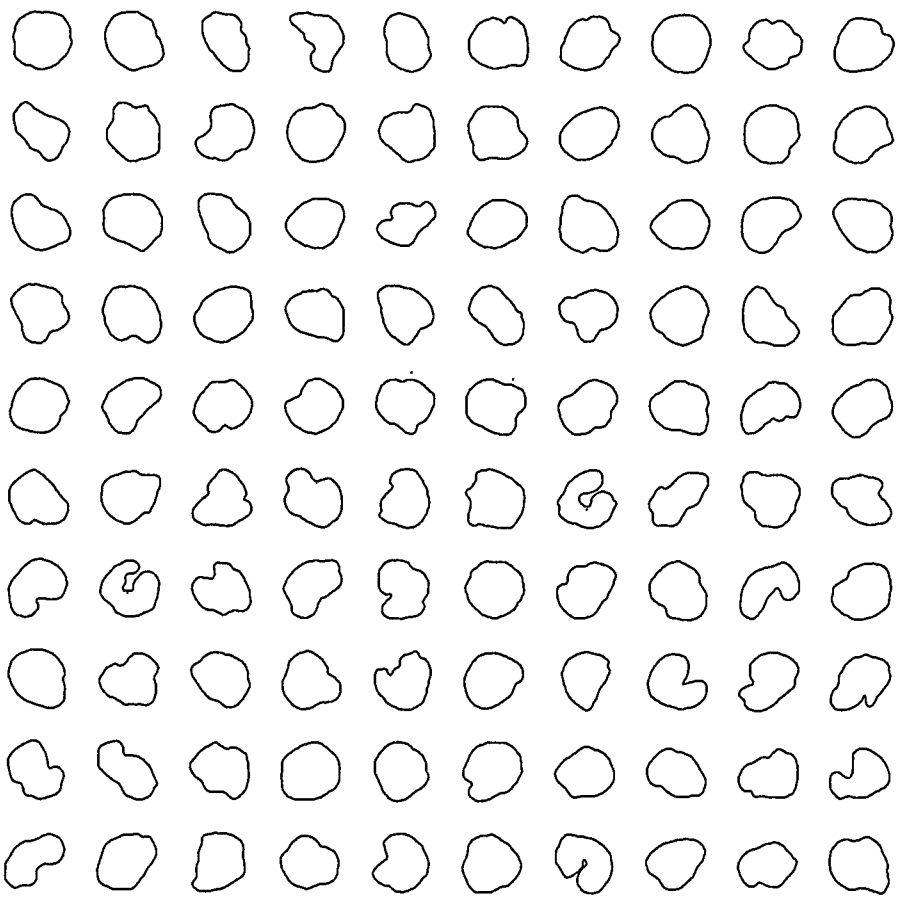


Figure 2.1: Profiles of nuclei from 50 malignant T-cell lymphomas and 50 normal T-lymphocytes

Figure 2.1 shows the 100 profiles all together. As can bee seen in the figure, the profiles vary in form and orientation. Figure 2.2 shows a thicker point indicating the way in which the profiles were read when digitised from the microscope. The nuclei were first outlined from a fixed point with integer coordinates in the device used to view them. That is to say that there is no statistical nor biological reason for starting to outline the nucleus profile from what seems to be the uppermost point at the "centre" of the horizontal plane diameter of the profile. Even if a specific location was desired, it would have been very difficult to "line up" the nuclei by eye. These nucleus profiles were kindly provided by Hobolth and Jensen.

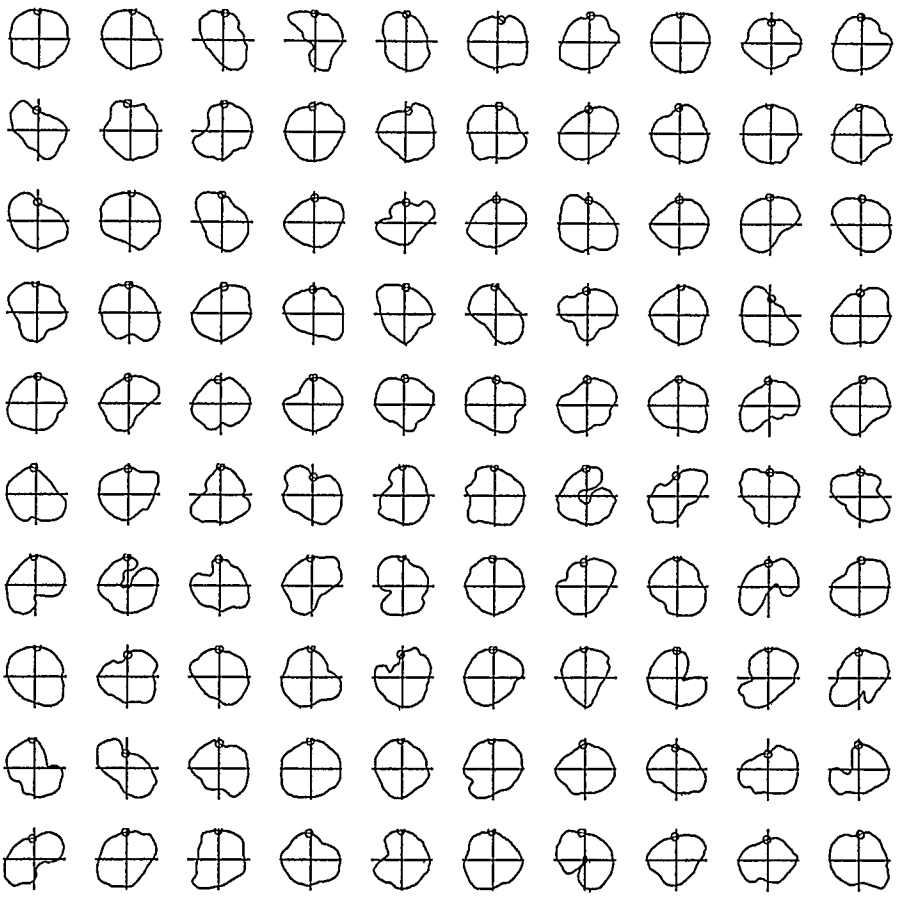


Figure 2.2: 100 profiles with centres

The profiles, in this thesis, were to be analysed using FDA; and in such analysis we wanted to have a true "reference" point or time t_0 that was not only meaningful in the sense of being the first point of the nucleus profile, but that would be, although arbitrary, determined by the same criteria for each profile. In such an effort, each nucleus was fitted with an ellipse via least squares to obtain information on the rotation, if any, of such corresponding ellipse.

Ellipses are used for the fit as they are the next step in geometrical complexity after the circle and they are still simple enough to manage.

For the first stages of our analysis, where we are interested in overall shape, the profiles are "aligned" to avoid creating fictitious variability. If there are two ellipses that differ slightly, but only on their eccentricity, and if one of them is viewed as having its semimajor axis rotated by $\pi/2$ and the other is viewed as a lying on its semimajor axis, more variability between the shapes seems to exist. If both ellipses are resting horizontally on their semimajor axis then we induce no fictitious variability.

Some controversy sorrounds the alignment or registration procedures. There are two main tendencies regarding shape analyisis, the landmark based approach and the outline based approach. We follow an outline based approach.

Macleod [22] states "hard distinctions between landmark and outline morphometric data/analysis are illusory and damaging to the entire enterprise of morphometrics". He argues that although biological correspondence for measurements is legitimate, it does not address or avoids in itself the potential source of error. In his article he states that any comparison that is meaningful happens at the landmark to landmark comparison which is as good as the curve to curve comparison in comparing outlines.

In palentobiology [?] and mathematical geology [6], amongst other sciences, eigenshape analysis consists of tracing a closed curve from a starting point, which is as a standard, a landmark in the shape and the angular differences between equally spaced points are taken to represent the shape function. Singular value decomposition procedures are applied to these. This procedure is analogous to PCA.

Lohmann [20] argues that when there is no obvious or common landmark in its biological or physiological sense, objects are matched or aligned by adjusting the starting point to be a location determined by the maximal correspondence of the outline to that of a reference outline. This is analogous to the alignment done in this thesis by using the leftmost point in a profile that is aligned to the semi major axis of the fitted ellipse.

Ferson, Rohlf and Kohen [9] describe the use of an invariant approach to Fourier elliptical decomposition. The Profiles are traced and then their Fourier coefficients "normalised" by a rotation, size and starting point transformation that makes the coefficient consistent. This work drives the first three Fourier coefficients to be 1 and 0, 0 which are the coefficients that represent an elliptical form, namely an ellipse with either a horizontal or vertical semi major axis. The procedure appears to be invariant to the orientation of the profiles and to the starting point as the profiles are traced in any given position and at any starting point. After the transformation, the profiles end up being 'aligned'. This is equivalent to rotating the profiles to lie on their semi major axis and the starting point to be as we chose in this thesis.

It is worth mentioning that we are not searching for the biological reason that makes the shapes of the profiles to be the way they are. We are not assuming any biological homology. We measure shape itself as Ferson *et al* [9] do and therefore, quoting them "it is valuable to quantify shape variation sensu stricto".

2.0.4 Fitting the ellipse

The fitting of the ellipse was done with the method discussed by Fitzgibbon *et al* in their 1999 paper [10] which is based on solving a generalised eigenvector problem.

Let D be the matrix that has as rows the vectors $x_i = [x^2\ xy\ y^2\ x\ y\ 1]^T$, let S be the square dispersion matrix obtained from $D^T D$ and let $\mathbf{a} = [a\ b\ c\ d\ e\ f]^T$ such that $ax^2 + bxy + cy^2 + dx + ey + f = 0$. Then solve the generalised eigenvector problem:

$$\begin{aligned} Sa &= \lambda Ca \\ a^T Ca &= 1 \end{aligned}$$

where C is the constraint matrix that guarantees elliptical results by constraining the system to have $4ac - b^2 = 1$. In matrix form C is

$$\begin{pmatrix} 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The solution to this system minimises the algebraic distance from the points to an ellipse, resulting in the coordinates for the centre, the angle of rotation and the semimajor and semiminor axes' lengths.

2.1 Alignment

The alignment and standardisation of the profiles is obtained by rotating the profiles in such a way that the best fitting ellipse will be resting on the semimajor axis. In order to have more clearly comparable profiles, after being rotated, the profiles

are centered and scaled in such a way that their caliper diameter, measured parallel to the semimajor axis, ranges from -1 to 1. This standardises the range of the X coordinates to be in $[-1, 1]$. The Y ranges are scaled by their corresponding X factor to preserve perspective and ratio between X and Y in each of the profiles.

This approach seems to be arbitrary and artificial, but is indeed preventing the introduction of variability due to rotation or size in the study. This normalisation is performed in the same spirit as Ramsay and Silverman do for the bone shapes and the intercondilar notch in their case study publication [30].

It is thought that performing this centering and scaling might have the same effect that registration would have on the profiles. If registration in the FDA sense is to be performed on the profiles via the bivariate $(X(t), Y(t))$ function, we must keep in mind the fact that these two are not independent and hence the same time warping function should be used for both functions to keep the original shape.

Figure 2.3 shows one of the nucleui with its corresponding fitted ellipse. Knowing the centre and rotation of these fitted ellipses gave the needed information for "aligning" all nuclei to lie in their corresponding ellipse's major axis. This action allowed for all nuclei profiles to have a well-defined reference point $(X(t_0), Y(t_0))$.

After rotating the nuclei according to their corresponding ellipse, we determine the point that will be deemed as $(X(t_0), Y(t_0))$. This point is chosen as the leftmost point that lies on the semimajor axis; see Figure 2.4. If the desired point is not one of the sampled points, as it is possible that there might be no point whose Y coordinate is exactly zero, the needed point is obtained by interpolation.

Figure 2.5 shows all the nuclei rotated and time zero t_0 is shown as the point in the plot.

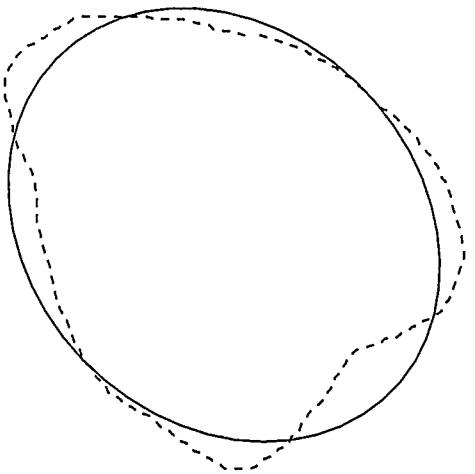


Figure 2.3: Ellipse fitted to nucleus profile

The starting point or reference point,although arbitrary, has been determined in the same way for all nuclei now. For the linear interpolation in the profiles we can start measuring the arc length from t_0. Each profile is represented by 150 equidistant points. The distance between points refers to arc length in the profiles. For example if the perimeter of the profile is 10 length units, then the points were chosen to lie at distances $0, 10/149, 10*2/149, ..., 149*10/149)$ so the that total perimeter will be covered.

Figure 2.6 shows the parameterisation of the $X(t), Y(t)$ coordinates of the first normal T-lymphocyte nucleus with the equidistant approach.

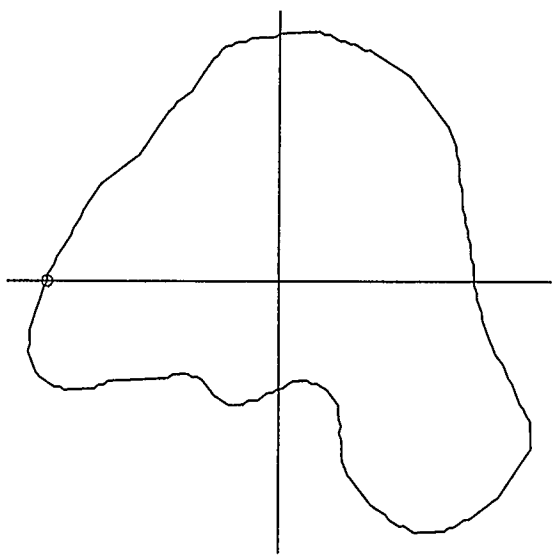


Figure 2.4: Rotated profile

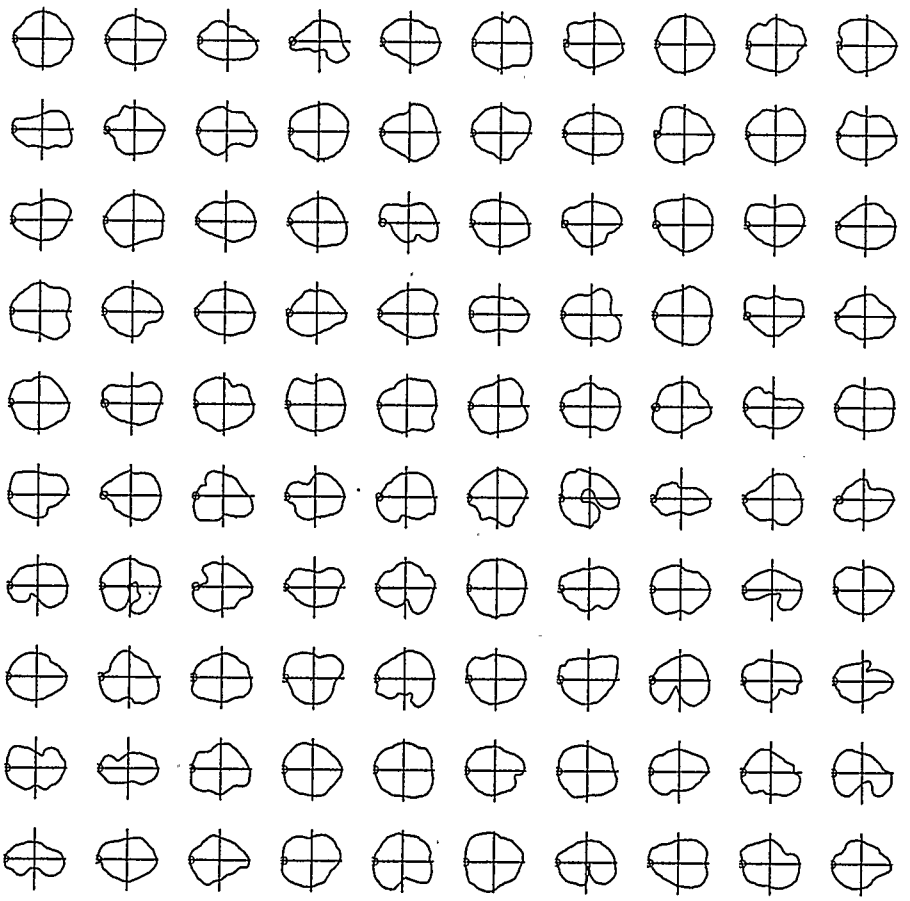


Figure 2.5: Rotated nuclei with starting points

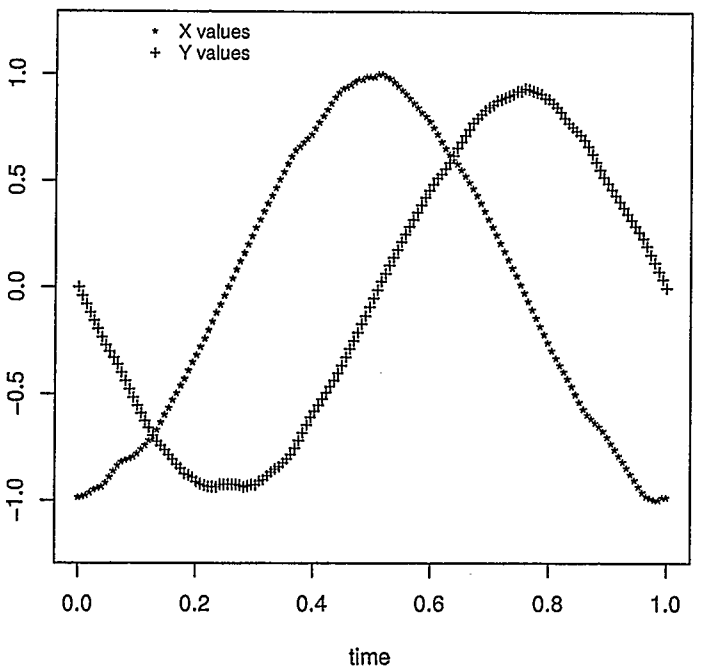


Figure 2.6: $X(t), Y(t)$ based on equidistant points

The points can also be chosen by normalising the total time interval for each profile, hence having the points $0, i/n$ where n is the number of sampled points that define the profile and $i = 1, ..., n$, and then interpolating linearly the X and Y values for 150 equally spaced time points between $[0, 1]$. Figure 2.7 shows the same profile as in Figure 2.6 parameterised by taking the 150 points as uniformly distributed on the $[0, 1]$ time interval. The number of sampled points differs from profile to profile, however this is not a problem since we have determined time zero t_0 beforehand.

Although the values of the X, Y coordinates will not necessarily be the same for

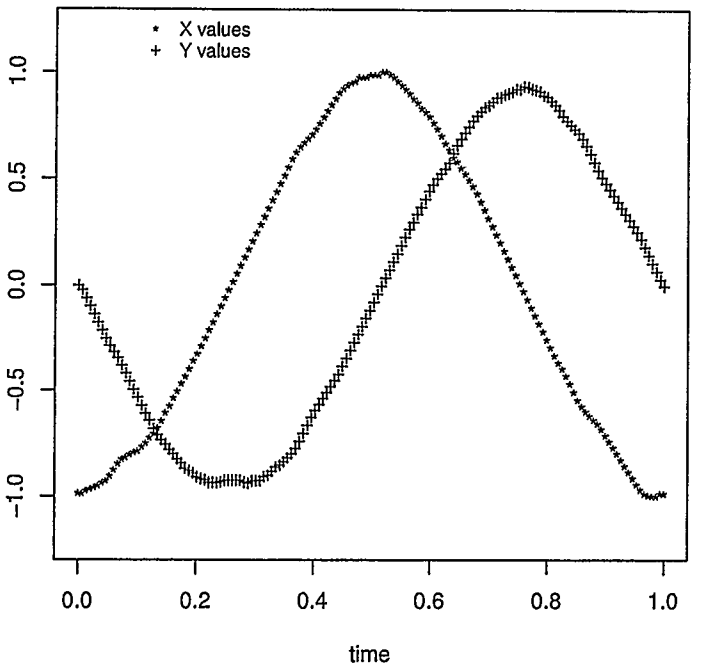


Figure 2.7: $X(t), Y(t)$ based on uniform time points

the two choices of determining times $t_0 = 0$ to $t_n = 1$ we can expect that the choice used will not affect the analysis. Comparing Figure 2.6 and Figure 2.7 the difference seems negligible to the naked eye. The principal component analysis is performed for each of the choices and compared for sensitivity.

All the smoothing done in the thesis is performed using these points as the basic data for analysis. In many applications of FDA the functional data are registered in order to better 'align' characteristics and enable the analyst to find differences. In our case, however, registration was not performed as such. The profiles were rotated

to rest in their major axis so that analysis on the $X(t)$ and $Y(t)$ as functions of time would not be artificially affected by having some "horizontal" and some "vertical" profiles.

At further stages, where derivative information is needed and analysed, the data are smoothed and approximated by basis functions. Given the cyclic nature of nuclei, as they are closed curves, approximation for each of the coordinates in the $X(t), Y(t)$ process is based on Fourier expansions for the underlying cyclic structure and compared to the spline fit in order to extract residuals information.

Chapter 3

Analysis with linear interpolations

3.1 Principal Components Analysis

The profiles of the nuclei are formed by the X, Y pairs at each time t, and in this manner each of the pairs contributes to the variability of the profile at specific positions in the profile. Based on this, the profile can be seen as having the 150 points as variables and then Principal Components Analysis can be performed to discover the type of variation that affects each of the types of profiles the most.

As mentioned in section 2.1, there are two approaches to the selection of the $n = 150$ generated points. One approach is to take arc-length equidistant points in the profile so that at in-between times $t_i = i/(n-1)$, $t_{i+1} = (i+1)/(n-1) \in [0,1]$ we have walked the same distance along the profile. The other approach is to set $n = 150$ time points uniformly distributed over $[0,1]$ and interpolate on the existing X, Y sampled points for their respective $X(t_i)$ and $Y(t_i)$ values.

In order to perform PCA, each bivariate $X(t), Y(t)$ datum is taken as separate in each of its coordinates. The data from the 100 profiles are arranged in 100 rows with 300 columns, (150 for each of X and Y coordinates) and multivariate PCA is performed on these [30]. The resulting matrix of rotations or loadings is rearranged as a three-dimensional array for easier access and interpretation. This array has in its first two dimensions 150×2 matrices of loadings for the 2-vector X, Y pairs, and its third dimension accounts for the 100 'pages' corresponding to the 100 profiles.

The purpose of performing PCA on the data is to try to detect differences in the two groups while reducing the data dimensionality. We expect to find differences in the components' scores for the two different types of profiles. Apart from the score obtained from the 'linear combination' or score, we gain interpretability from the principal components in a graphical sense by investigating the possible effect that each of them has on the geometry of the mean profile.

The effect of the principal components on the shape of the profiles is captured graphically by adding and subtracting a fixed amount C times the standard deviation of the component to the mean profile (obtained by averaging out the values of $X(t), Y(t)$ for each fixed t).

The PCA routine returns 100 components and we are interested in keeping the ones that account for the highest percentage of the variability. The following screeplot shows the decrease in percentage of variance accounted for per component, and hence is a guide for choosing the number of components to study.

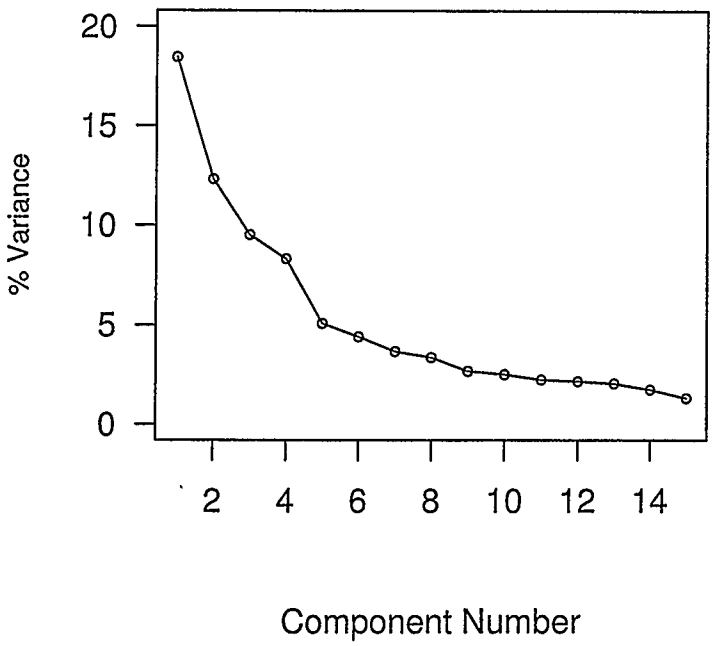


Figure 3.1: Plot of variance accounted for by PCs

Figure 3.1 indicates that there would be no real gain in the variability accounted for after using about 6 principal components.

The effect of the principal components on the shape of the profiles is shown in Figure 3.2.

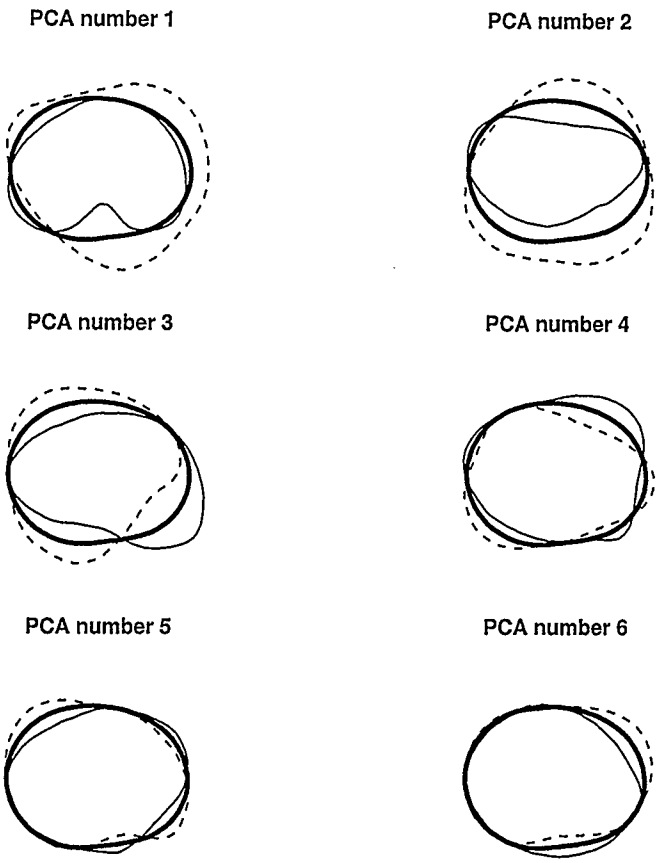


Figure 3.2: Effect of first 6 principal components on the mean profile; the thick line is the mean profile, the dotted line shows mean minus pca effect and the solid thin line shows mean plus pca effect.

Figure 3.2 shows the effect of having a component being negative or positive for profiles. The first principal component is regulating the behaviour of the convexity or concavity of the bottom part of the profile. A positive first principal component tends to make the bottom of the profile cut into the profile making it concave, whereas a negative first component tends to create a convex bump in the lower part

of the profile as well as having the profile exceed the borders of the mean profile in most directions.

The second component regulates what could be seen as the eccentricity of the profile, its roundness or its tendency to look more like a horizontal potato. A positive value on the second principal component will shrink the semiminor axis of profile, and hence the profile has a narrower Y range than the mean profile. If looking at the profile as an ellipse, the positive value of the second component creates an ellipse with greater eccentricity than the mean profile. A negative value on the second component will create a shape closer to that of a circle. The eccentricity is smaller than that of the mean profile.

For the third component, a positive value shrinks the semiminor axis, causing most of the profile to be encased inside the mean profile, except for the fourth quadrant where a positive value for the component creates a protuberance that exceeds the borders of the mean profile both in the X and Y coordinates. A negative value on the third component has the opposite effect, that is, the profile is enlarged on the vertical scale almost everywhere except at the fourth quadrant, where the profile is concave and inside the mean profile.

The fourth component affects mainly the first and third quadrant. A positive value in this component causes the profile to grow in the first quadrant, while the third quadrant effect is to pull the profile slightly to the inside of the mean profile without loss of convexity. The profile flattens on the right side. A negative value of this component has the effect of shrinking the profile in the first quadrant and protruding it on the right side, while also expanding the boundary outside the mean profile in the third quadrant.

The fifth and sixth components regulate the smaller scale bumpiness of the profile all around. In both cases the negative of the component does the opposite of the positive, hence where there is a bump in the positive part, there is a dip in the negative and vice versa.

We expect normal profiles and lymphoma profiles to have different values on some of these components. Figure 3.3 shows the summary distributions of the first six components.

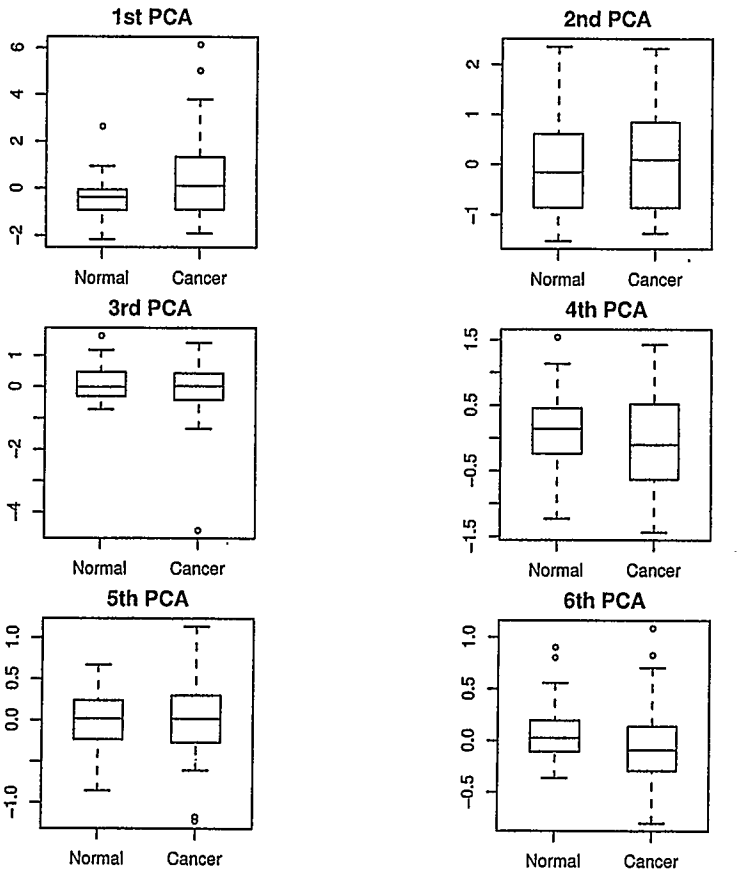


Figure 3.3: Comparison of PCA scores by profile type

Profiles of non-cancer nuclei tend to have a negative value for the first principal component and those of cancerous nuclei tend to have a positive value in this component. For the second component, normal profiles tend to have negative values while cancerous ones tend to have a positive value. For components 3 to 6 the normal nuclei tend to have a positive value and the cancerous a negative value. We want to know if these differences are significant.

Performing Welch's T test on the means of each type of profile, we see that the mean value of the first component for Normal profiles is significantly smaller than the mean value for first component of cancerous profiles (p-value < 0.002) with a 95% confidence interval of $(-1.3637, -0.2808)$. Wilcoxon's rank sum test also yields a significant difference (p-value $< .02$). So the first principal component is useful in separating normal and cancerous profiles.

Means for components 2 through 5 do not show to be significantly different. However, normal profiles have a significantly higher mean for component 6 than that of the cancerous profiles (Welch's: p-value < 0.04, Wilcoxon's: p-value < 0.03)

A sensitivity analysis for different number of points and for different starting points was performed. Different numbers of points were taken, from as low as $n = 50$ to as many as $n = 150$ points. The first component proved to be significantly different for normal and cancerous profiles for all values of n. The starting point was also shifted from $t_0 = 0$ to analyse if it would have any effect on the calculations and hence make any difference in the conclusion. This was done by dividing the interval between t_0 and t_1 by 10 and performing the analysis again for every shift of one tenth of the interval. The analysis gave the same conclusions as before.

3.1.1 Were we really doing MDA instead of FDA?

In the previous sections of this thesis the reader may be misled into thinking that there have been no FDA techniques used because of the fact that multivariate-like techniques and procedures were used. This is not the case. The fact that for the first analyses we decide to use linear interpolation on the data for conversion to functional form resulted in allowing us to use the MDA techniques on the data, almost as if doing only MDA. The very important interpretation part is the construction of vectors of the form $(X(t), Y(t))$. When using the multivariate techniques of PCA and LDA the fitting mechanics behind these methods lets us "trick" them into doing the analysis for the bivariate $X(t), Y(t)$ multivariate process as if it was processing the multivariate problem on one long string of variables.

Formally, PCA analysis looks for the linear combination of the variables that will have the largest mean square error, that is

$$PC_i = \sum_{j=1}^{p} \xi_j x_{ij}, \quad i = 1, ..., N \tag{3.1}$$

where the ξ_j are normalised weight coefficients that determine the linear combination of the i-th observed values of the j-th variable. In terms of inner products of the vectors ξ and x_i, we write $PC_i = \langle \xi, x_i \rangle, \quad i = 1, ..., N$. The process of PCA is well known and documented as in Johnson and Wichern [17], so little time will be spent on the detail of the process. The interesting part comes when dealing with functional data; here the counterparts of the vectors x_i are functions $x_i(t)$ and the linear combinations become:

$$PC_i = \int \xi(t) x_i(t) dt = \langle \xi, x_i \rangle. \tag{3.2}$$

The problem of finding the adequate ξ's is equivalent to solving an eigenvalue/eigenvector problem for the covariance matrix of the x's. The functional form is analogous except that we are dealing with weight *functions* rather than weight *vectors*. One concern is the fact that in the functional case the number of values of the functional datum is infinite as this is the counterpart for the dimensionality p in MDA and therefore, the concern for the maximum number of different eigenvalue-eigenfunction pairs arises. However, if the functional data $x_i(t)$ are not linearly dependent, the covariance operator has rank $N-1$ and hence there will be $N-1$ non-zero eigenvalues.

Let us now return to the bivariate functional case. In our case we are dealing with X and Y coordinates both as functions of time. Then a typical principal component is defined by the 2-vector $\xi = (\xi^x, \xi^y)'$ of weight functions, with ξ^x taking into account the variability of x and ξ^y the variability of y. Then (see Ramsay and Silverman [29]) the inner product is:

$$\langle \xi_1, \xi_2 \rangle = \langle \xi_1^x, \xi_2^x \rangle + \langle \xi_1^y, \xi_2^y \rangle. \tag{3.3}$$

This is equivalent to concatenating the two functions to create a new function. The same procedure can be applied to the data and have one-vectors (with double the length of the original) representing each of the profiles and perform the analysis on these. The resulting components and vectors can be rearranged as 2-vectors to be used for graphical analysis. An analogous "trick" of stringing the data can be used for performing linear discriminant analysis or other functional version of MDA.

3.2 Discriminant analysis

Assuming that we can obtain a value δ_i for each profile, such that

$$\delta_i = \begin{cases} 1 & \text{if cancerous} \\ -1 & \text{if normal} \end{cases}$$

we would want to be able to calculate such δ to be able to allocate each profile to one of Lymphoma or normal group. In multivariate analysis, where we would have vectors $W^{(i)}$ corresponding to populations $i = 1, 2$, we would be interested in finding the vector a that would determine the linear combination or linear discriminant for these profiles and a value $\hat{m}$, such that we could know from which population a vector W came by calculating whether $a^T W > \hat{m}$. In the MDA case a takes the following form

$$a = \hat{S}^{-1}(\overline{W}^{(2)} - \overline{W}^{(1)})$$

and

$$\hat{m} = \frac{1}{2}a'(\overline{W}^{(2)} + \overline{W}^{(1)})$$

where $\overline{W}^{(i)}$ is the mean for population i and $\hat{S}$ is the pooled estimator for the covariance matrix, assuming equal population covariance matrices. The method is based on a comparison of within-group sum of squares and between-group sum of squares [17].

In the case of functional data analysis, we have a function of time, say $\alpha(t)$, and we have functions $W_i(t)$ instead of the multivariate vectors.

In the bivariate functional case, the case of our profiles, we equivalently propose to have $\alpha(t) = (\alpha_X(t), \alpha_Y(t))$ in such a way that our discriminant function becomes

$$\hat{\delta}_i = \int_0^T \{X_i(t)\alpha_X(t)dt + Y_i(t)\alpha_Y(t)dt\} \tag{3.4}$$

Recalling that the functional form of the data has so far been linear interpolation, we can perform the linear discriminant analysis simply on the sampled X, Y points from each profile. The classification of the profiles can be done in this manner, but the interpretation about the directions that the profiles should deviate from the mean in order to discriminate is cumbersome and even senseless.

The analysis is carried out on the stringed X, Y points, having a matrix with 100 rows, one row per profile, where the first $n/2$ points are the X's and the last $n/2$ are the Y's.

For each of the points in the profile we obtain a direction for discrimination from the *normal* mean profile. Figure 3.4 shows with the arrows the direction of the projections for the discriminant. This figure is difficult to interpret and to follow if our aim is to classify a new profile or to look for characteristics that a profile should follow to be classified into one of the types.

The approach that proves more interpretable is to carry out discriminant analysis based on a subset of the principal components obtained from the former analysis. We first find the discriminant as a function of these components. In our case we find the discriminant based on the first six principal components that accounted for 90.78% of the variability. The choice of 6 principal components for the discriminant calculation was based not only on the variability accounted for by the components. In addition, leave-one-out crossvalidation was performed for the discriminant analysis based on $K = 2, \ldots, 12$ components and the false positives and false negatives

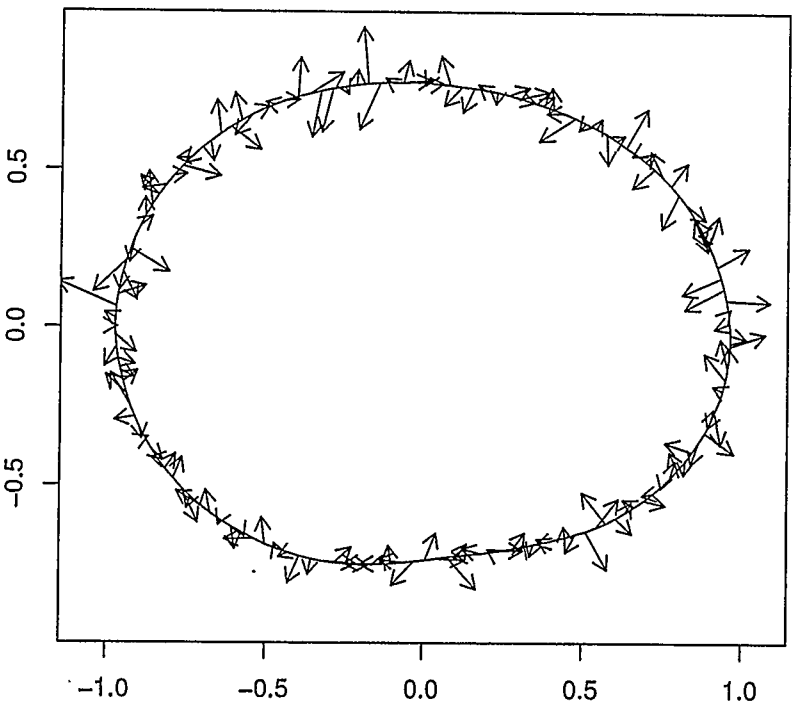


Figure 3.4: Discriminant directions for the naive approach

calculated for each. The summary for each of these values of K is shown in Table 3.1.

The discriminant is created based on the first 6 components and the obtained matrix of weights, the α's, is applied to the matrix of the corresponding rotations of the 6 chosen components and then rearranged in the X, Y coordinates.

Figure 3.5 shows the normal mean profile with the arrows pointing in the direction of discrimination. This means that profiles following, on average, the shape formed

K	2	3	4	5	6	7	8	9	10	11	12
False positives	12	10	14	13	16	18	18	18	18	17	17
False negatives	24	26	23	24	20	21	23	23	23	26	21

Table 3.1: Number K of PC's used with corresponding false positives and false negatives

by the tips of the arrows, will have different value for the discriminant than that of the normal mean profile.

Figure 3.6 is a comparison of the use of principal component based discriminant and the discriminant based just on the difference of the group means. This figure shows boxplots for the discriminant scores. Both methods were adjusted to have a mean value of 1 for Lymphoma profiles and a mean value of -1 for normal profiles.

Welch's t test was performed on both methods and the discriminant values obtained from the analysis based on components yielded a significant difference between the means of the discriminant values (p-value < 0.001), while that based on the group means was not significant.

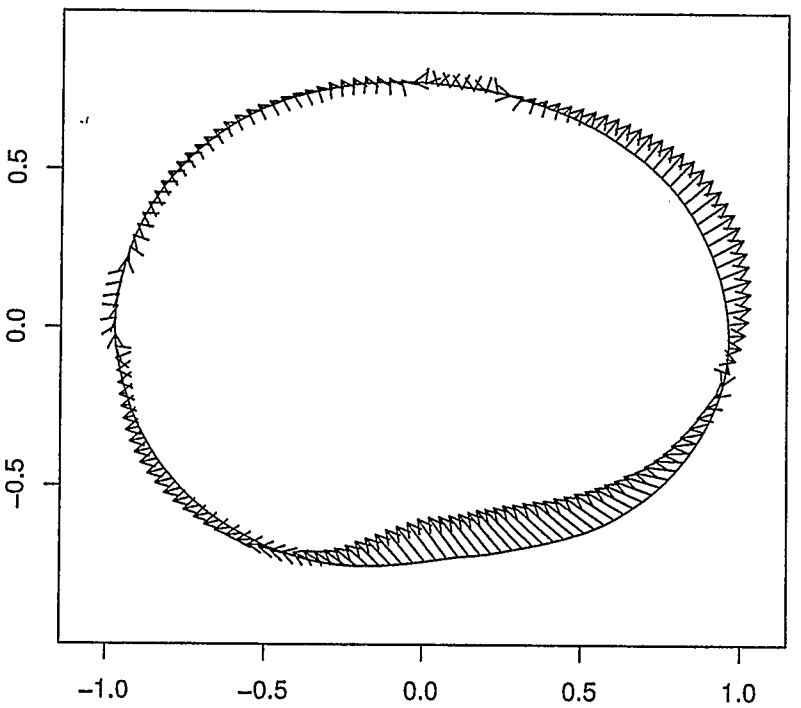


Figure 3.5: Discriminant directions for the PC approach

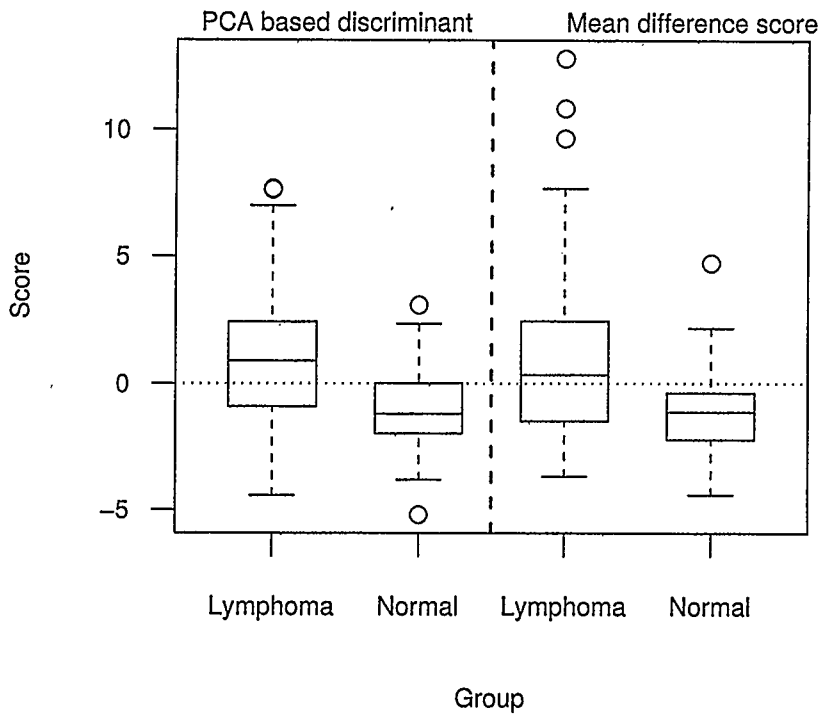


Figure 3.6: Comparison of methods for discriminant values

Chapter 4

Basis function expansions

In the previous parts of this thesis we created each functional datum via linear interpolation between the observed values of the profiles. We were able to perform principal components analysis and linear discriminant analysis directly on the chosen points using the existing MDA approach, as there was no smoothing performed on the profiles. It is worth mentioning that performing the smooth functional version of any multivariate analysis involves more than performing MDA and then smoothing the results. Ramsay and Silverman [29] and Green and Silverman [11] discuss the subject. Previously we had no smoothing and hence the results followed in a rather straightforward form.

In this thesis we are interested in variability at different scales. So far, the analysis has been concerned with overall shape and the analysis was performed by first having to insure that extraneous variability was not introduced because of rotation or size effects. We want to study the variability of the profiles at the level of their derivatives, that is to study the speed at which the border of the profiles changes and compare measures of curvature. It is our assumption that a normal cell will tend to have a smooth nucleus and will have smaller total curvature measurements than that of a malignant one which we assume will tend to be a "squiggly" nucleus; this curvature can be measured locally.

On the other hand, two nuclei that are smooth but whose profiles differ in their roundness should also yield different speed for profile border change and different cur-

vature. For example, taking the first profile from the group of normal T-Lymphocytes and the seventh from the Lymphoma group (see Figure 4.1), it is clear that the nucleus that does not "cut" into itself will have a total sum of local curvature smaller than the Lymphoma one that is shaped like a croissant.

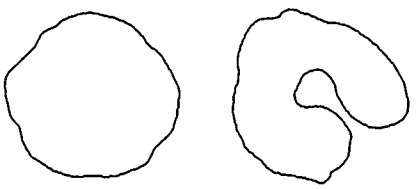


Figure 4.1: Examples of normal and malignant profiles

Now we are talking about curvature and therefore we are not only thinking of continuous functional forms, but of basically smooth continuous functional data. We create each functional datum as a smooth function of the observed points via basis functions expansion such as B-splines or Fourier expansions.

With this in mind, we would like to examine the behaviour of the coordinates with respect to time. We start by graphing the raw discretised data as a function of time. Here we are assuming the data were taken at equal time intervals, that is to say that the first point in the data will be considered to happen at time $t_0 = 0$, the second point at $t_1 = 1/n$ and so on until the last point in the data was taken at time $t_n = 1$. It is important to point out that the number of points per profiles are not the same and that although the time points are spaced equally over $[0, 1]$ this does not mean that they represent equally spaced points.

Figure 4.2 shows one of the normal profiles and the X and Y coordinates as

functions of time. Figure 4.3 shows one of the malignant profiles and its X and Y coordinates as functions of time. In both graphs, the solid line is for the $X(t)$ function and the dashed line is the $Y(t)$.

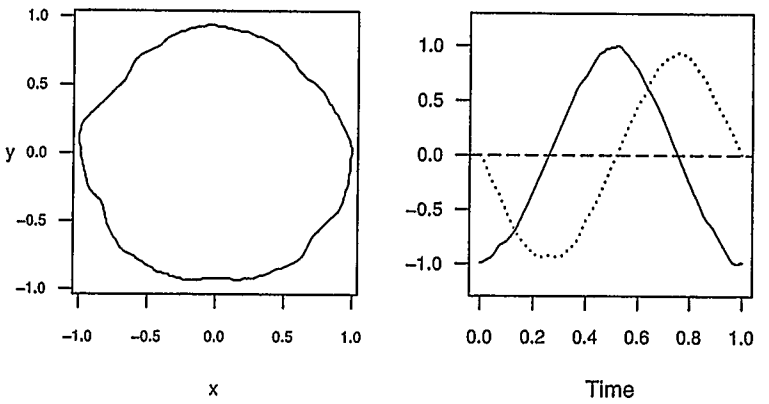


Figure 4.2: Normal profile with coordinates as a function of time

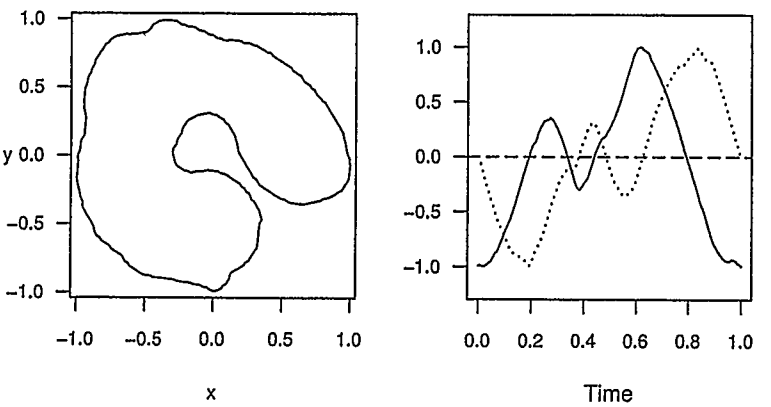


Figure 4.3: Malignant profile with coordinates as a function of time

These graphs present what could be taken as extreme cases, but are illustrative of the kind of variation that is observable. Our emphasis is on trying methods that will measure local variability, given that, as Peura and Iivarinen discuss [25], some

known descriptors, such as convexity ratio A/A_{CH} where A is the area of the planar object and A_{CH} is the area of the convex hull, prove not to be useful in distinguishing a planar object with a smooth boundary from another with irregular boundary if both happen to be non-convex.

In order to try to perceive, if only graphically, differences between types of profiles, and also to aid in the decision of which type of basis functions to use for the basis function expansion, we present in Figure 4.4 the data for all the profiles' raw data and their corresponding coordinates as functions of time.

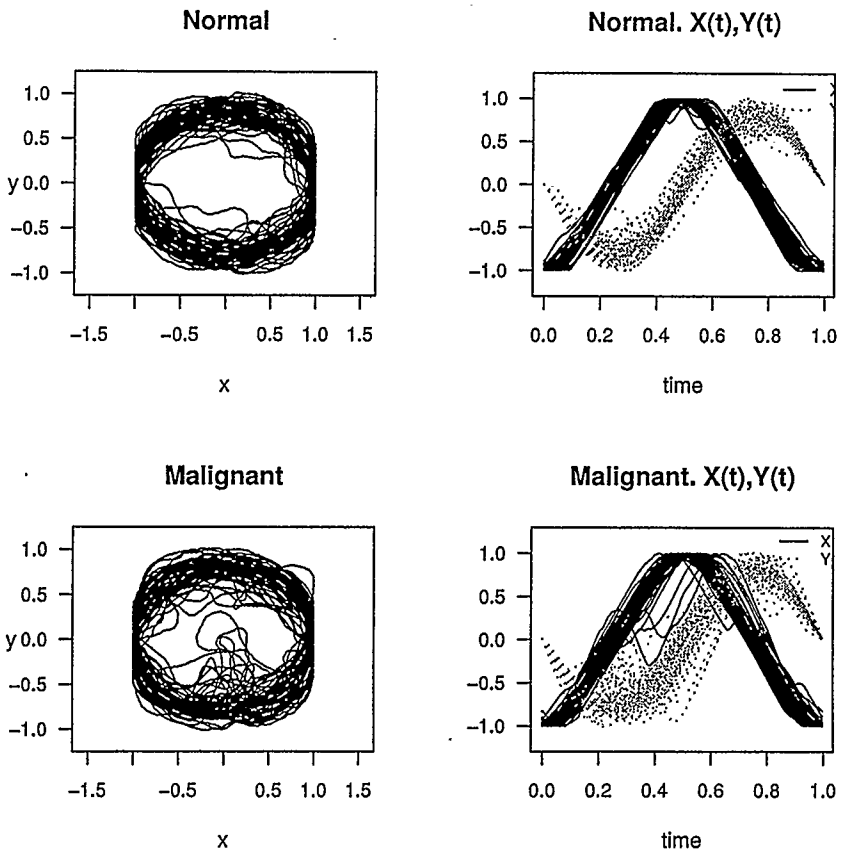


Figure 4.4: Raw profiles and their coordinates as function of time

The nuclei are, by nature, closed curves and hence cyclic; this would suggest the use of Fourier series expansion for the profiles. In Figure 4.4 we observe that in the $X(t), Y(t)$ graphs for the malignant profiles, both coordinates seem less regular and are prone to greater phase shifts than for the normal profiles.

4.1 Creating the functional data

Recall that transforming the discrete data, say z_i, into functional form ($x(t)$) involves representing the function by a linear combination of a fixed number K of known basis functions, usually denoted by ϕ_k,

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t). \tag{4.1}$$

In the creation of the functional data from the discrete observed values via basis functions, we have the option of smoothing the function. The choice of type of basis functions and the number of basis functions on which to base the transformation has as a consequence smoothing. Ramsay and Silverman [29] and Efromovich [8] discuss the difference between interpolating with smooth functions and smoothing. In our case, if we are to take each point in the profile for our basis expansion and pass through them, we will introduce no smoothing, only interpolation. On the other hand if we base our approximation on fewer points than the total, or on $K < n$ basis functions we might miss getting exactly the observed values, hence smoothing. If we have n points and no smoothing is wanted, by using $K = n$ in the linear combination 4.1 it is possible to choose the coefficients c_k to yield $x(t_j) = z_j$ for each j.

The choice of number and type of basis functions is important for the estimation of the derivatives of the function. Bases that work well on representing accurately the observations may push the estimated $x(t)$ to have high frequency oscillations.

Fourier series expansion, where the basis functions ϕ_i are given by

$$\begin{aligned} \phi_0(t) &= 1 \\ \phi_{2r-1}(t) &= \sin r\omega t \\ \phi_{2r}(t) &= \cos r\omega t \end{aligned}$$

and where ω determines period $2\pi/\omega$, yield a simple derivative estimation since the coefficients of the derivative function can be found by multiplying individual coefficients c_k by powers of the period-argument and the appropriate change of signs and interchange of sine and cosine function

$$\begin{aligned} \frac{\partial}{\partial t}\sin r\omega t &= r\omega \cos r\omega t \\ \frac{\partial}{\partial t}\cos r\omega t &= -r\omega \sin r\omega t \end{aligned}$$

Thus the first derivative has coefficients $(0, -\omega c_2, \omega c_1, -2\omega c_4, 2\omega c_3, ...)$ and the second derivative has coefficients $(0, -\omega^2 c_1, \omega^2 c_2, -4\omega^2 c_3, -4\omega^2 c_4, ...)$.

This basis generally yields smooth expansions but might be not so good when there are rather strong local features.

In response to the need of reflecting both global and local features polynomial splines were developed [5]. In a nutshell, polynomial splines are functions constructed by smoothly joining polynomials at chosen points called *knots* spread in the interval $[a, b]$ (in our case $[0, 1]$) increasingly. Between adjacent knots the spline is a polynomial of fixed degree, but at the knots where the polynomials meet they are required to match in the values of a fixed number of derivatives. A spline of degree 0, or order 1, is a step function and hence discontinuous at knots; a spline of degree 3, order 4, is piecewise cubic with continuous second derivative. B-splines are versions

of splines defined on a smaller support, or truncated over certain values, usually the interval $[t_{j-2}, t_{j+2}]$.

A useful approach for creating the expansion is the penalised smoothing approach [11, 28, 29]. The penalised smoothing yields a functional datum from discretised data, but instead of basing the fit only on minimising the sum of squares of the fit, a penalty is added to the object function; the idea behind penalising is to control roughness in the resulting function. One of the most used penalties is based on the integrated norm of the estimate of the second derivative of the function, that is the curvature of the function. There is a parameter λ to regulate the degree of this penalisation. Formally we want to minimise the penalised sum of squares $PENSSE$:

$$PENSSE_\lambda(x|y) = \sum_j \{y_j - x(t_j)\}^2 + \lambda \|D^2 x\|^2 \tag{4.2}$$

where the penalisation is the integrated square second derivative of the function $x(t)$, formally $\|D^2 x\|^2 = \int \{D^2 x(s)\}^2 ds$ and the operator D^m is the m^{th} derivative operator.

If the penalisation parameter $\lambda \to \infty$ then we end up with a linear fit; if $\lambda \to 0$ we end up with interpolation. As Ramsay and Silverman mention [29], even in the case where $\lambda \to 0$, the resulting function is the smoothest twice-differentiable curve that fits the data.

Since we are interested in the variability of derivatives, and we are assuming that normal and malignant profiles differ on their borders locally, we are not interested in smoothing the data too much (at first) and will not penalise the fit. By not penalising the fit we will try to preserve the variability in the curvature rather than

smoothing it out. We will create the functional data based on all the observed points for each profile, but with a fixed number of basis functions to be consistent in the approximation.

We created the functional data based on a Fourier expansion with 17 basis functions. We use 17 bases in order to capture the local variability and approximate the observed data closely. Recall that the Fourier approximation would be:

$$\hat{x}(t) = c_0 + c_1 \sin \omega t + c_2 \cos \omega t + c_3 \sin 2\omega t + c_4 \cos 2\omega t + ... \tag{4.3}$$

Figures 4.5 and 4.6 show the approximation applied to two normal and two malignant profiles. The difference between the observed data and the approximation is not visible.

If we assume the profiles to be ellipses, then we would expect the first and second coefficients of the Fourier expansion, c_0 and c_1, for X to be relatively close to zero, the third coefficient (c_2 corresponding to the first cosine term) to be negative and close to -1 and also that coefficients of order higher than 3 (c_3, c_4, ...) would tend to be zero or have averages of zero with observed variability due only to noise introduced by digitisation. For the Y coordinates we would expect the second coefficient (c_1 corresponding to the first sine term) to be negative and close to -1, and those coefficients of order higher than 3 to behave as for the X expansion.

Based on the shapes observed in graphing the X and Y coordinates with respect to time and assuming that profiles are basically deformations from an elliptical template, we can expect coefficients of order higher than 3, for the normal profiles, to have averages not so close to zero with small variability or close to zero but with high variability. Such behaviour is also expected for the normal Y. These are expected

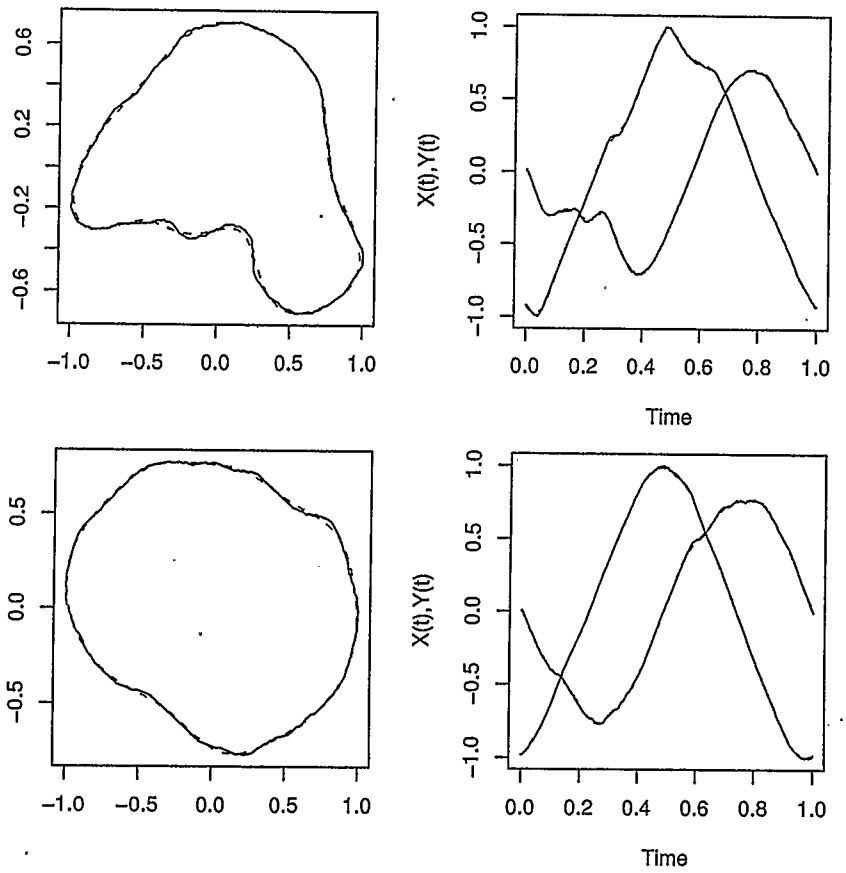


Figure 4.5: Two normal profiles and approximations

because the normal profiles seem to have a shape that is closer to elliptical and seem to experience less local variability than malignant profiles do. In the case of the malignant profiles we expect that coefficients of order higher than 3 tend to be greater in magnitude than those for the normal profiles due to the presence of concavity and tendency to cut into themselves (found in the PCA section of this thesis).

The following figures, Figures 4.7, 4.8, 4.9, 4.10 show boxplots of the coefficients for the expansions.

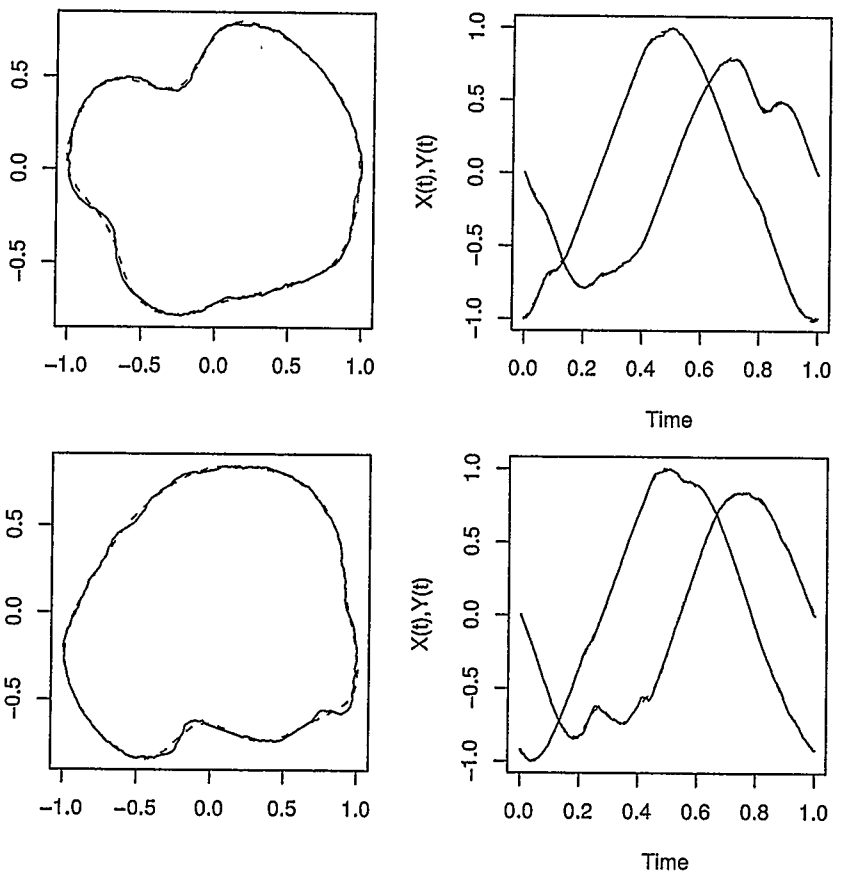


Figure 4.6: Two malignant profiles and approximations

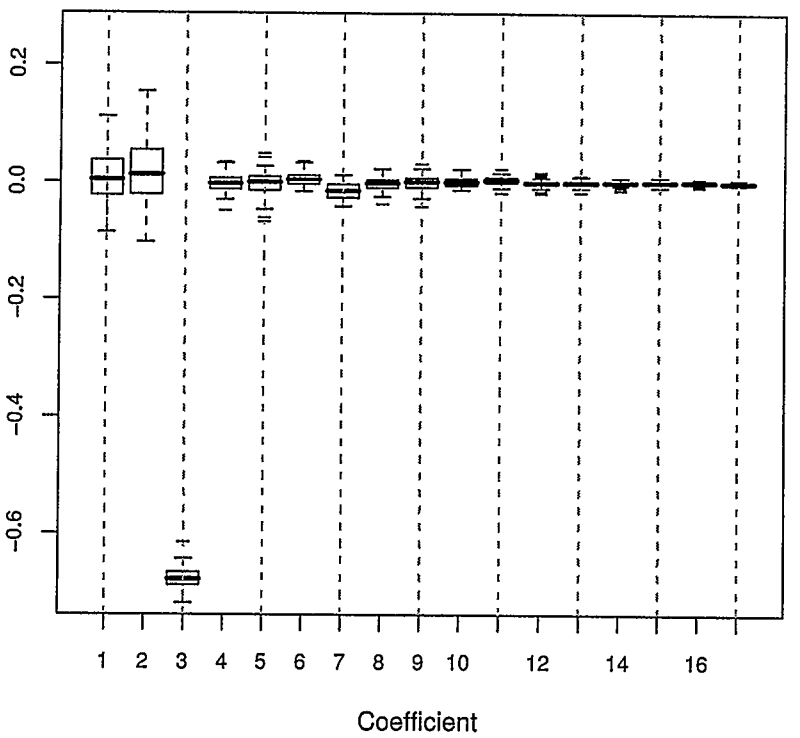


Figure 4.7: Boxplots of coefficients for X in normal profiles

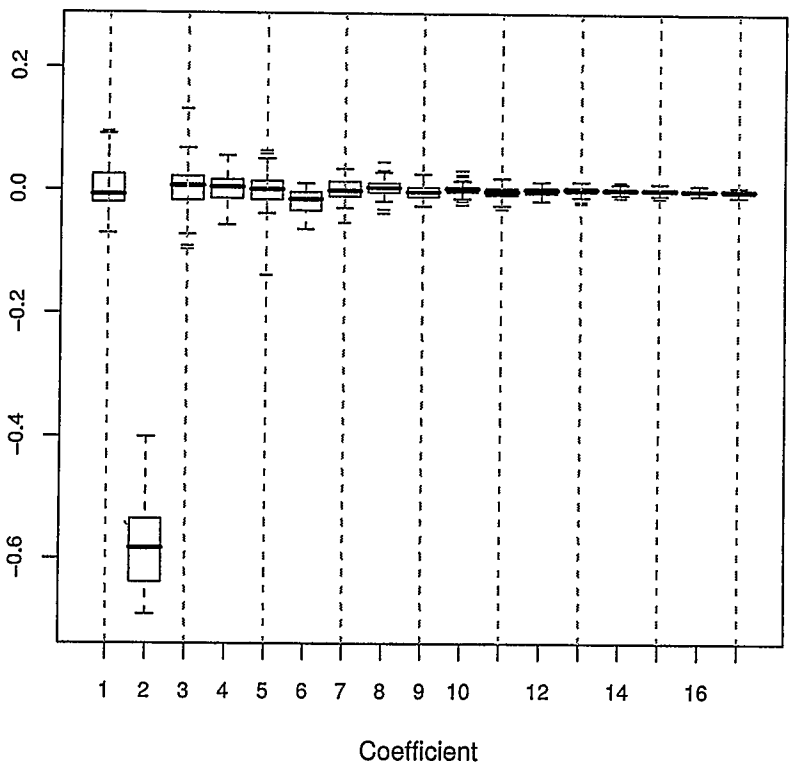


Figure 4.8: Boxplots of coefficients for Y in normal profiles

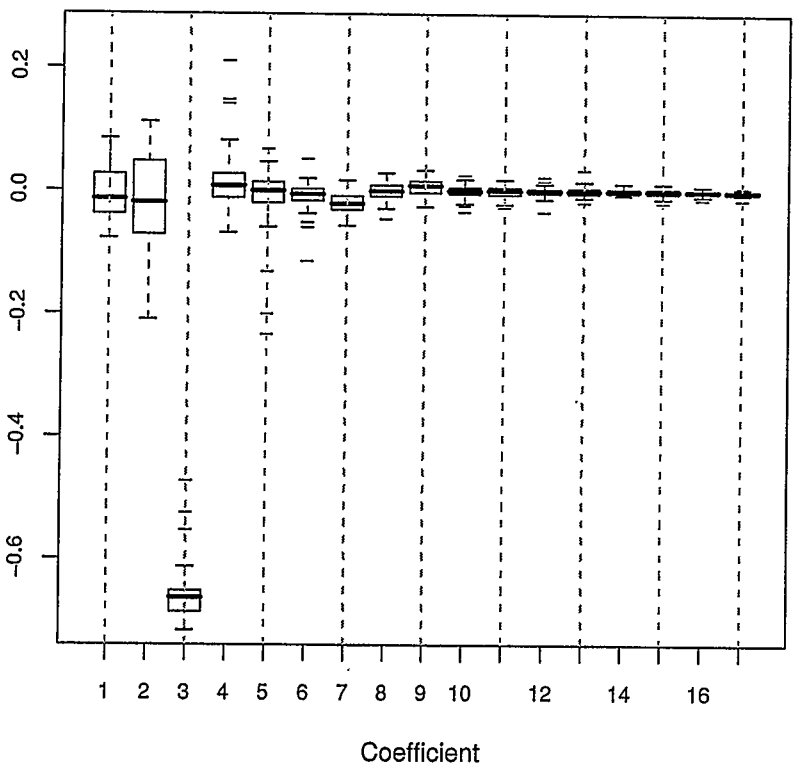


Figure 4.9: Boxplots of coefficients for X in malignant profiles

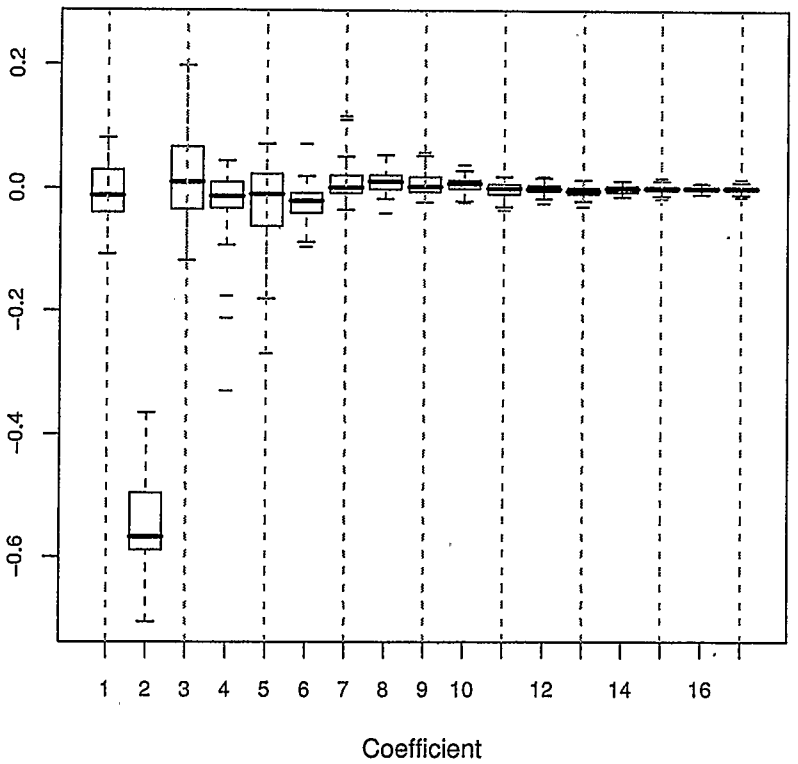


Figure 4.10: Boxplots of coefficients for Y in malignant profiles

Figures 4.7, 4.8, 4.9, 4.10 show that for the normal profiles the coefficients of order higher than seven start being concentrated around zero, for both the X and the Y coordinates. For the malignant profiles the coefficients start dampening at the ninth or tenth coefficient for X and Y.

Having the boxplots for the first three coefficients together with the boxplots for the higher order coefficients obscures the variation of the higher order coefficients. Therefore we exhibit Figures 4.11, 4.12, 4.13, 4.14 without them. Wilcoxon's tests

were performed to test if the mean of coefficients of higher order is significantly different from zero.

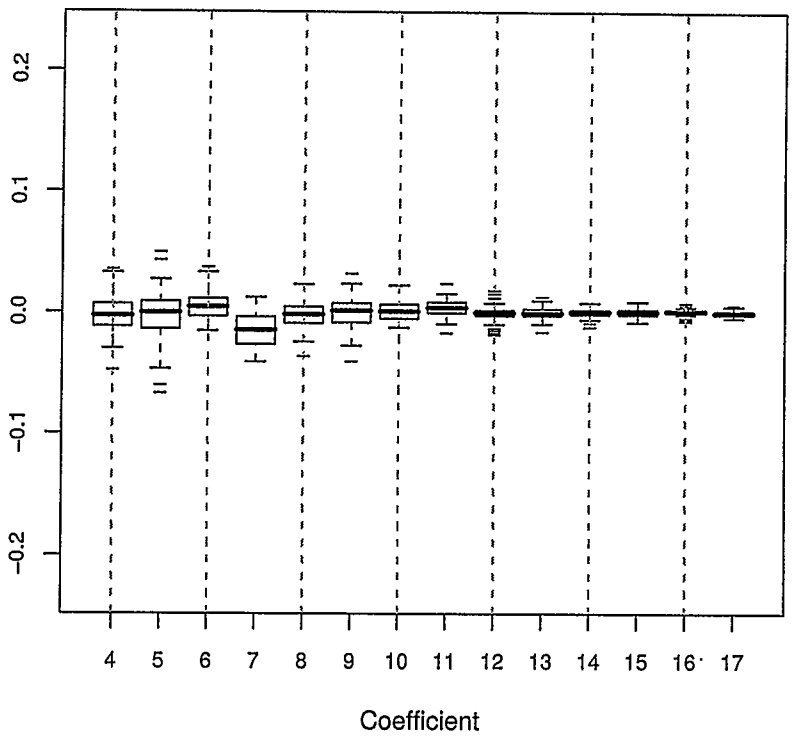


Figure 4.11: Boxplots of order > 3 coefficients for X in normal profiles

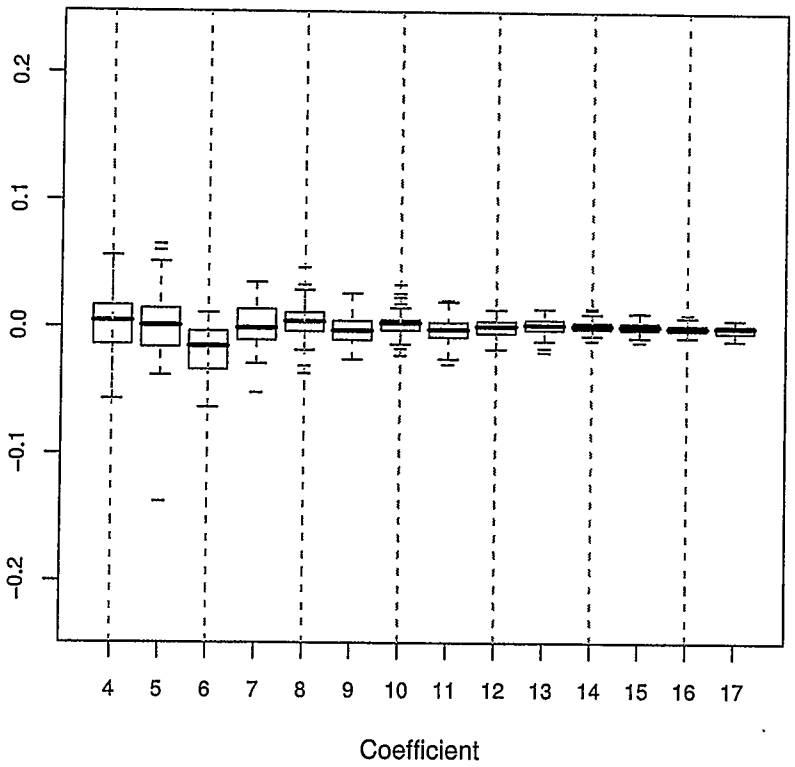


Figure 4.12: Boxplots of order > 3 coefficients for Y in normal profiles

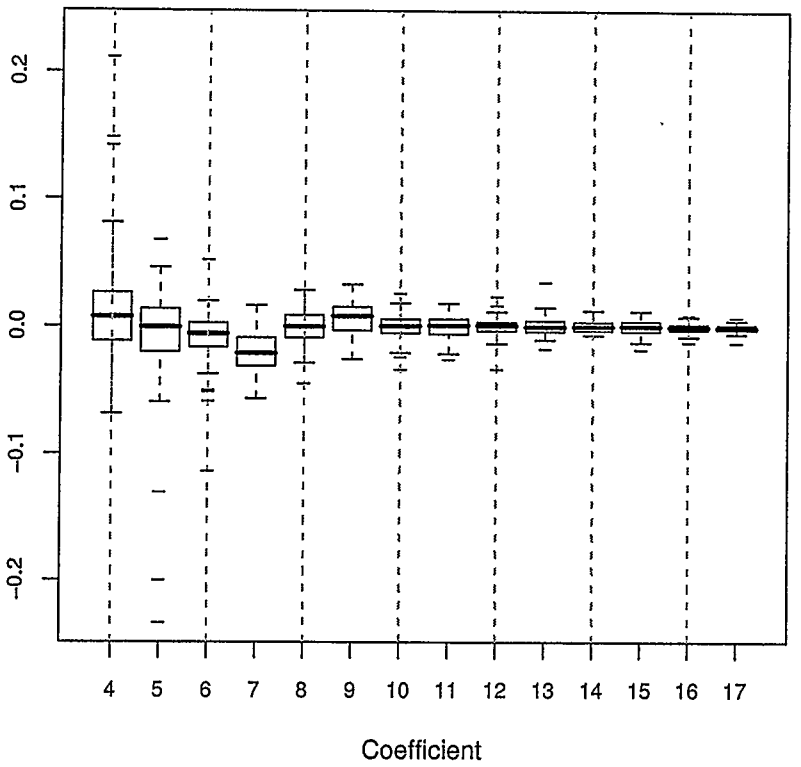


Figure 4.13: Boxplots of order > 3 coefficients for X in malignant profiles

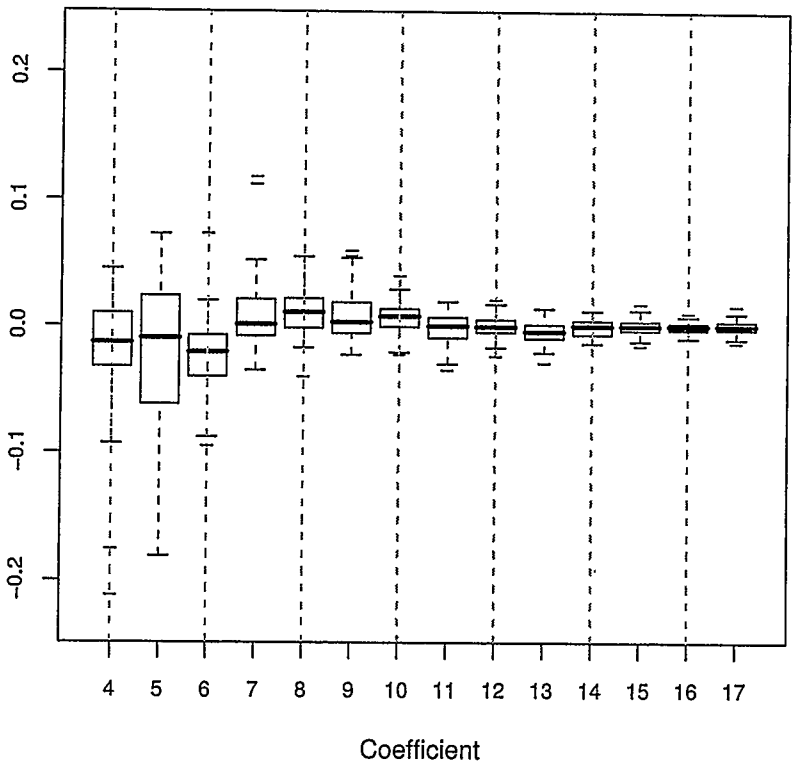


Figure 4.14: Boxplots of order > 3 coefficients for Y in malignant profiles

	X normal	Y normal	X malignant	Y malignant
c_0	0.40	0.79	0.26	0.27
c_1	0.05	0.00	0.15	0.00
c_2	0.00	0.39	0.00	0.13
c_3	0.19	0.59	0.09	0.01
c_4	0.37	0.92	0.64	0.04
c_5	0.06	0.00	0.01	0.00
c_6	0.00	0.71	0.00	0.17
c_7	0.26	0.04	0.93	0.00
c_8	0.72	0.14	0.00	0.02
c_9	0.56	0.02	0.67	0.00
c_{10}	0.00	0.22	0.55	0.64
c_{11}	0:34	0.74	0.45	0.96
c_{12}	0.62	0.02	0.58	0.00
c_{13}	0.99	0.02	0.30	0.78
c_{14}	0.32	0.08	0.86	0.58
c_{15}	0.04	0.64	0.95	0.28
c_{16}	0.43	0.69	0.45	0.21

Table 4.1: P-values for Wilcoxon's test

Table 4.1 shows the p-values from Wilcoxon's test for $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ for the Fourier coefficients of $X(t)$ and $Y(t)$ for the normal profiles and for the malignant profiles.

We observe that the means of the coefficients of orders higher than 3 are not all zero which is indicative of greater deformations from elliptical templates and indicative of greater local variability existing in the profiles. This local variability would be best captured with a basis expansion based on splines.

Figure 4.15 shows the correlation contour plots of the X coordinates for each of the normal and malignant profiles, Figure 4.16 shows correlations for the Y coordinates and Figure 4.17 shows the crosscorrelation between X and Y for each.

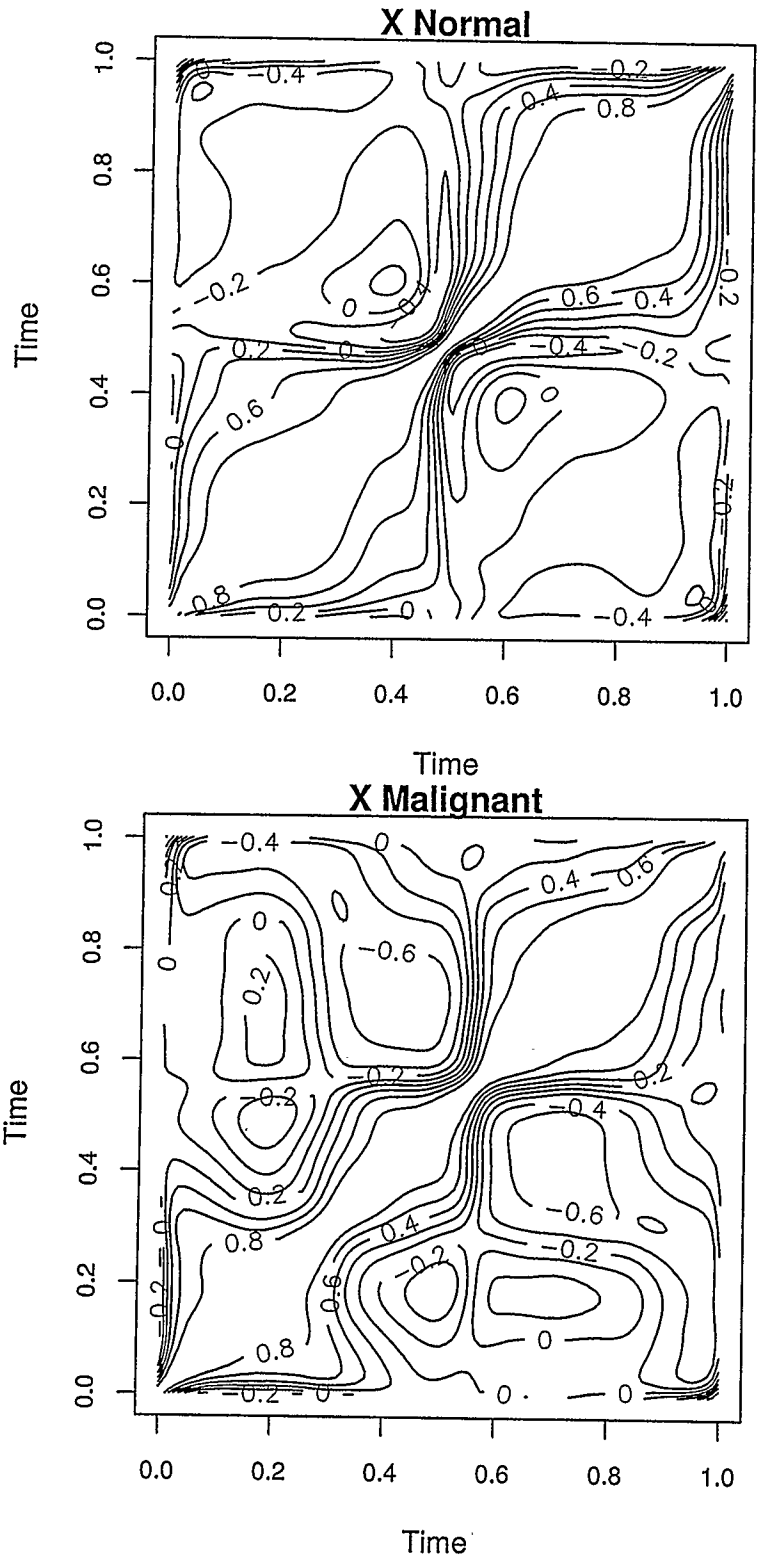


Figure 4.15: Correlations of X for normal and malignant profiles

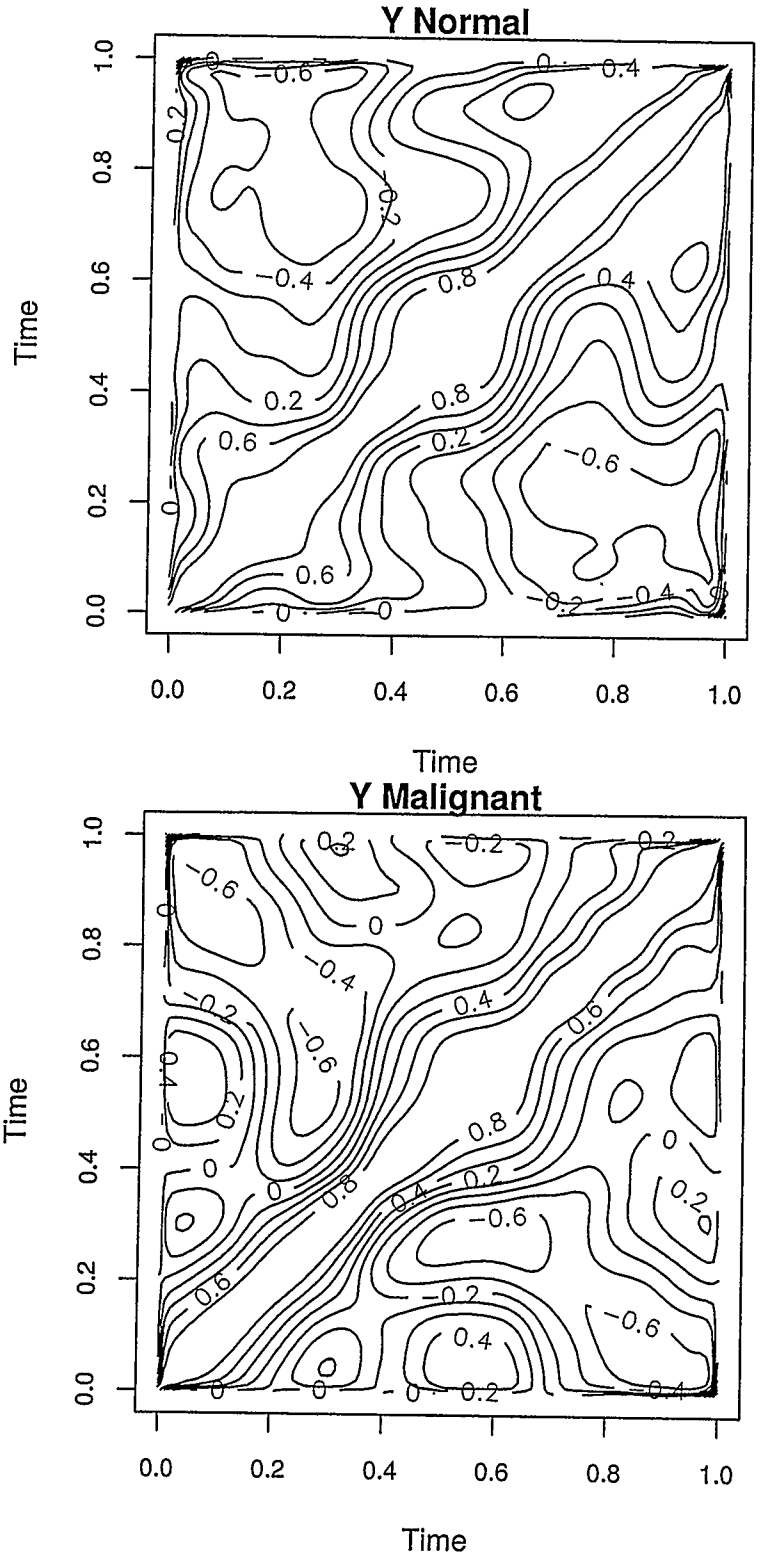


Figure 4.16: Covariance of Y for normal and malignant profiles

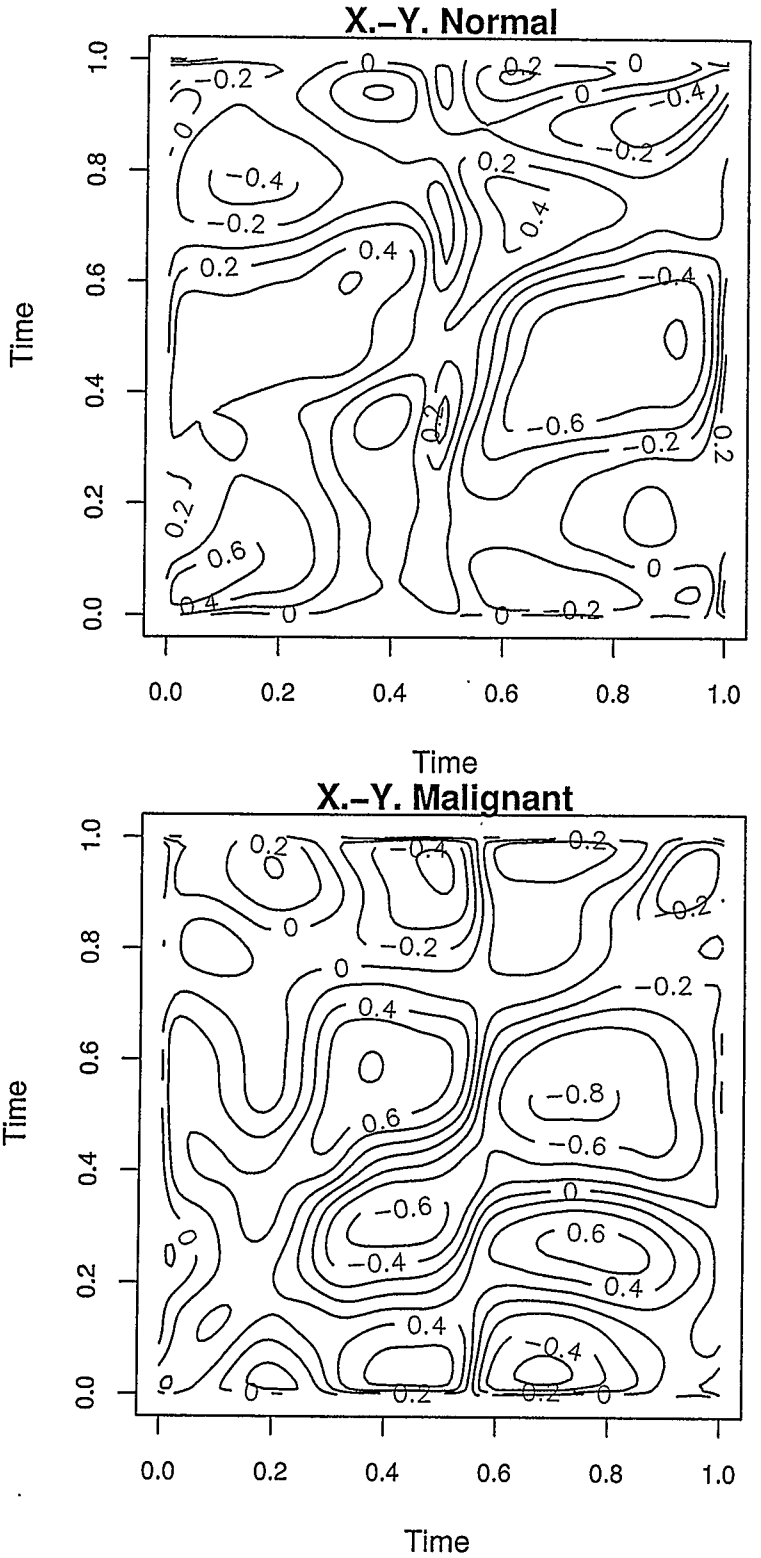


Figure 4.17: Cross-covariance of X and Y for normal and malignant profiles

The normal profiles' $X(t)$ present high correlation along small steps off the diagonal times for times less than 0.45 or greater than 0.55. Around the $0.45 - 0.55$ range the correlation in the smallest step from the diagonal rapidly decreases. This type of correlation is expected and is explained by two facts:

- The starting time t_0 was taken to be the minimum x value along the semimajor axis and then we moved counterclockwise, hence the first half of the time $X(t)$ is increasing to close to its maximum. One quarter of the way we are still in negative values, closing in on 0 still increasing. Given that we have the profiles resting on their semimajor axis, we have the variable increasing for half of the time. Once we move away from the maximum x it starts decreasing. That is why around $t = 0.5$ the correlation is so tight. If we compare the values of the x at times $t_j = 0.2$ and $t_i = 0.7$ we are comparing negative values with positive values that have not much in common.

- The normal profiles tend to follow a shape closer to that of an ellipse and hence we observe the symmetry not only with respect to the line extending from $(0, 0)$ to $(1, 1)$ but close symmetry with respect to the line extending from $(0, 1)$ to $(1, 0)$ as an ellipse would have.

In the X for the malignant profiles the correlation contour plot appears to have three sections along the diagonal, instead of having two as does the normal profiles' correlation. This is indicative of the fact that malignant profiles tend to experience concavity and therefore the $X(t)$ seems to be closing in into the area of small change, but then it increases again. The fact that there is no symmetry about the $(0, 1)$ to $(1, 0)$ line indicates that the X is not behaving the same way before and after time

0.5.

Regarding Y for the malignant profiles the correlation contour plot appears asymmetric with respect to the line spanned between $(0,1)$ and $(1,0)$. The correlation plot for normal profiles appears symmetric about the such line. This is indicative of the fact that malignant profiles tend to experience concavity and given the orientation of the profiles, $Y(t)$ increases and then decreases rapidly.

The cross-correlation plot of $X-Y$ is difficult to interpret, but the variability in the plot corresponding to malignant profiles is greater than it is in the plot for the normal profiles.

4.2 Using the functional data

The functional data, as constructed with Fourier or splines basis functions, enables us to analyse the total curvature in a profile, as we now have twice differentiable data. If we had univariate data we could calculate the integral of the second derivative as our estimate of curvature. This would amount to calculating the default penalisation term when smoothing is performed.

In our case we are defining the profiles as bivariate functional data that describe a planar object. We are looking at the profiles in terms of the arc length s parameterised by t. We have $X(t), Y(t)$ defining the profiles. The construction of the functional form of the data guarantees that the obtained planar curve (profile) is closed and twice differentiable [29]; hence we can use known calculus results to express the profiles' shape and curvature.

Let Z be a profile in $\mathcal{R}^2$, parameterised by t in such a way that $Z(t) = (X(t); Y(t))$

for $0 \leq t \leq 1$. The arc length measured from our starting point t_0 to a point $Z(s)$ in the profile is

$$S = \int_0^s \left(X'(t)^2 + Y'(t)^2\right)^{1/2} dt$$

and hence

$$\frac{dS}{dt} = \left(X'(t)^2 + Y'(t)^2\right)^{1/2}$$

Now the curvature $\kappa(t)$ at some point t in the curve is:

$$\kappa(t) = \frac{X'(t)Y''(t) - X''(t)Y'(t)}{(X'(t)^2 + Y'(t)^2)^{3/2}} \tag{4.4}$$

And the total curvature $Curv(Z)$ of the planar profile takes the form:

$$Curv(Z) = \int_Z |\kappa(s)| ds = \int_0^1 \left| \frac{X'(t)Y''(t) - X''(t)Y'(t)}{X'(t)^2 + Y'(t)^2} \right| dt \tag{4.5}$$

For the calculation of curvature there is no need for registration or alignement of the data since the integration is over the entire C^2 curve.

We assume that the normal profiles will tend to have smaller values for the total integrated curvature since we are assuming that they will tend to be convex and with smoother borders. The malignant profiles are assumed to be nonconvex more often than the normal profiles and also exhibit more variability locally in the borders.

We are interested in testing the hypothesis

$$H_0 : \mu_{Curv(z),\text{Normal}} = \mu_{Curv(z),\text{Malignant}}$$

against the hypothesis

$$H_1 : \mu_{Curv(z),\text{Normal}} < \mu_{Curv(z),\text{Malignant}}$$

Performing Welch's T test we conclude that the curvature is significantly smaller for the normal profiles than it is for malignant ($p = 0.00029$) and performing Wilcoxon's rank sum test takes us to the same conclusion ($p = 0.00067$). Figure 4.18 shows the density estimates for the curvatures of profiles and Figure 4.19 shows boxplots for the curvatures.

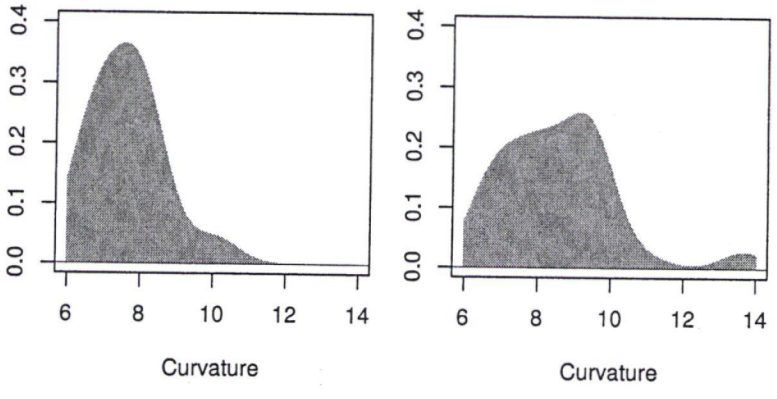


Figure 4.18: Density estimates for the curvatures of normal (left) and malignant (right) profiles

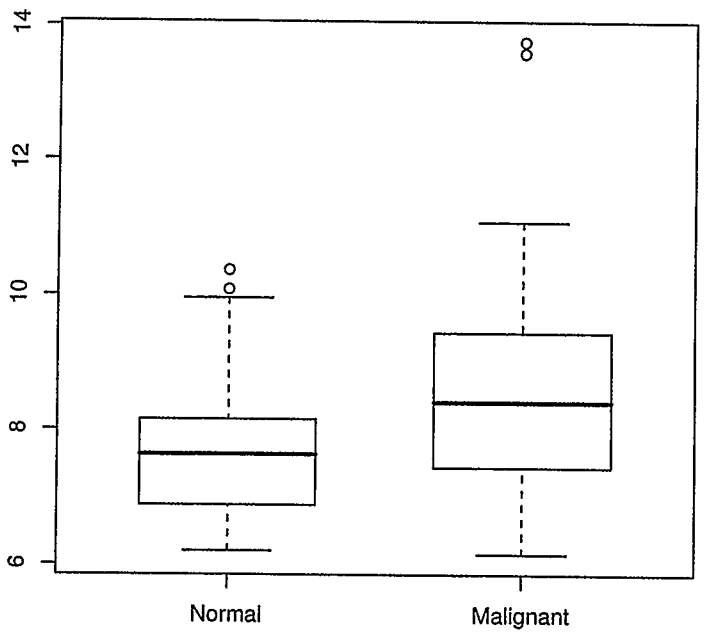


Figure 4.19: Boxplots for the curvatures of normal and malignant profiles

Chapter 5

Principal Differential Analysis

The aim of this thesis has been to find a process that will help us in classifying profiles into malignant or benign types. Moreover, we want to be able to provide some uncertainty measurement or assessment of this classification. Such a procedure should not only characterise the existing profiles, but be able to classify new profiles.

Hobolth and Jensen [14, 16] assume in their modelling, and conclude in their results, that malignant and benign profiles differ in the amount and type of variability or deformation from the templates. They also show that local variability plays a significant role in the shape of the profiles [15]. This local variability, as measured by assessing curvature in the previous chapter, has proved to be significantly different in the two groups.

A procedure that can be used to try to classify a new profile consists of the following steps:

- Create the empirical cumulative distribution of curvature for the normal group and for the malignant group.
- Calculate the curvature of the profile via the bivariate functional form as done previously.
- Based on the empirical distributions, observe where the curvature for the new profile stands and calculate the probability of having a higher(lower) value in the normal and in the malignant group.

The purpose of this procedure is based on the findings and assumptions that normal profiles have lower curvature values than those of the malignant profiles. A profile that is malignant would yield a low estimated probability of being larger than the observed value under the distribution for the normal profiles. It would yield a large probability of being greater than the observed value under the distribution for the malignant.

If we have a new set of profiles coming from one tissue, we can calculate the mean of the observed data and perform Hypothesis testing against the means of normal profiles and against the mean of malignant profiles. Given the variability in the curvatures of the two types, this would be subject to a large Type I or Type II error.

We observed, at the time of creating the functional data via Fourier series, that the data have high variability at local levels. Recall that if the profiles had been perfectly smooth ellipses, then we would have observed coefficients of higher order than three to be exactly zero. This was expected as the parameterisation of an ellipse in polar coordinates is well known to be

$$x(t) = a\cos t$$

$$y(t) = b\sin t$$

for t in $[0, 2\pi]$. Hence the corresponding coefficients would necessarily be, for the $X(t)$ series: $c_0 = 0$, $c_1 = 0$, and $c_2 = a$ and the rest of the coefficients would be zero; as for the $Y(t)$ series, the coefficients would be $c_0 = 0$, $c_1 = b$, and $c_2 = 0$ and the rest zero.

In the case of small deformations from the elliptical template we would expect the coefficients of high frequencies to be nonzero but small. Observing the boxplots

from the previous chapter, we perceive that there is variability attributable to low frequencies and hence we detect that there is more structure than just that of periodical or sinusoidal nature in the X, Y coordinates. There is, apart from the sinusoidal structure, what may be considered a residual process.

At this stage we would like to assess and possibly characterise the underlying structure of both the normal and malignant profiles and also to compare their residuals. We would like to be able to decide whether a profile comes from a malignant or a normal nucleus based on the underlying structures of the $X(t), Y(t)$ functions. We want to determine the structure in an objective and reproducible way for each type from the profiles we have. With this, we will be able to test a new profile and determine where it 'fits'.

The variability structure of the coordinates can be assessed by the behaviour of their derivatives and the relationship between different orders of derivatives. Borrowing concepts from the differential equations world, we can define a Linear Differential Operator (LDO) that determines the relationships between the derivatives of different orders. This LDO annihilates the primary structural form of the functions.

We use the notation:

$$D^m x(t) = \frac{\partial^m x}{\partial t^m}$$

for the mth derivative of the function $x(t)$ where D is the derivative operator and when $m = 0$ then we have the identity, mainly $D^0 x = x$

In this way, define a Linear Differential Operator by:

$$L = \sum_{j=0}^{m} \beta_j D^j \tag{5.1}$$

In the functional case, β_j is a function $\beta_j(t)$.

To determine the structure of the functions we want to find a linear differential operator such that $LX_i(t) = 0$. In reality, this would be very difficult given the local variability of the profiles. Hence, instead of assuming a homogeneous differential model, we assume a more realistic model, a nonhomogeneous system where there exists a forcing function, say $\alpha(t)$, and also some error structure : $LX_i(t) = \alpha(t) + \epsilon_i(t)$. If the LDO captures most of the structure we expect the error terms to oscillate very closely around zero. Given a specified LDO, we expect the normal profiles to have weight functions $\beta_i(t)$ different from those of the malignant ones. Moreover, we expect that the weight functions will be characteristic of the type of profile. The residual processes or error functions should oscillate around zero, assuming that their LDO's are a good fit, for both types of profiles.

Once the full form of the LDO is determined, and having these being characteristic of each type, we want to determine the group to which a new profile would belong. We do this under the assumption that applying the weight functions for the normal profiles to a normal profile will result in a "nice" residual process, whereas applying it to a malignant profile will give erratic residuals. This procedure should yield similar results if we are to apply the malignant weight functions to the normal and malignant types of profiles.

This type of analysis can be related to that of linear regression applied to two different populations to give a more intuitive idea of our analysis. In linear regression, when having two populations or two regions, say A and B, the linear relationship of a set of variables $X_1, X_2, \cdots, X_p$ to a given response Y in each region is fitted by the linear model (in matrix form) $\mathbf{Y}^k = \mathbf{X}^k\beta^k + \epsilon^k \quad k = A, B$

The separate fitting of β's assumes that the X_j variables in region A relate to the

response Y differently from the way in which the corresponding variables X_j relate to the response in region B. In this type of model, the error structure for regions A and B is assumed to be Gaussian with mean zero and constant variance for each region. A good fit of the data should yield residuals that satisfy the assumptions in each of the regions. Techniques for testing the equality of the p-vector of β's, namely $\beta^A = \beta^B$, are based on F-ratios of the difference of residual sum of squares and their degrees of freedom. Basically, they are based on the concept that the residuals will tend to increase their magnitude and variability if the fit is not good.

Based on this latter concept, and contemplating true difference in the β's, the residuals for region A should be small when the Y^A is well fitted by the corresponding β^A, and should be comparatively larger when fitted by significantly different β's. The β coefficients for each region or population can be thought of being characteristic of each population.

Returning to the functional setup, in order to estimate the weight functions for the operators, we need the data to be aligned or registered to avoid any phase shifts that would introduce exogenous variability to the derivatives and therefore to the estimated structure.

5.1 Registering the data

Registration or alignment of functional data can be performed according to two criteria regarding the target function. In some cases, we might be interested in aligning special characteristics such as local maxima that represent specific characteristics of the process in question. In other cases, we might be interested in a continuous and

overall alignment, that is, avoiding phase and amplitude shifts and trying to have the functions follow a target function as close as possible.

The alignment or registration of the data is based on the creation of a "time warping" function that has the effect of stretching and/or shrinking the time axis so that the values of $X_i(t)$, $X_j(t)$ for $t_{k'} \neq t_k$ align according to some criterion. The time warping function $h(t)$ is then such that $X_i(h(t_k)) \approx X_j(h(t_{k'}))$. This function is constrained to be strictly increasing and complies with having $h(0) = 0$ and $h(T) = T$ where the functional datum is originally defined over $[0, T]$.

The time warping function for the continuous registration is based on minimising a measure of shape similarity, a functional form of sum of squares criterion. Such a measurement is expressed as:

$$FSSE(h, A) = \int_0^{T_0} \{x[h(t)] - Ax_0(t)\}^2 \, dt, \tag{5.2}$$

where h is the time warping function, $x_0(t)$ is the target function, and A is an amplitude factor. If the target curve and the registered curve are very closely proportional then the matrix:

$$\begin{bmatrix} \int \{x_0(t)\}^2 dt & \int x_0(t)x[h(t)]dt \\ \int x_0(t)x[h(t)]dt & \int \{x_0(t)\}^2 dt \end{bmatrix}$$

will be singular and hence only one of its eigenvalues is nonzero. This minimisation can be penalised also and hence we can obtain a smoothed time warping function.

We need to have a target function or curve defined for the registration. For this purpose we calculate the mean of the normal profiles using the normalised data, that is the 150 linearly interpolated values of the $X(t), Y(t)$ functions, based on the

equidistant time points for the rotated and centred profiles.

Figure 5.1, shows the mean curves for the $X(t)$ and $Y(t)$ curves to which we register.

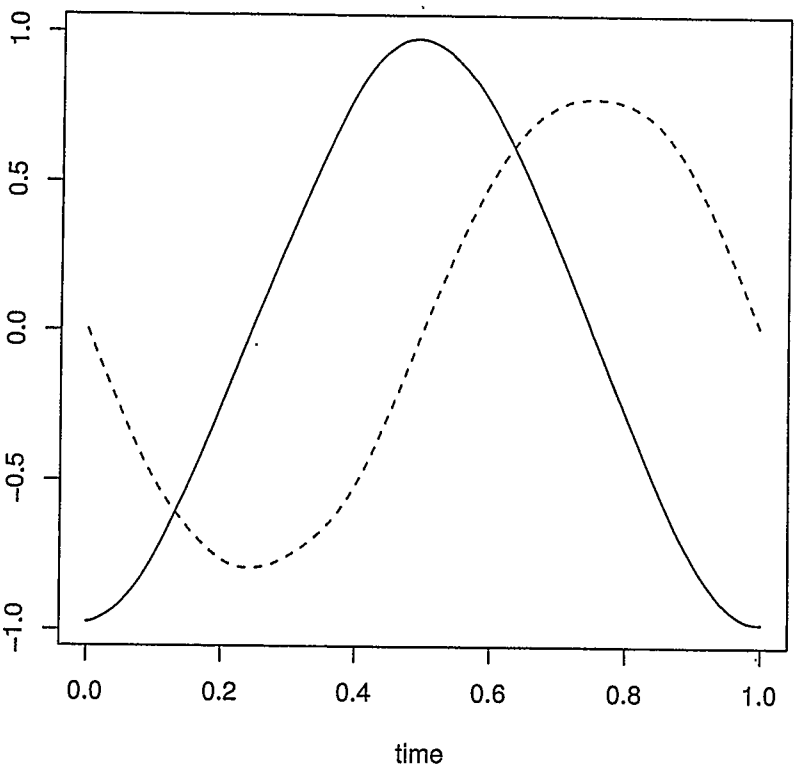


Figure 5.1: Mean $X(t)$, solid, and $Y(t)$, dashed, curves

We register the original rotated and centred data to the mean of the normalised data. Rather than registering the data directly, we perform registration based on the first derivative of the data because the derivatives usually exhibit more variability and they oscillate around zero. Once these are registered we register the data itself

using the time warping function (say $W(t)$) calculated for the derivatives.

One can make the registration process an iterative one by re-calculating the warping function using the mean of the resulting data from the first registration. We iterated a second time with no gain in registration.

Figures 5.2, 5.3 show the $X(t), Y(t)$ data for the normal profiles before registration and the mean for each coordinate. Following these are Figures 5.4, 5.5 that show the registered data and the target function to which they were registered.

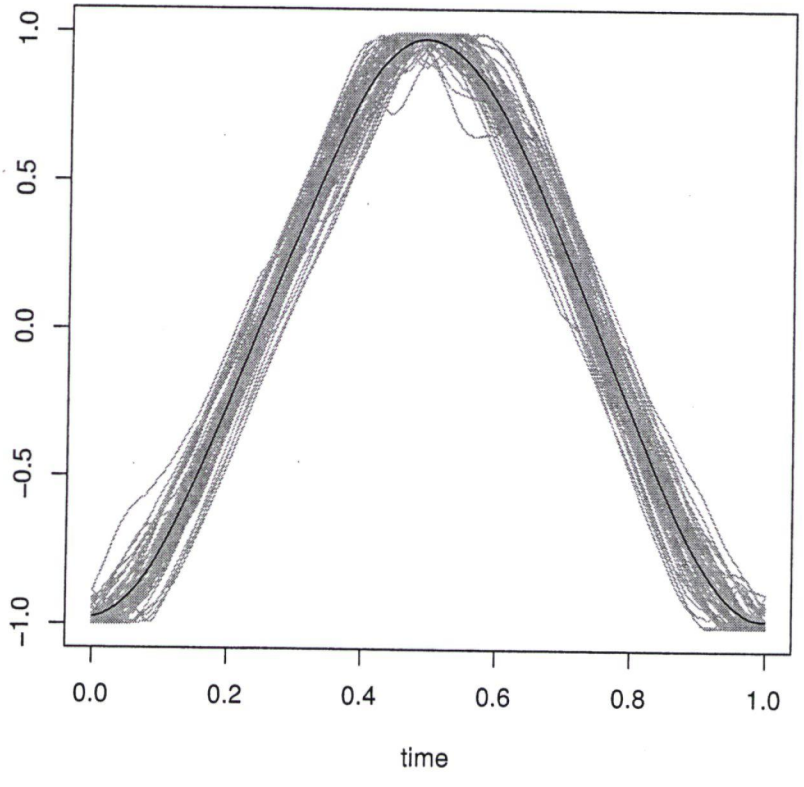


Figure 5.2: $X(t)$, grey, and mean $X(t)$, black, curves

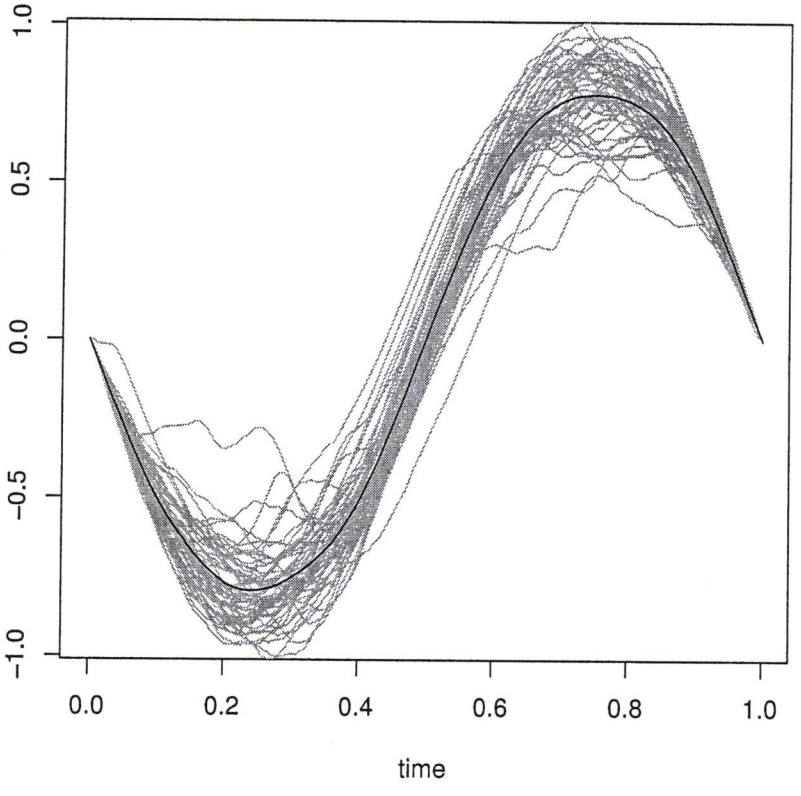


Figure 5.3: $Y(t)$, grey, and mean $Y(t)$, black, curves

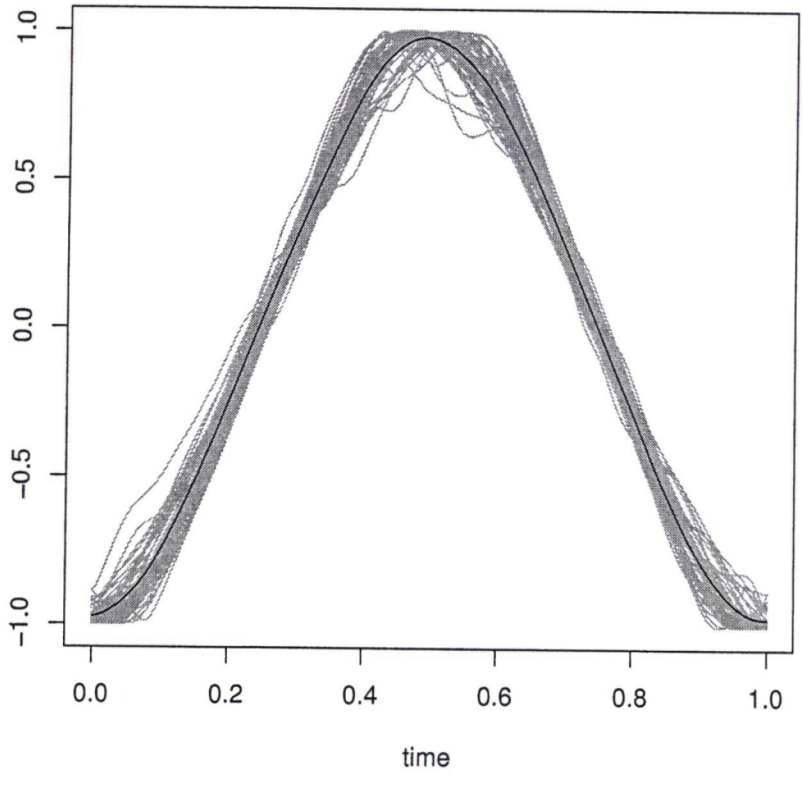


Figure 5.4: Registered X, grey, and mean $X(t)$, black, curves

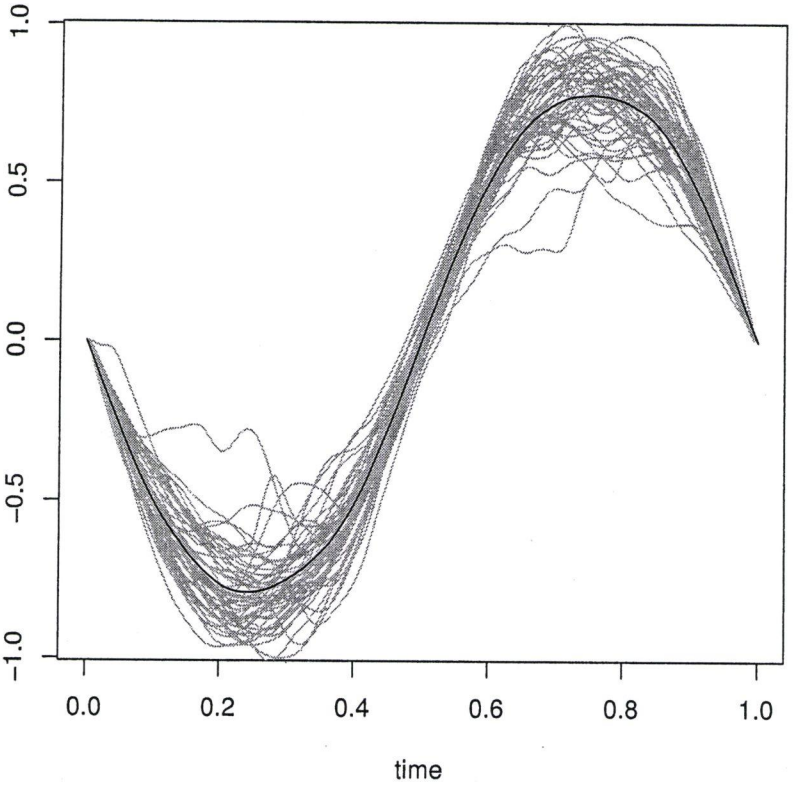


Figure 5.5: Registered Y, grey, and mean $Y(t)$, black, curves

We can observe that there is only a small change in the graphs, and we can observe the change by looking at the different spread on the horizontal. That is to say that we see the horizontal variability to be less, the lines of the 50 profiles are tighter in the registered data. For the purposes of the PDA we registered both normal and malignant to the normal mean profile.

The registration procedure was done for overall shape rather than for landmarks such as the local maxima or minima of the $X(t), Y(t)$ functions because the profiles are not, indeed, perfect ellipses where we would certainly expect $X(t)$ to reach its local minimum at time 0, then reach 0 at time $T/4$, reach its maximum of 1 at time $T/2$, then again 0 at time $3T/4$ and finally the local minimum of -1 at time T. Given the possible nonconvexity of the profiles, we would be forcing the profiles to really change shape. We use the continuous overall registration that tries to align curves overall even if there is no exact alignment at minima or maxima.

5.2 Principal Differential Analysis applied

The structure that we have observed in the $X(t), Y(t)$ functions is that of a sinusoidal nature. We are interested in the rate of change of the profile as this is one of the levels at which local variability exists. Given the sinusoidal structure, and our interest in velocity of $X(t), Y(t)$ we propose to use the linear operator on it that would annihilate the structure of such velocity. The LDO we propose and use is

$$Lx = D^3x + \beta_2 D^2x + \beta_1 Dx \tag{5.3}$$

which can be seen as a second order operator on the derivative of x.

This operator would definitely annihilate the structure in an exact sinusoidal structure for a homogeneous differential system, that is to say that $Lx = 0$ if we assume no forcing function is driving the variability and if x was, say $\sin t$. In this way we would have

$$\begin{aligned} Dx &= \cos t \\ D^2x &= -\sin t \\ D^3x &= -\cos t \end{aligned}$$

and hence $D^3x + 0 \times D^2x + 1 \times Dx = 0$ ($\beta_2 = 0, \beta_1 = 1$). We know that the $X(t), Y(t)$ functions are not exactly a *sin* or *cos* function as they have added variability and we are assuming that there is a forcing function $\alpha(t)$ that yields the nonhomogeneous differential model as $Lx = \alpha(t)$. Moreover we know that there is variability at higher frequencies (higher order Fourier basis functions) but it is considered to be error-like. We propose that we the weights $\beta_j(t)$ for the LDO will be the functions that will characterise each type of profile.

The name of Principal Differential Analysis was coined by Ramsay as the process is, in its motivation at least, comparable to that of principal component analysis. The motivation or question is: 'Can we use a set of N functional observations x_i to create a very small set of m functions on which we can approximate efficiently the observed functions?'

In the case of the LDO, we want to have the LDO (defined by its weights) that comes as close as possible in satisfying the homogeneous equation $Lx = 0$.

Once we have decided on the operator L, we can define linearly independent functions, say u_i, that will span its null space. Any function x, satisfying $Lx = 0$

can be expressed as a linear combination of such u_i.

Then we want to minimise:

$$SSE_{PDA}(L) = \sum_{i=1}^{N} \int [Lx_i(t)]^2 dt \tag{5.4}$$

and we minimise to find the weights.

The calculation of such weights is outlined in the appendix, and are results from Ramsay and Silverman [29].

We present first the model for the change in $X(t)$:

$$LX(t) = \alpha(t) + \epsilon(t) \tag{5.5}$$

where $LX(t) = \beta_1(t)DX(t) + \beta_2(t)D^2X(t) + D^3X(t)$ and so can be expressed as

$$D^3X(t) = \beta_1(t)DX(t) + \beta_2(t)D^2X(t) + \alpha(t) + \epsilon(t) \tag{5.6}$$

here, we write β_i instead of $-\beta_i$ as the β_i are to be estimated.

In our calculations we estimate the forcing function $\alpha(t)$, the weight functions $\beta_1(t)$, $\beta_2(t)$ simultaneously and from these we estimate the residual process $\epsilon(t)$. We used 47 Bspline basis of order 8 for creating the functional forms of the data. The order might seem high, but the reader is reminded that we are to calculate third derivatives and we will do penalised smoothing for the creation of the functional data. Hence, we will be dealing with 5th derivatives and hence we fit with 2 degrees more; this results in degree 7 and therefore the order (degree of local polynomial +1) has to be 8. The choice of 47 basis functions yields 41 knots which gives, in the case of the smallest number of points for a profile (189), about 5 internal points between knots, and in the case of the greatest (343), some 8 internal points.

Given the bivariate nature of the profiles' process we end up estimating six functions for the normal profiles and six for the malignant: Two forcing functions: $\alpha_X(t)$ and $\alpha_Y(t)$, and the four weight functions β_{1X}, β_{2X}, β_{1Y}, β_{2Y} for each type.

Figure 5.6 shows the estimated forcing functions $\hat{\alpha}_X(t)$ and $\hat{\alpha}_Y(t)$ for the normal profiles.

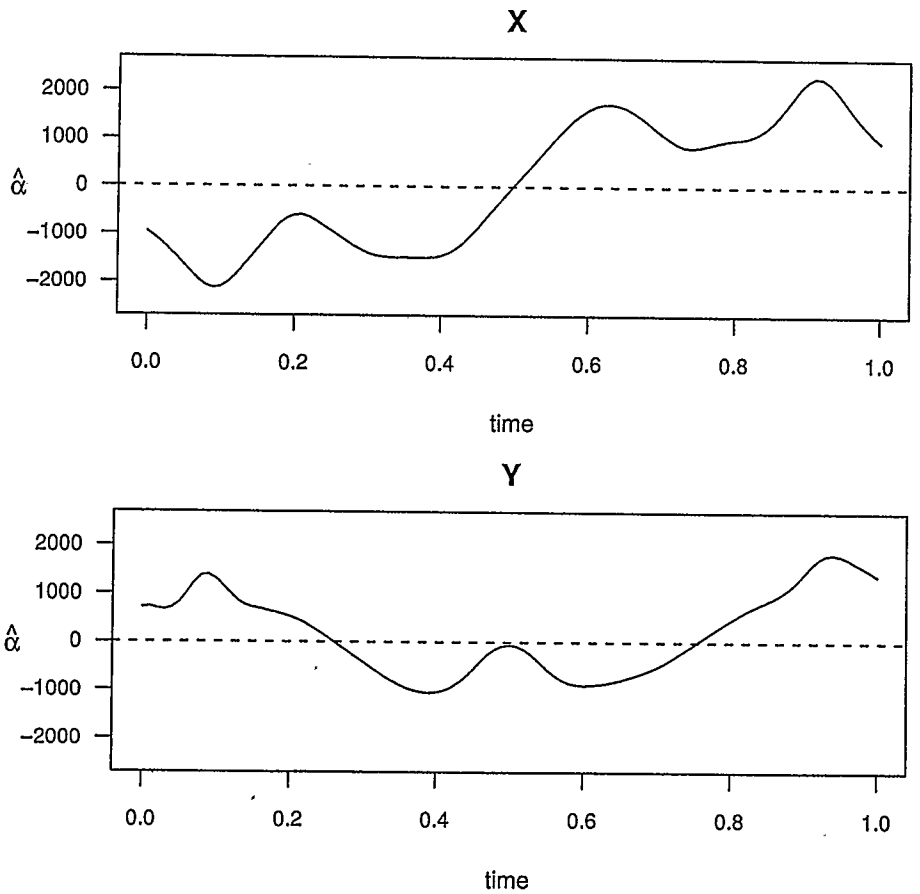


Figure 5.6: Forcing functions for normal profiles

Figure 5.7 shows the estimated weight functions $\hat{\beta}_{1X}(t)$ and $\hat{\beta}_{1Y}(t)$ for the D^1 term of L in the normal profiles and Figure 5.8 the estimated weight functions $\hat{\beta}_{2X}(t)$ and $\hat{\beta}_{2Y}(t)$ for the D^2 term of L in the normal profiles.

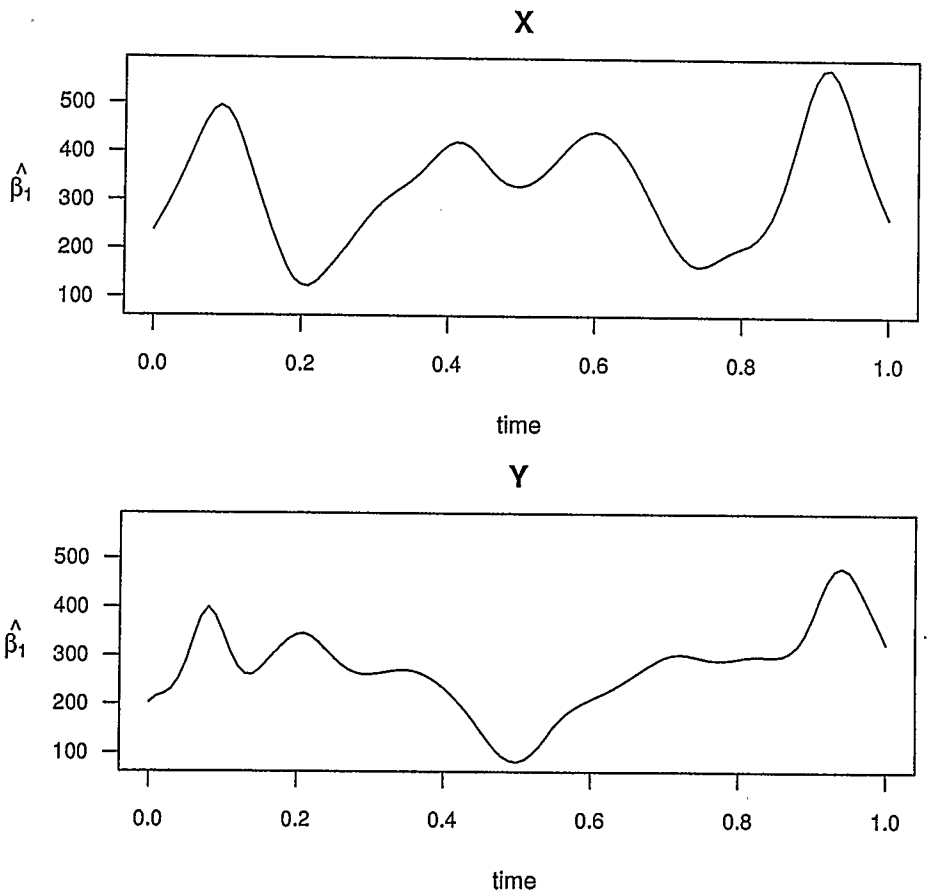


Figure 5.7: First weight functions for Normal profiles

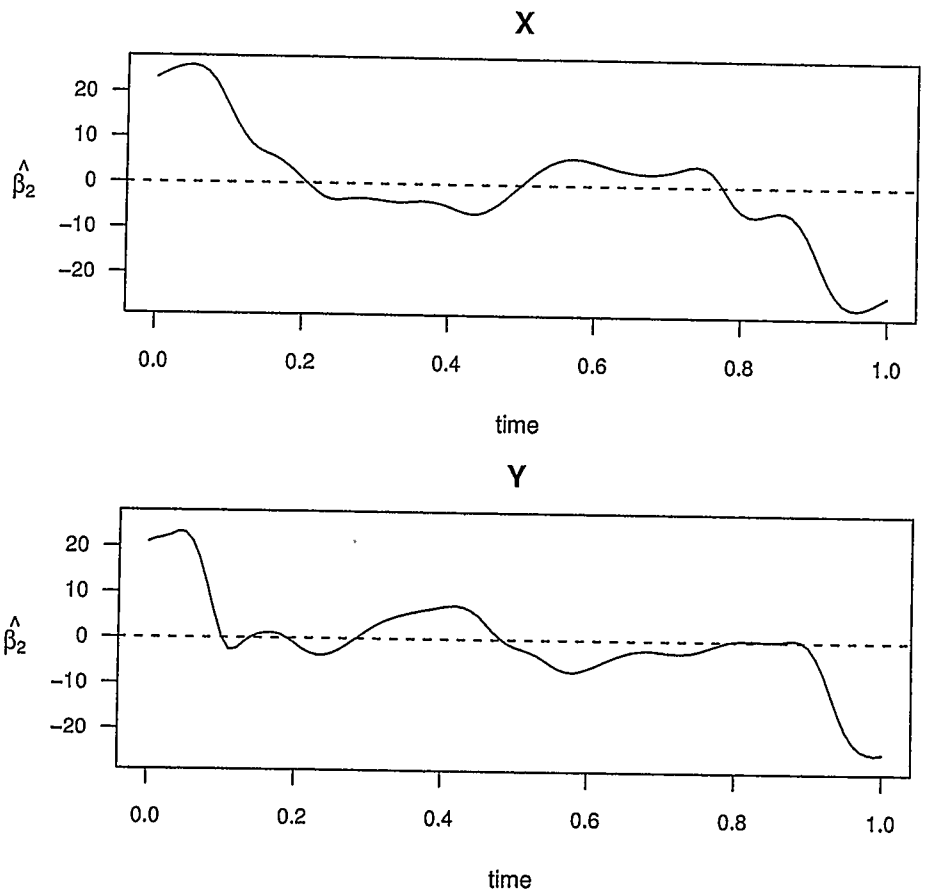


Figure 5.8: Second weight functions for Normal profiles

We present the forcing and weight functions together in Figure 5.9 to show that the forcing function is the largest source of variation and how the first and second derivatives have smaller impact in such variability.

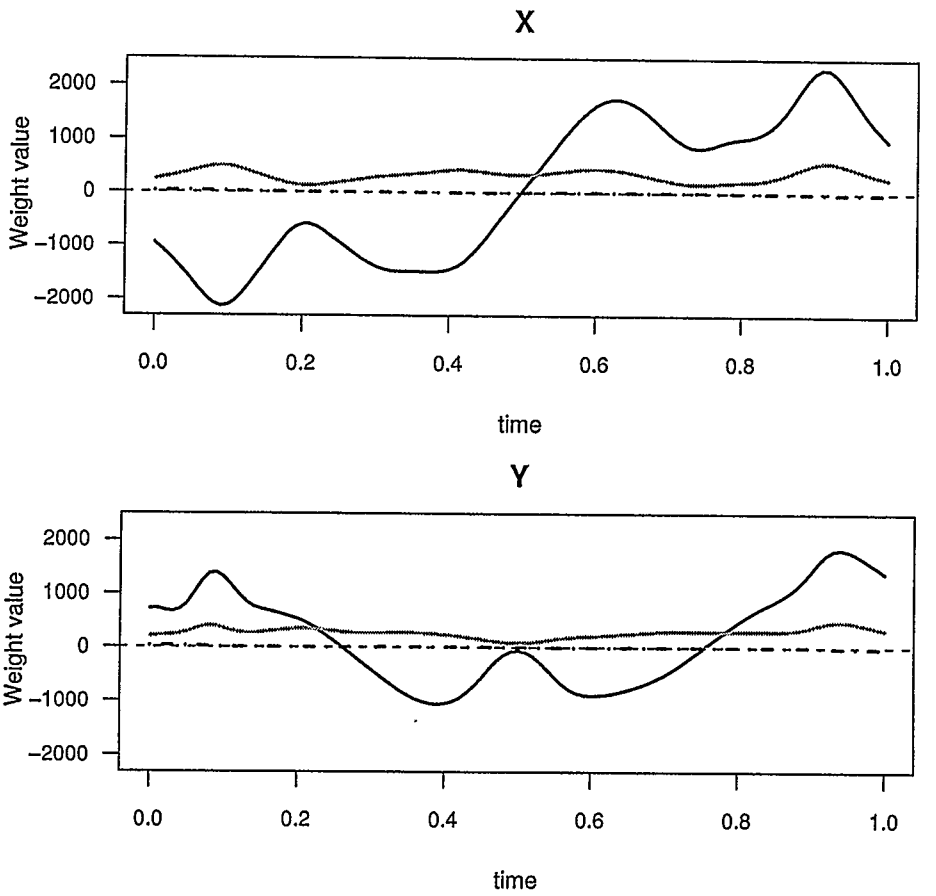


Figure 5.9: Forcing and weight functions for Normal profiles. Solid black line is the forcing function, Grey line is $\hat{\beta}_1$ and dashed line is $\hat{\beta}_2$

The object of fitting the weight function to the LDO was to try to characterise the normal profiles and be able to have residual functions which we expect to be small and oscillate around zero. Figure 5.10 shows the residual functions obtained from applying the LDO with weights calculated from all 50 normal profiles to the normal profiles via crossvalidation. Since the aim is to classify a new profile, we mimic this approach by calculating the residuals for the normal profiles by crossvalidation.

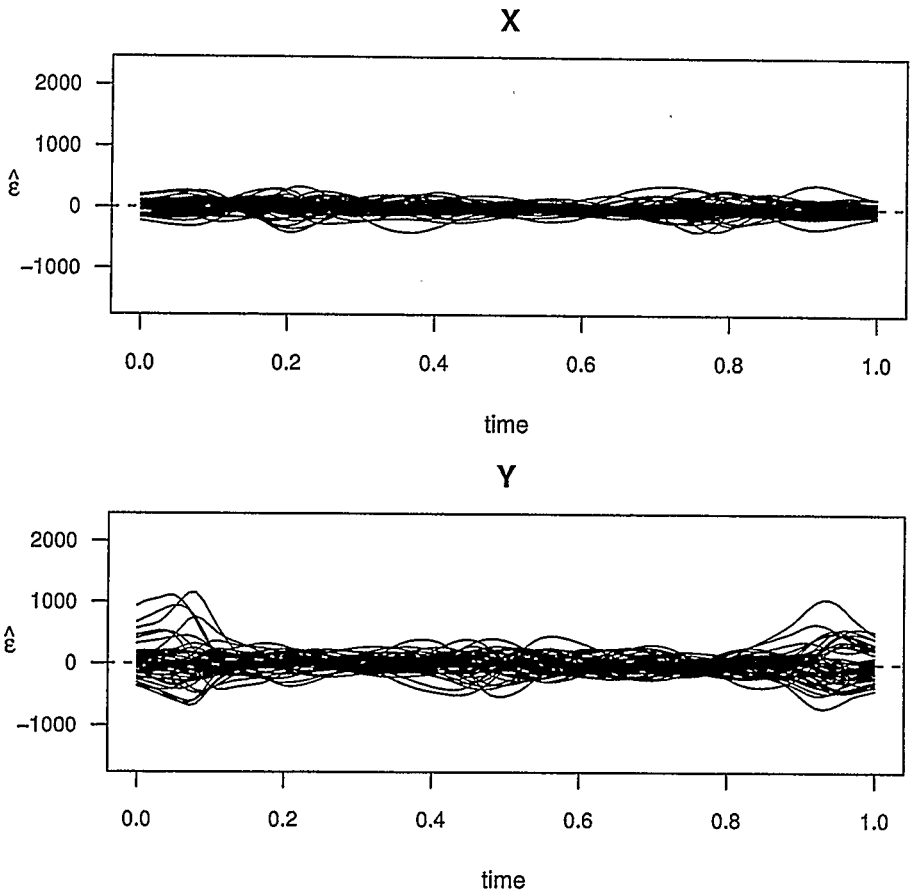


Figure 5.10: Residual functions for Normal weights on Normal profiles via cross-validation

Residual functions calculated for malignant profiles using the weight functions from the normal profiles should be significantly greater than the ones obtained for the normal profiles using the same weight functions. Figure 5.11 shows these residuals. The scale on the vertical axis has been set to be the same in Figure 5.11 and in Figure 5.10 for better comparison.

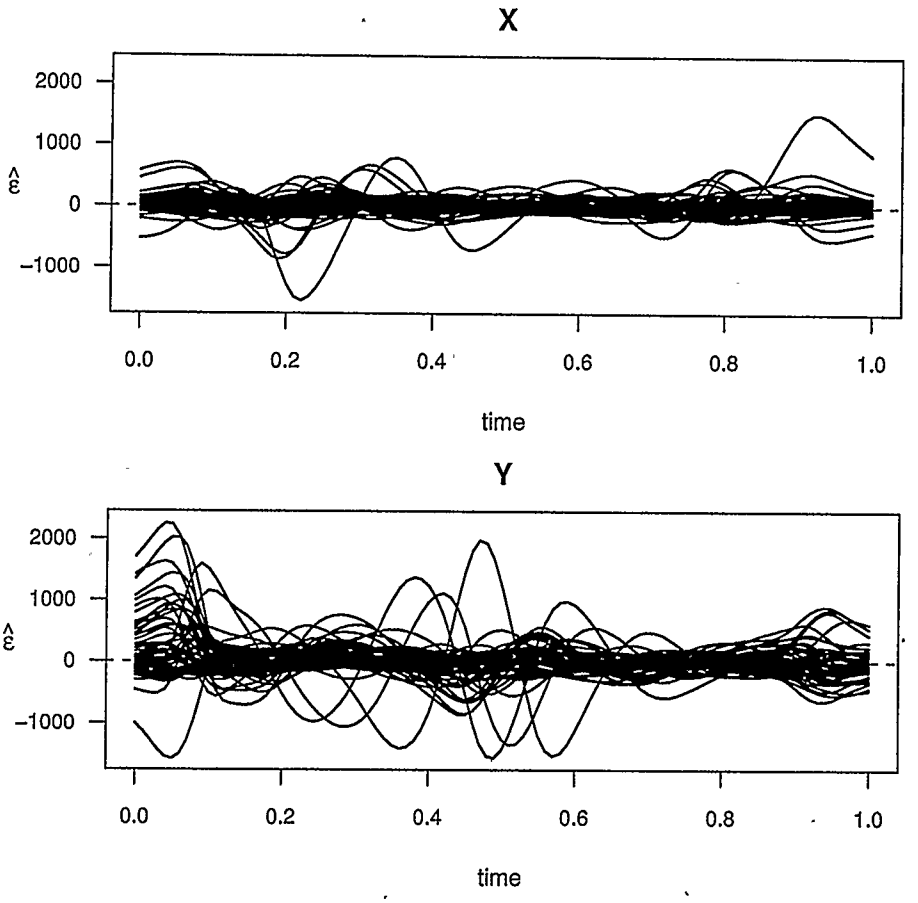


Figure 5.11: Residual functions for Normal weights on Malignant profiles

Figure 5.12 shows the mean of the residual functions when the normal weights are applied to normal profiles, while Figure 5.13 shows the mean of residual functions when the normal weights are applied to malignant profiles.

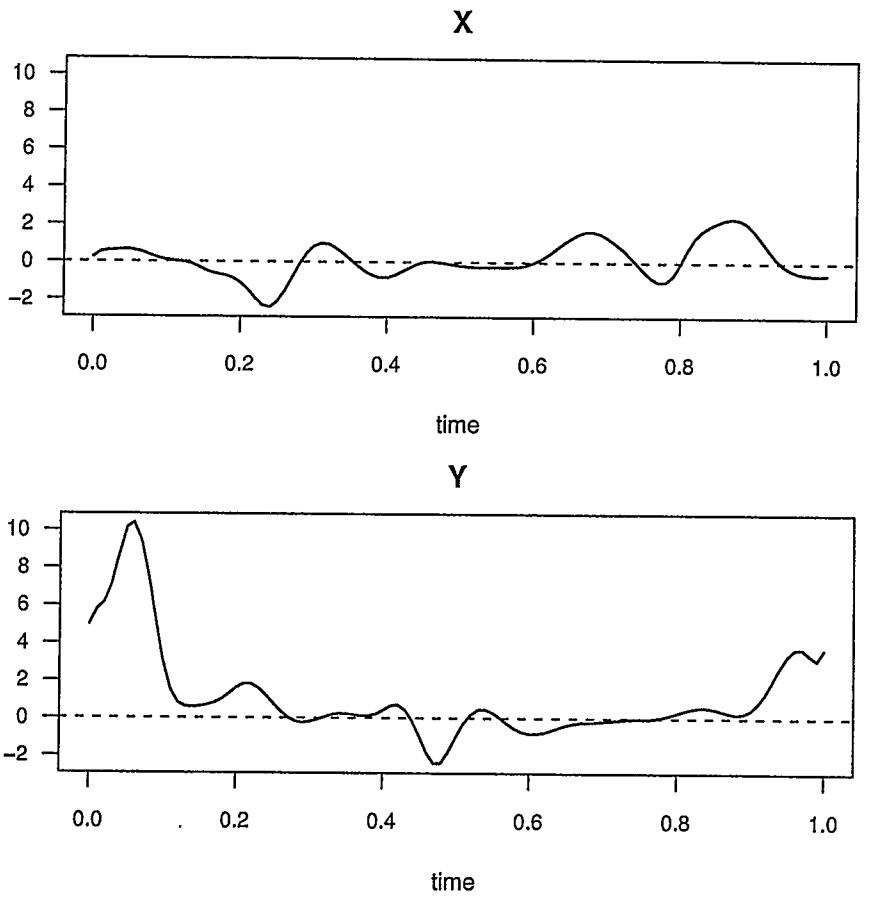


Figure 5.12: Mean residual functions for Normal using Normal weights (via cross-validation)

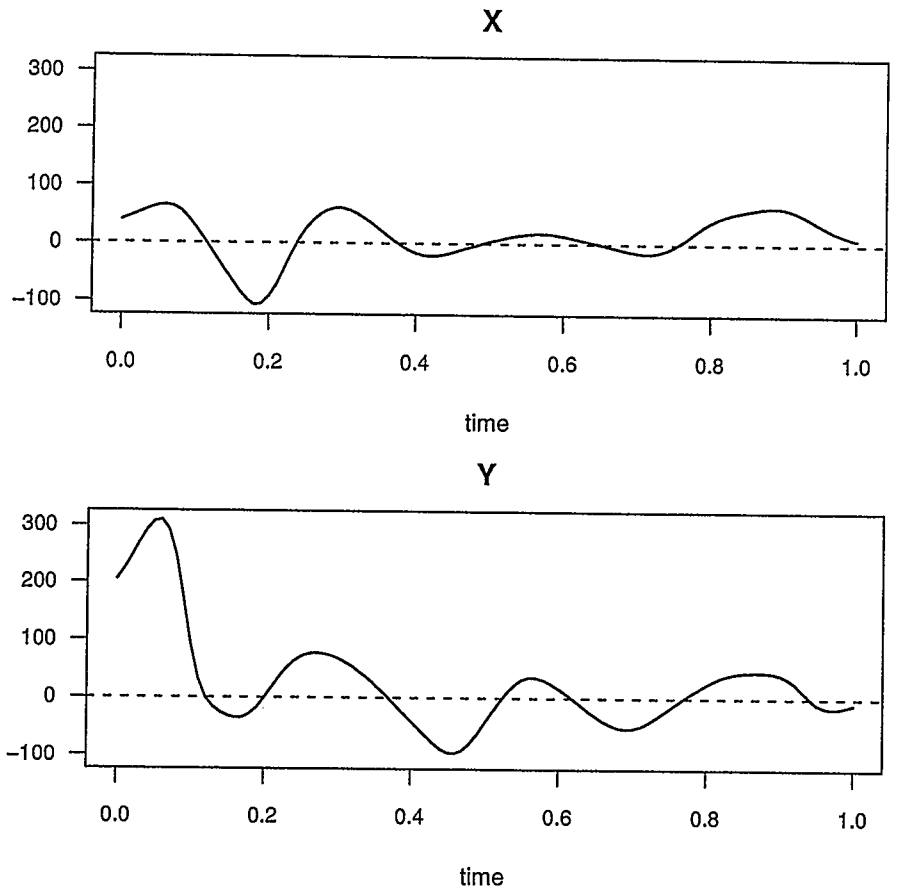


Figure 5.13: Mean residual functions for Malignant using Normal weights

The range for the values of the vertical axis in Figure 5.13 is of greater magnitude than that in Figure 5.12.

Figure 5.14 shows the estimated forcing functions $\hat{\alpha}_X(t)$ and $\hat{\alpha}_Y(t)$ for the malignant profiles.

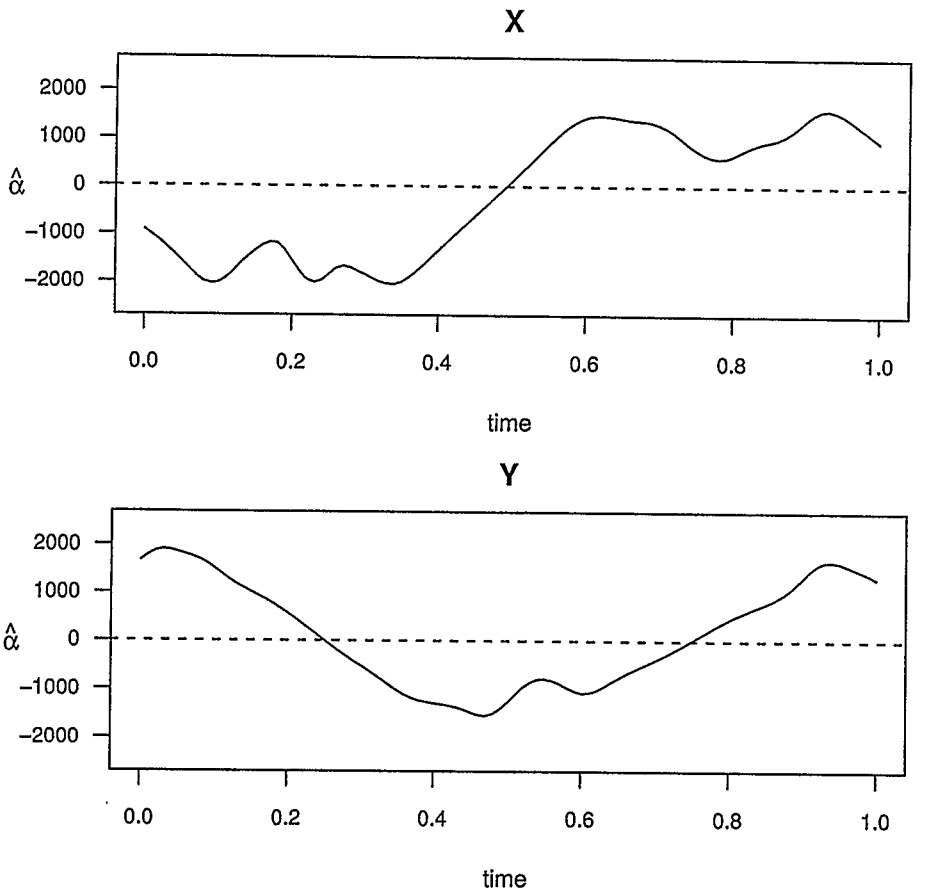


Figure 5.14: Forcing functions for Malignant profiles

Figure 5.15 shows the estimated weight functions $\hat{\beta}_{1X}(t)$ and $\hat{\beta}_{1Y}(t)$ for the D^1 term of L in the malignant profiles and Figure 5.16 the estimated weight functions $\hat{\beta}_{2X}(t)$ and $\hat{\beta}_{2Y}(t)$ for the D^2 term of L in the malignant profiles.

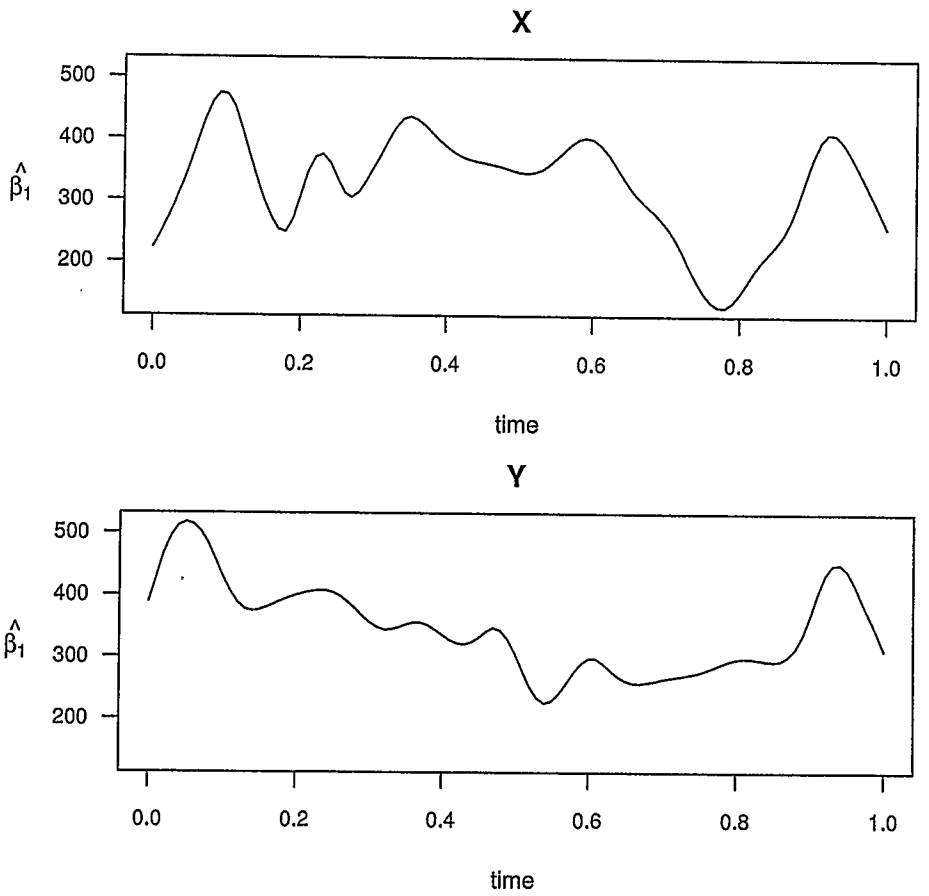


Figure 5.15: First weight functions for Malignant profiles

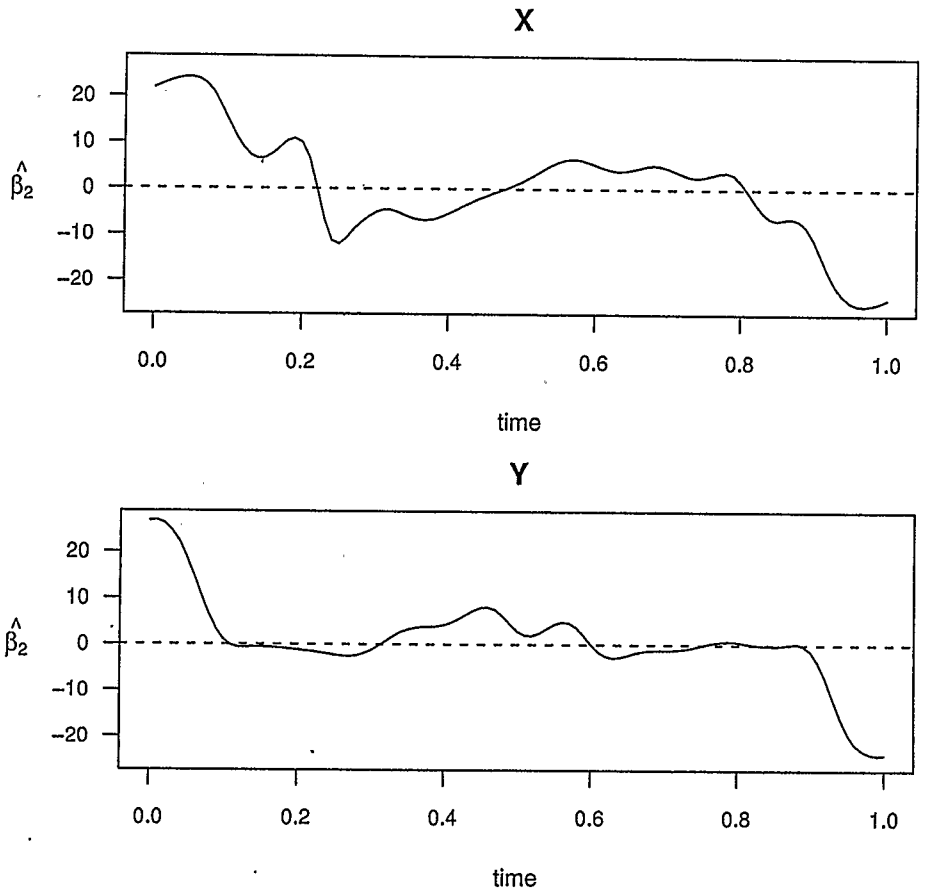


Figure 5.16: Second weight functions for Malignant profiles

We present the forcing and weight functions together in Figure 5.17 to show that the forcing function is the largest source of variation and how the first and second derivatives have smaller impact.

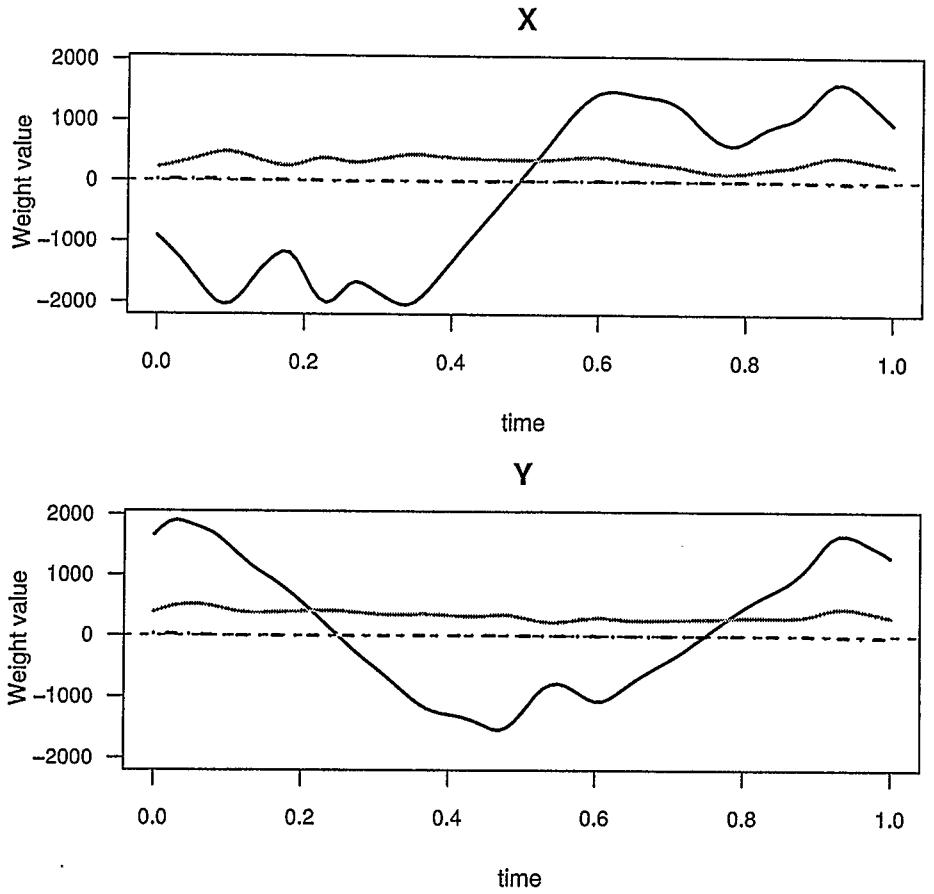


Figure 5.17: Forcing and weight functions for Malignant profiles. Solid black line is the forcing function, Grey line is $\hat{\beta}_1$ and dashed line is $\hat{\beta}_2$

Figure 5.18 shows the residual functions obtained from applying the LDO with weights calculated from all 50 malignant profiles to the malignant profiles via cross-validation as done for applying normal profiles' weights on normal profiles.

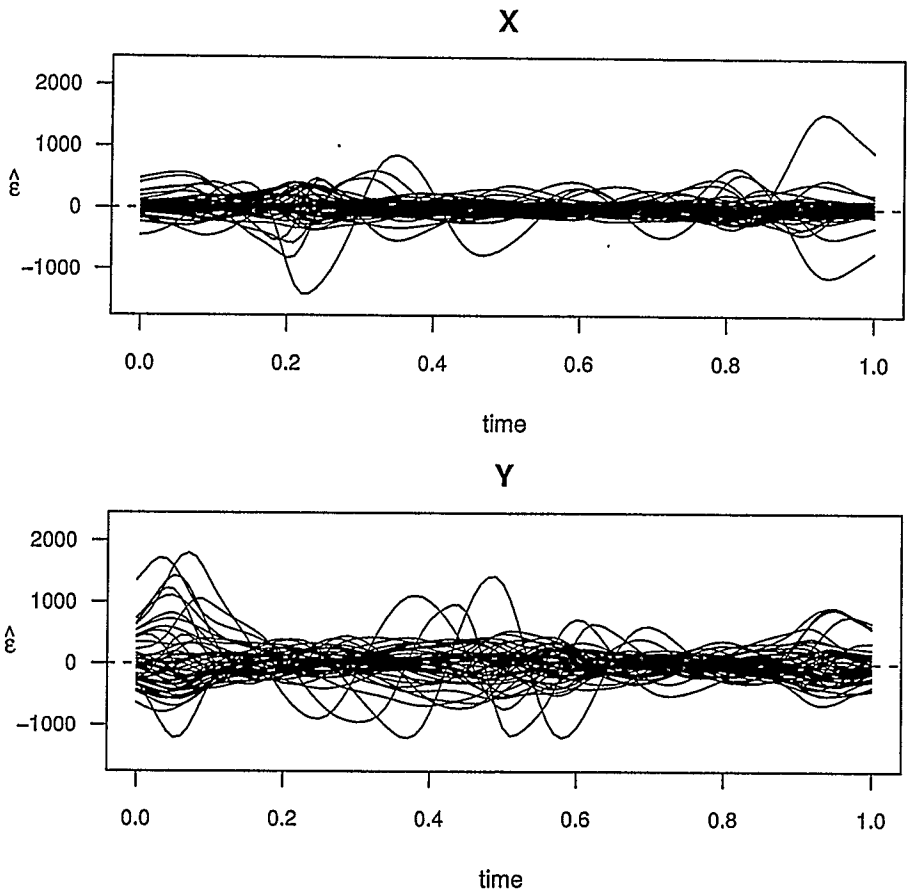


Figure 5.18: Residual functions for Malignant weights on Malignant profiles (via crossvalidation)

Residual functions calculated for normal profiles using the weight functions from the malignant profiles significantly deviate from 0, more so than the ones obtained for the malignant profiles using the same weight functions. Figure 5.19 shows these residuals. The scale on the vertical axis has been set to be the same in Figure 5.19 and in Figure 5.18 for better comparison.

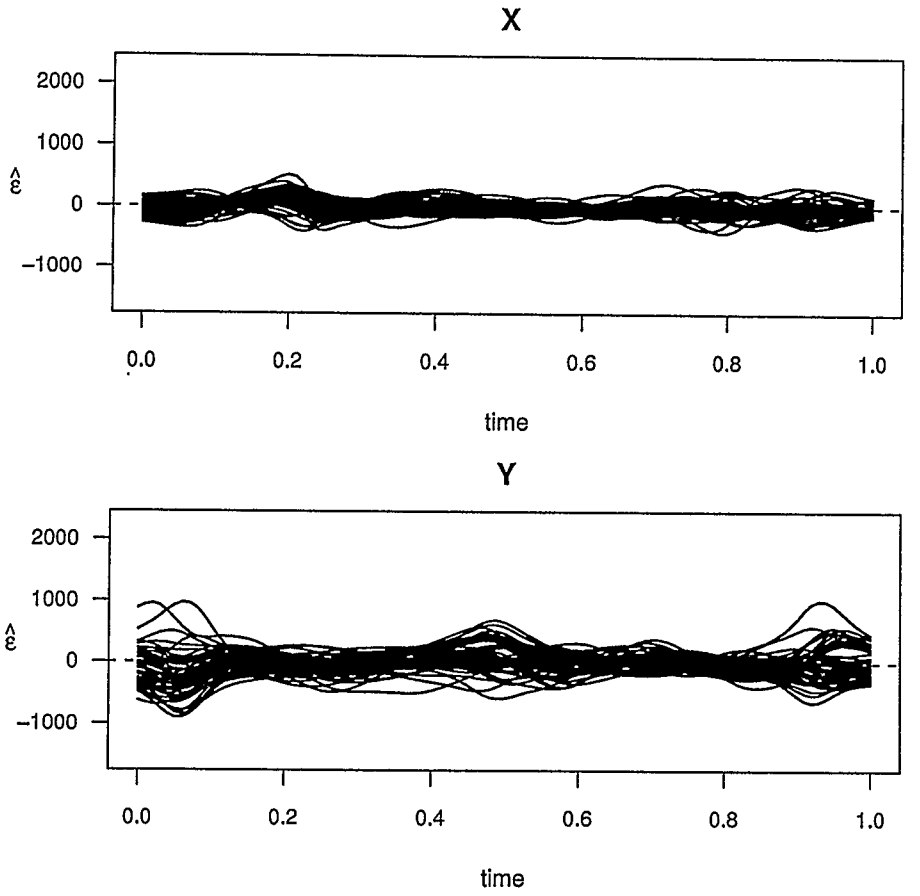


Figure 5.19: Residual functions for Malignant weights on Normal profiles

When looking at Figure 5.18 and Figure 5.19 the reader might think that the residuals from malignant weights on malignant profiles deviate more from 0 than the residuals from malignant weights on normal profiles do. This appears to be so because the variability in Figure 5.18 is more than the variability in Figure 5.19 at some t's. By taking a closer look at the Figures it can be seen that although the variability is greater in Figure 5.18 most the curves are around the zero line whereas curves in Figure 5.19 are mainly and under the zero line. It is worthmaking a note

of this, otherwise the reader would tend to believe there is a contradiction between the assumptions of the behaviour of the residuals and the results obtained.

Figure 5.20 and Figure 5.21 seem to be inconsistent with Figure 5.18 and Figure 5.19, but when taking into consideration the issue mentioned in the above paragraph, the reader will find that they *are* consistent.

Figure 5.20 shows the mean of the residual functions when the malignant weights applied to malignant profiles, while Figure 5.21 shows the mean of residual functions when the malignant weights are applied to normal profiles.

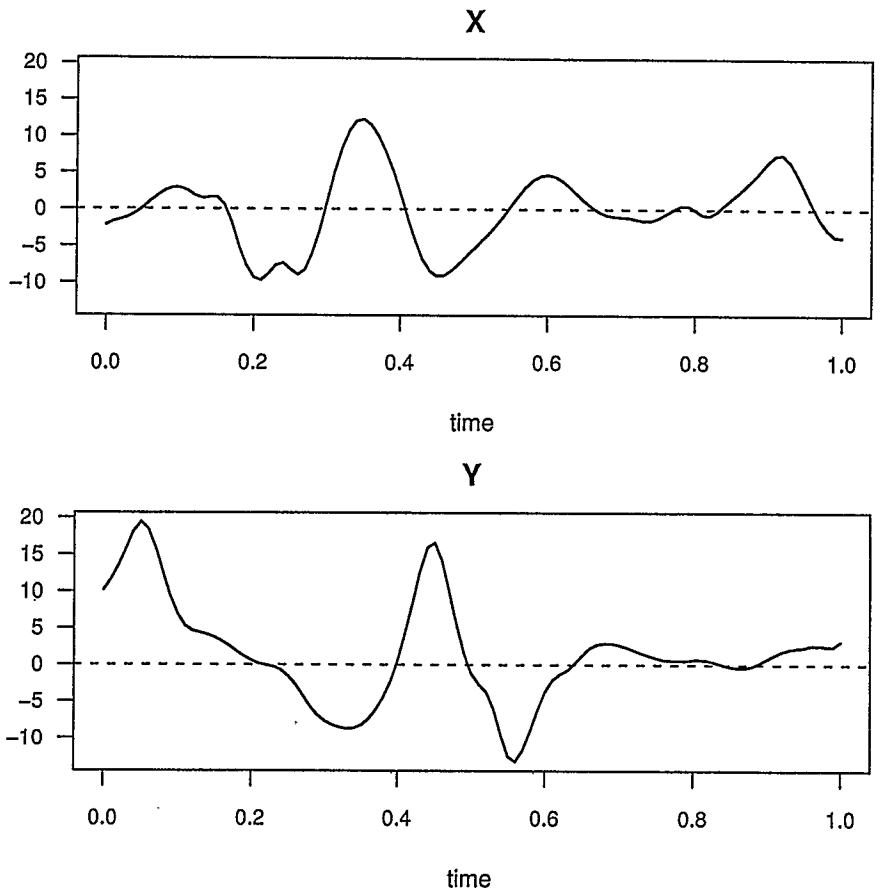


Figure 5.20: Mean residual functions for Malignant using Malignant weights (via crossvalidation)

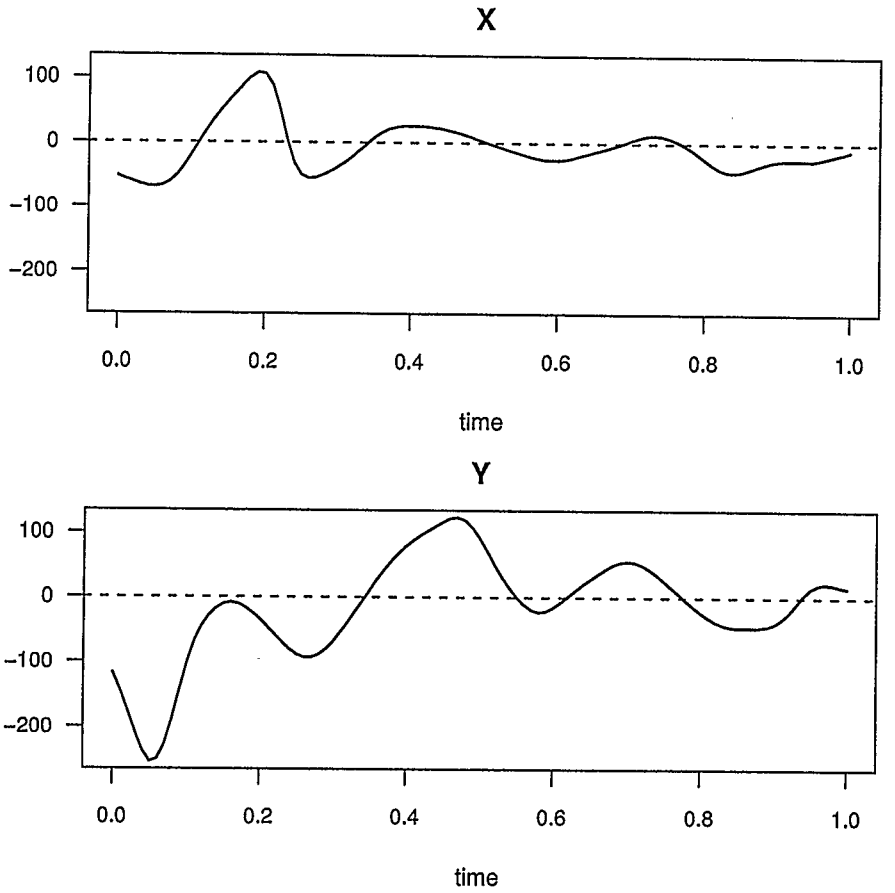


Figure 5.21: Mean residual functions for Normal using Malignant weights

The range for the values of the vertical axis in Figure 5.21 is of greater magnitude than that in Figure 5.20.

We assumed that the residuals obtained from applying normal weight functions to normal profiles (normal on normal) and malignant weight functions to malignant profiles (malignant on malignant) would be distributed closer to zero than those obtained from applying normal weight functions to malignant profiles and vice versa. The results we observe are consistent with this assumption. We have seen that the

mean of the residuals of normal on normal and those of malignant on malignant are not significantly different from zero at any time t for $t \in [0, 1]$.

Figure 5.22 shows the functional 95% confidence intervals for the $X(t)$ and $Y(t)$ for the mean of normal on normal and for malignant on malignant and it is clear that zero is always inside the interval.

These intervals are calculated in an analogous way as confidence intervals for point estimates, the only difference is that the mean and standard deviation of the curves are curves themselves. The mean and the standard deviation are functions of the parameter t.

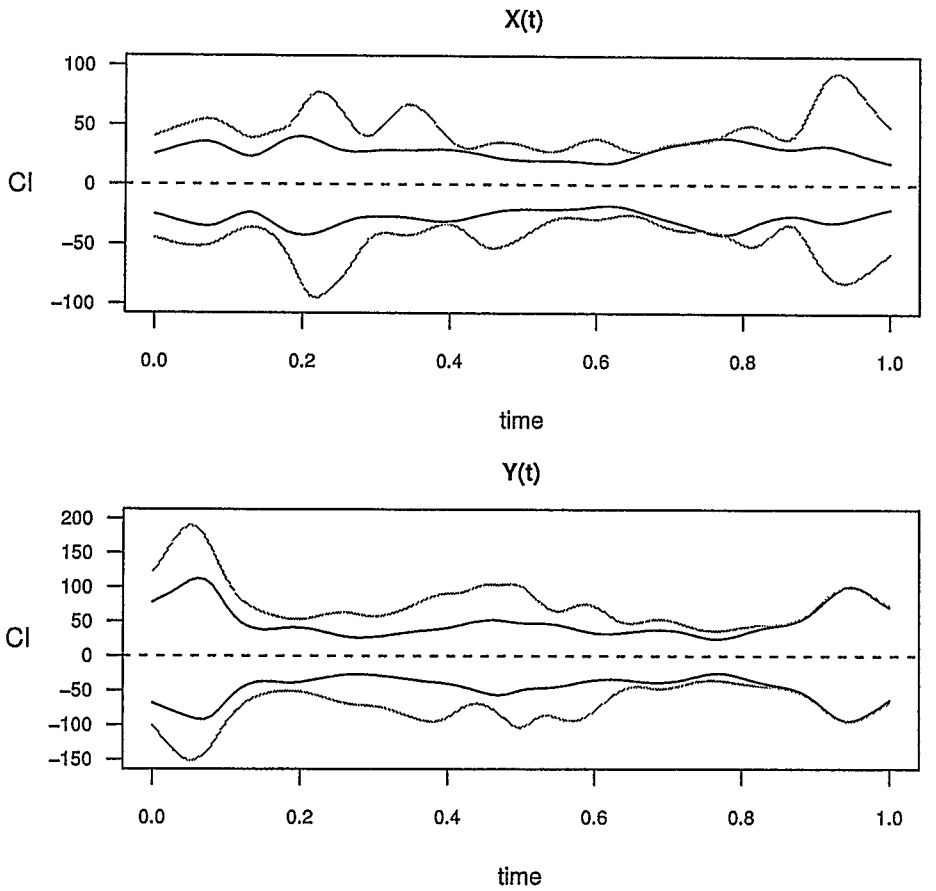


Figure 5.22: 95% Confidence-like interval for the mean of residuals. Black line: Normal on Normal, grey line: Malignant on Malignant

Figure 5.23 and 5.24 show the p-values for the Wilcoxon test for the location parameter of zero. This figure shows that the p-values for normal on normal are not less than 0.53 for $X(t)$ residuals and not less than 0.33 for $Y(t)$; the p-values for malignant on malignant are not less than 0.21 for $X(t)$ and not less than 0.29 for $Y(t)$ and hence it can be concluded that the residuals are centred at zero at all times $t \in [0, 1]$.

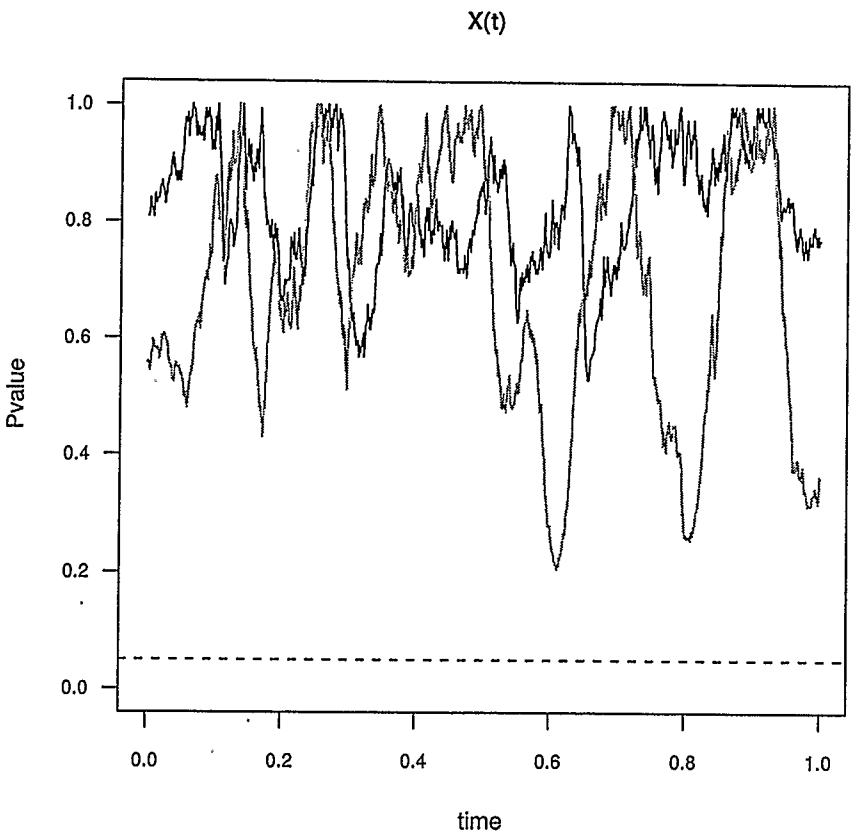


Figure 5.23: Pointwise (fine grid of 1000 times t_i) P-values of testing $\mu = 0$ for residuals (X). Black line: Normal on Normal, grey line: Malignant on Malignant. Dashed line: P-value=0.05

Figure 5.25 shows the functional 95% confidence intervals for the $X(t)$ for the mean of normal on malignant and for malignant on normal and it is clear that zero is not always inside the interval. For normal on malignant, zero is not within the interval 28.8% of the time for X and 21.1% of the time for Y. For malignant on normal, zero is not within the interval 39.8% of the time for X and 52.5% of the

time for Y.

Figures 5.26 and 5.27 show the p-values for the Wilcoxon test for the location parameter of zero. This figure shows that the p-values are sometimes less than 0.05 and in those time periods it can be concluded that the residuals are centred at a value significantly different from zero.

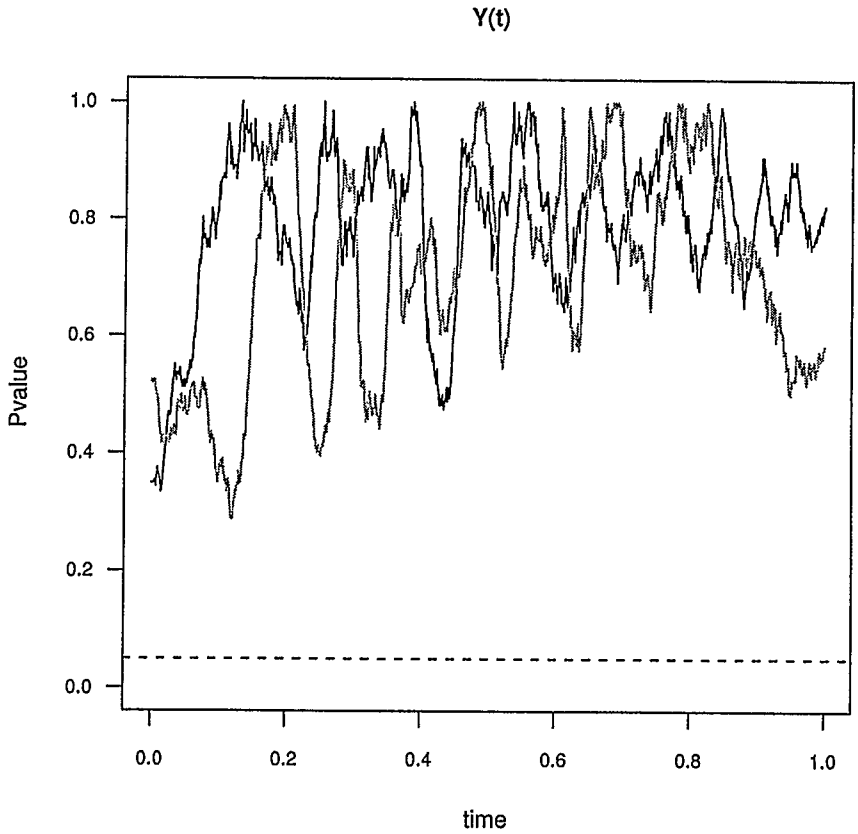


Figure 5.24: Pointwise (fine grid of 1000 times t_i) P-values of testing $\mu = 0$ for residuals (Y). Black line: Normal on Normal, grey line: Malignant on Malignant. Dashed line: P-value=0.05

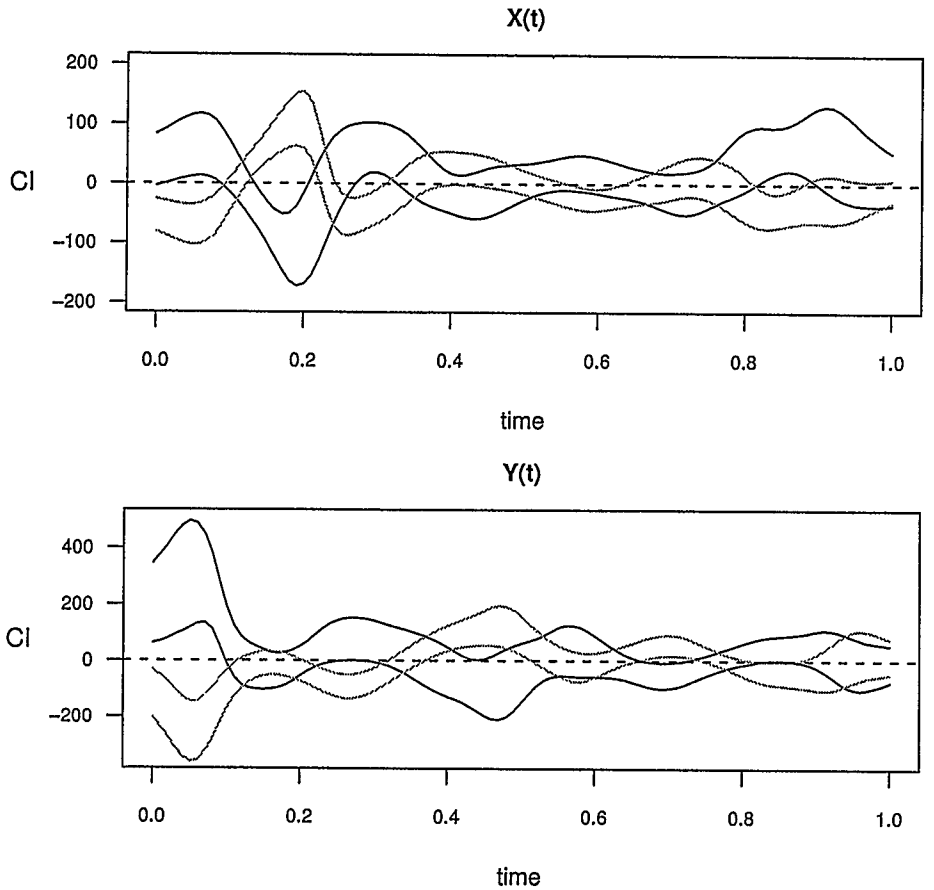


Figure 5.25: 95% Confidence-like interval for the mean of residuals. Black line: Normal on Malignant, grey line: Malignant on Normal

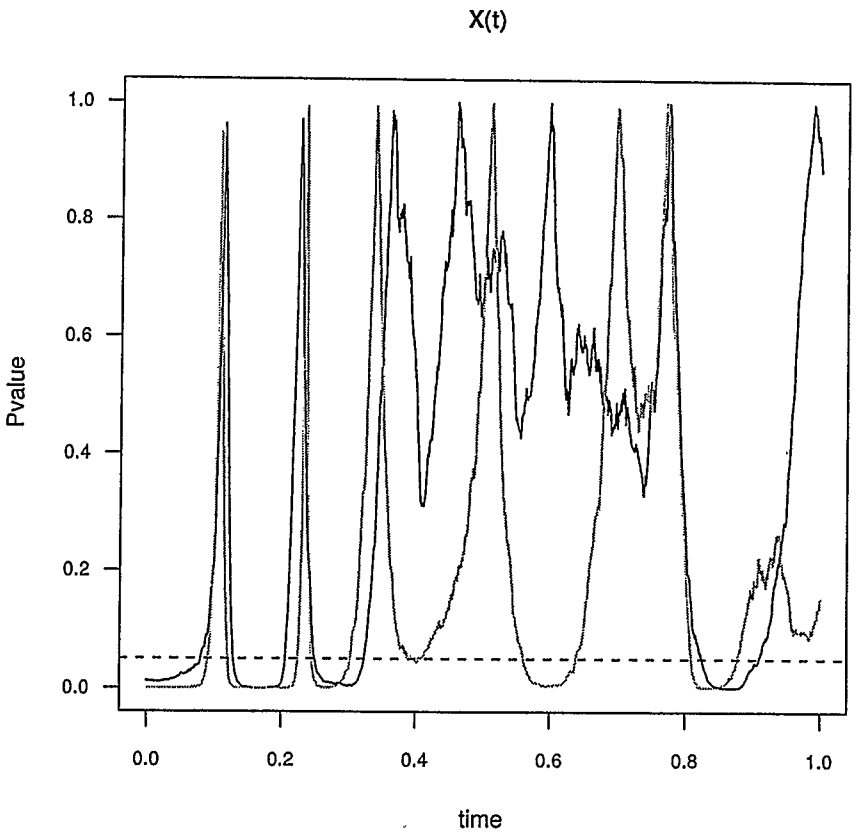


Figure 5.26: Pointwise (fine grid of 1000 times t_i) P-values of testing $\mu = 0$ for residuals (X). Black line: Normal on Malignant, grey line: Malignant on Normal. Dashed line: P-value=0.05

The analysis has shown that the residual processes obtained by applying weight functions of the same type as the profile type (normal on normal or malignant on malignant) are "well behaved" in both of the coordinates X, Y and their confidence intervals always cover zero. On the other hand, when applying weight functions of different type than that of the profiles (normal on malignant, malignant on normal)

the residual processes are "ill behaved" in at least one of the coordinates X, Y, having the confidence intervals not covering zero over nonnegligible proportions of time spanning from 21.1% to 52.5%.

Based on this analysis and given the fact that profiles are obtained in batches, say from a biopsy, a new batch of profiles can be digitised, converted into functional data, registered to the normal profiles' mean function and then have the weight functions applied to each of the profiles to obtain the residual processes. Once these are obtained, the confidence intervals and/or the Wilcoxon tests can be performed to obtain a diagnostic of normal or malignant.

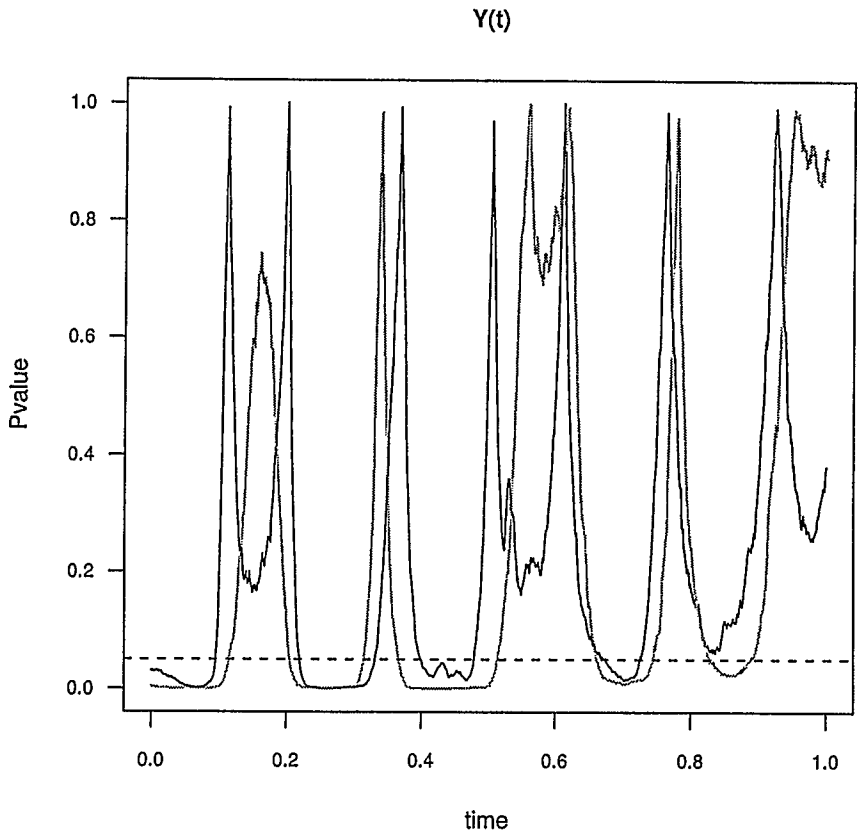


Figure 5.27: Pointwise (fine grid of 1000 times t_i) P-values of testing $\mu = 0$ for residuals (Y). Black line: Normal on Malignant, grey line: Malignant on Normal. Dashed line: P-value=0.05

Chapter 6

Conclusions

The methods and analyses used in this thesis have been based on many important and scientifically significant ideas and methods created recently by renowned people in their respective fields of research such as Hobolth, Jensen [14, 16, 15], Ramsay and Silverman [29, 30, 28], and on equally important methods that were created previously and that are now known to many, but that have been of great impact to the study of statistics and hence are still used and built on, such as Fourier transforms, splines approximations and some of the basic multivariate statistical procedures.

The chapters in this thesis present a brief introduction and overview of some of the methods and look at some others in a more detailed manner. This is an attempt to give the reader some understanding for each of the methods used.

The purpose of this thesis is to combine techniques from the methods presented in a new approach that surpasses constraints faced when applying the methods individually. The approach used is of an exploratory nature in search for a possible aid in the diagnosis.

The alignment and registration of the profiles is, from a biological point of view, arbitrary and has no physiological meaning. It is, however, a protocol followed to analyse all profiles in a consistent way. The 'reference' point is reached in each profile by following a fixed criterion.

The different fitting procedures used in this thesis are performed using least

squares fitting. The use of other methods such as absolute error may give different inferences, for example, in the principal differential analysis.

The ways in which we used the methods in this thesis allowed us to deal with the profiles as continuous functions which better represent the continuous form and nature of the nuclei profiles. We were able to study nuclei profiles without the constraint of restricting our study to that of objects that are star shaped with respect to their centre of mass [14].

If we had not been able to apply our methods to non-star shaped profiles, we would have lost 11 profiles from the Malignant set and 1 from the Normal set. Having started with a set of 50 profiles for each type, this would have not only impacted our sample size, but we would have lost information on the possible relationship of shape and type. Excluding the non-star shaped profiles could have affected the discovery of the characteristics pointed out by the first principal component in section 3.1. Section 3.1 shows that one of the graphical characteristics where the scores of principal components differ significantly relates to the non-convexity of the shape and to not having star shaped objects.

The bivariate approach, the use of $X(t), Y(t)$ for the profiles, also opens the door for future analyses where weight functions such as those in section 5 would not only be calculated for the coordinates separately, but would be modelled to assess variability and crosscorrelation between variables. Mainly the set of $\beta(t)$'s would be of greater dimension as we would have $\beta_{XY}(t)$ and $\beta_{YX}(t)$ plus the $\beta_X(t)$ and $\beta_Y(t)$. Here the subindexes XY and YX would indicate the influence that X has on Y and vice versa.

This thesis enabled us to have a better understanding in a tangible graphical way

of the shape differences between the two types of profiles. These are observable in the results obtained from Section 3.1 and Section 3.2. In these sections we observe results that support the assumptions set forth at the beginning of the thesis which state the belief that healthy nuclei would tend to be more convex than malignant ones.

We gained diagnostics of malignancy and a discriminator between the two types of profiles. Although the use of discriminant analysis in section 3.2 gave False Positive and False Negative rates higher than what could be desired, the procedure not only yields a label for classification but it also provides the misclassification rates which are a useful measurement of accuracy.

We also gained a useful tool in the principal differential analysis, namely the criterion for classification based on the intervals for the residuals.

When it comes to having some measure of uncertainty, we can relate to the confidence intervals and the p-value function. It is important that the reader remembers that the profiles, although they are presented individually, usually belong to a set of nuclei which comes from one tissue sample such as a biopsy. In this sense, an analyst will not be facing the problem of having only one profile to diagnose or to classify, as there will be a set of profiles and hence the sample means and standard deviations of the residuals obtained by applying the weight functions from section 5 and the construction of the confidence intervals is possible.

Based on these intervals, we can be 95% confident in the proposed decision rule of classifying as Normal if the confidence intervals for the residuals obtained from applying the Normal weight functions contain zero throughout the whole $[0, 1]$ interval, or classifying as Malignant if there are periods of time in which the confidence

intervals do not contain zero. Once the new set of profiles is classified as Normal or Malignant the weight functions can be recalculated by using the full set of each type of nuclei.

The methods used here open the doors to other possible diagnostics or classifiers. Some ideas came into mind during the realisation of this thesis. One of them has been mentioned above, that is the assessment of a linear differential operator that relates the influence of one coordinate on the other. This is proposed as a more realistic approach to assess the nuclei shapes, as opposed to having coordinates separate from each other.

Another possibility is to impose statistical and distributional assumptions on the coefficients of the Fourier expansions of the $X(t)$ and $Y(t)$ functions. The distributional assumptions could be similar to those imposed by Hobolth and Jensen on the coefficients of their radius vector function. The results would not follow directly as one of the main advantages in their work is that the X,Y are combined into a radius vector function and this is possible because of having star shaped objects. The intuitive feeling for these distributions is that we would have to consider multivariate distributions.

An extension of logistic regression came to mind, given the fact that functional data analysis enables the creation of linear models of functional predictors on scalar responses. This logistic regression could use the functional $X(t)$ and $Y(t)$ as predictors, or their derivatives and have the binary response to be "Normal" or "Malignant" type of nucleus.

There is still work to be done in creating some form of functional significance testing method. We have had pointwise testing on fine grids of values obtained from

functional data, but there has not been a functional testing method per-se.

Other future work would include a transformation on the $X(t), Y(t)$ functions into polar coordinates $r(t), \theta(t)$. This might bypass some of the issues with orientation, registration and landmarks as well as not being restricted to star shaped objects. Tests based on the derivatives of $\theta(t)$ and derivatives of $r(t)$ could be useful. The function $\theta(t)$ would identify those non-starshaped objects. Tests on $r\prime(t)$ would capture the local variability of profiles.

Three-dimensional analysis of nuclei via parameterisation on two "time" parameters (s,t) and the use of spherical coordinates $(r(s,t),\ \theta(s,t),\ \phi(s,t))$ is another approach. With this approach, we would concentrate on $\theta(s,t)$, $\phi(s,t)$ and first derivatives of $r(s,t)$ in various directions as an approximation to measuring curvature in three-dimensional space.

The inclusion of "expert opinion" is an important issue to address. Based on pathologist's expertise, for example, we could search for special characteristics that are important because of their physiological function. Another way of using expert opinion would be to incorporate it from a Bayesian point of view and assume prior distributions on the parameters that address issues raised by such expert opinion.

Bibliography

[1] Y. Araki. Functional data analysis of human gait. Master's thesis, University of Calgary, 2002.

[2] Leo Breiman. Reply to comment on "Statistical modeling: The two cultures" (Pkg: P199-231). *Statistical Science*, 16(3):226–231, 2001.

[3] Leo Breiman. Statistical modeling: The two cultures (Pkg: P199-231). *Statistical Science*, 16(3):199–215, 2001.

[4] C. C. Chang, S.M. Hwang, and D.J. Buehrer. A shape recognition scheme based on relative distances of feature points from the centroid. *Pattern Recognition*, 24:1053–1063, 1991.

[5] C. De Boor. *A practical guide to splines.* Springer, 2001.

[6] E. Dommergues, J.L. Dommergues, F. Magniez, P. Neige, and E.P. Verrecchia. Geometric measuremente analysis versus fourier series analysis for shape characterization using the gastropod shell (trivia) as an example. *Mathematical Geology*, 35(7):887–894, 2003.

[7] I.L. Dryden and K.V. Mardia. *Statistical shape analysis.* Wiley, 1998.

[8] S. Efromovich. *Nonparametric curve estimation.* Springer-Verlag, 1999.

[9] S. Ferson, F.J. Rohlf, and R.K. Kohen. Measuring shape variation of two-dimensional outlines. *Systematic Zoology*, 34(1):59–68, 1985.

[10] A. W. Fitzgibbon, M. Pilu, and R. B. Fisher. Direct least-squares fitting of ellipses. *Pattern Analysis and Machine Intelligence*, 21(5):476–480, May 1999.

[11] P.J. Green and B.W. Silverman. *Nonparametric Regression and Generalized Linear Models (A roughness penalty approach).* Chapman & Hall, London, 1st edition edition, 1994.

[12] U. Grenander. *General Pattern Theory.* Oxford University Press, Oxford, 1993.

[13] U. Grenander and K.M. Manbeck. A stochastic model for defect detection in potatoes. *Journal of Computational and Graphical Statistics*, 2:131–151, 1993.

[14] A. Hobolth, J. Pedersen, and E.B. Vedel Jensen. A deformable template model, with special reference to elliptical templates. *Journal of Mathematical Imaging and Vision*, 17 (2):131–137, 2002.

[15] A. Hobolth, J. Pedersen, and E.B. Vedel Jensen. A continuous parametric shape model. *Annals of the Institute of Statistical Mathematics*, 55(2):227–242, 2003.

[16] A. Hobolth and E.B. Vedel Jensen. Modelling stochastic changes in curve shape, with an application to cancer diagnostics. *Advances in Applied Probability (SGSA)*, 32:344–362, 2000.

[17] R.A. Johnson and D.A. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall, Upper Saddle River, New Jersey, 4 edition, 1998.

[18] S. Lele and J.T. Richtsmeier. Euclidean distance matrix analysis: A coordinate free approach for comparing biological shapes using landmark data. *American Journal of Physical Anthropology*, 86:415–427, 1991.

[19] S.R. Lele and J.T. Richtsmeier. On comparing biological shapes: Detection of influential landmarks. *American Journal of Physical Anthropology*, 87:49–65, 1992.

[20] G.P. Lohmann. Eigenshape analysis of microfossils: a general morphometric method for describing changes in shape. *Mathematical Geology*, 15:659–672, 1983.

[21] Sven Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31-8:983–1001, 1998.

[22] N. Macleod. Generalizing and extending the eigenshape method of shape space visualization and analysis. *Paleobiology*, 25(1):10–138, 1999.

[23] M.I. Miller, S. Joshi, D.R. Maffitt, J.G. McNally, and U Grenander. Membranes, mitochondria and ameoba: shape models. *Advances in Applied Statistics*, II:141–163, 1994.

[24] H.J. Norris, R.L. Becker, and U.V. Mikel. A comparative morphometric and cytophotometric study of endometrial hyperplasia, atypical hyperplasia, and endometrial carcinoma. *Hum Pathol*, 20:219–223, 1989.

[25] M. Peura and J. Iivarinen. Efficiency of simple shape descriptors. In *Proceedings from the Third International Workshop on Visual Form*, pages 443 - 451, 1997.

[26] A.T. Popescu, C. Vidulescu, C.L. Stanciu, B.O. Popescu, and L.M. Popescu. Selective protection by phosphatidic acid against staurosporine induced neuronal apoptosis. *Journal of Cellular and Molecular Medicine*, 6,#3, 2002.

[27] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-00-3.

[28] J.O. Ramsay and C.J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society B*, 53:539–572, 1991.

[29] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis.* Springer-Verlag, New York, 1997.

[30] J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis,methods and case studies.* Springer-Verlag, New York, 1st edition, 2002.

Appendix A

PDA calculations by pointwise minimisation

This approach is based on having a very fine mesh of values and that the corresponding design matrix $\mathbf{Z}(t)$ (as in usual regression analysis) be of full rank for all t or , along the same line, the determinant of the dispersion matrix $\mathbf{Z}(t)'\mathbf{Z}(t)$ is bounded away from zero.

Let the pointwise criterion be

$$PSSE_{L_w}(t) = \frac{1}{N}\sum_i (Lx_i)^2(t) = \frac{1}{N}\sum_i \left[\sum_{j=0}^{m} w_j(t)(D^j x_i)(t)\right]^2 \qquad \text{(A.1)}$$

having $w_m(t) = 1$ for all t. When t is taken as fixed, then we have the least squares fitting criterion. Define $w(t)$ to be an m-vector of coefficients

$$w(t) = (w_0(t), \ldots, w_{m-1}(t))',$$

the $N \times m$ pointwise design matrix $\mathbf{Z}$ as

$$\left\{(D^j x_i)(t)\right\}_{i=1,\ldots,N;j=0,\ldots,m-1}$$

and the N-dimensional dependent variable vector as

$$y(t) = \{-(D^m x_i)(t)\}_{i=1,\ldots,N}$$

and following the usual least squares criterion, provided the conditions stated above are satisfied, we have

$$w(t) = [\mathbf{Z}(t)'\mathbf{Z}(t)]^{-1}\,\mathbf{Z}(t)'y(t).$$

Appendix B

PDA by basis expansion

Computing the solution for the linear equations for each value of t for high order m of the derivatives can be computationally intensive. A solution which can be approximate and quickly computed with rather smooth solutions is required. The pointwise method is sensitive to isolated singularities of the $\mathbf{Z}(t)'\mathbf{Z}(t)$ matrix which correspond to isolated singularities in the weight functions. By using sufficiently smooth weight functions we can bypass this problem.

Approximating the weight functions $w_j(t)$ via the use of basis functions is one strategy to use. Then with ϕ being the K-dimensional vector of the set of basis functions $(\phi_1, \ldots, \phi_K)$ we can get coefficients c_{jk} such that

$$w_j \approx \sum_k c_{jk}\phi_k$$

estimated form the data.

Now the fitting criterion for PDA can be written in terms of c in quadratic form:

$$\hat{F}(\dot{c} \mid \dot{x}) = C + c'\mathbf{R}c + 2c's$$

where the constant C does not depend on c and the estimate $\hat{c}$ is the solution to $\mathbf{R}c = -s$, where $\mathbf{R}$ is symmetric of order mK formed by $m \times m$ array of $K \times K$ submatrices of the form:

$$\mathbf{R}_{jk} = \frac{1}{N}\int \phi(t)\phi(t)' \sum_i D^j x_i(t) D^k x_i(t) dt$$

for $0 \leq j, k \leq m-1$. The mK-vectors has m subvectors s_j of length K defined as :

$$s_j = \frac{1}{N} \int \phi(t) \sum_i D^j x_i(t) D^m x_i(t) dt$$

for $j = 0, \ldots, m-1$.

Appendix C

numerical results for FPCA

Variability explained by components, expressed as %

[1]	44.442921	19.769229	11.845188	9.051975	3.371278	2.534699	1.763225
[8]	1.483607	0.937484	0.822239	0.655304	0.609027	0.553985	0.399082
[15]	0.228305	0.207018	0.181013	0.174710	0.139903	0.100226	0.096536
[22]	0.073265	0.067168	0.057585	0.054755	0.036247	0.034861	0.030879
[29]	0.028410	0.025589	0.020853	0.019667	0.017511	0.016812	0.013599
[36]	0.012377	0.010368	0.008894	0.008508	0.008202	0.008066	0.006607
[43]	0.005829	0.005303	0.004873	0.004644	0.004112	0.003748	0.003592
[50]	0.003153	0.002838	0.002488	0.002335	0.002145	0.001851	0.001742
[57]	0.001554	0.001422	0.001357	0.001341	0.001294	0.001123	0.001069
[64]	0.000986	0.000939	0.000897	0.000867	0.000783	0.000731	0.000696
[71]	0.000631	0.000598	0.000549	0.000514	0.000502	0.000474	0.000445
[78]	0.000411	0.000404	0.000359	0.000352	0.000331	0.000315	0.000301
[85]	0.000293	0.000272	0.000255	0.000245	0.000235	0.000214	0.000202
[92]	0.000200	0.000184	0.000166	0.000155	0.000139	0.000136	0.000119
[99]	0.000104	0.000000					

Appendix D

Numerical results for FLDA

Weights for the linear combination of the six principal components to calculate the discriminant are

PC1	PC2	PC3	PC4	PC5	PC6
0.3171116	0.1378731	-0.2338837	-0.2984332	-0.1181688	-0.8501197

which in turn yields the following discriminant values for each profile

	[,1]
[1,]	-0.42028766
[2,]	0.06376649
[3,]	0.01979505
[4,]	0.20899943
[5,]	-0.44176278
[6,]	-0.53490293
[7,]	-0.21860285
[8,]	-0.25557551
[9,]	0.36567275
[10,]	-0.13514305

```
[11,] -0.13024483
[12,] -0.41505898
[13,]  0.52145076
[14,] -0.35161937
[15,] -0.91112472
[16,] -0.03052841
[17,] -0.36330271
[18,]  0.13219491
[19,] -0.26978535
[20,] -0.03105860
[21,]  0.04214208
[22,]  0.04117799
[23,] -0.40710049
[24,] -0.91984566
[25,] -0.45560743
[26,] -0.60802034
[27,]  0.10395816
[28,] -0.72721960
[29,] -0.32203786
[30,]  0.05740435
[31,] -0.62997595
[32,]  0.23598660
[33,] -0.44235356
[34,]  0.71995799
```

```
[35,] -0.89614884
[36,]  0.44368194
[37,] -1.30483202
[38,] -0.85661026
[39,] -0.10181206
[40,] -0.11567428
[41,] -0.49237179
[42,] -0.49560579
[43,] -0.24149690
[44,] -0.44880228
[45,] -0.38639557
[46,] -0.42172850
[47,] -0.30105205
[48,]  0.59169814
[49,] -0.23493563
[50,] -0.25058912
[51,]  0.53624177
[52,] -0.93572547
[53,]  0.39248067
[54,] -0.21756093
[55,]  0.40479386
[56,] -0.17040201
[57,]  1.46647572
[58,]  0.54436286
```

```
[59,] -0.50109106
[60,]  0.91086324
[61,]  0.85756191
[62,]  1.85469110
[63,]  0.13192536
[64,] -0.08322844
[65,]  0.20429708
[66,] -0.48610425
[67,] -0.29208118
[68,]  0.33126534
[69,]  1.21442451
[70,] -0.15180907
[71,]  0.11769025
[72,] -0.31372668
[73,]  0.25061110
[74,] -0.06266782
[75,]  0.32676063
[76,] -0.36407293
[77,]  0.20155362
[78,]  1.68152856
[79,]  0.69445593
[80,] -0.44689770
[81,]  0.06772069
[82,]  0.28640130
```

```
 [83,]  0.27266485
 [84,] -0.15621011
 [85,] -0.70641371
 [86,] -0.07085139
 [87,] -0.38066811
 [88,]  1.03528491
 [89,] -0.09697904
 [90,]  1.01095709
 [91,]  1.05091386
 [92,] -0.20228581
 [93,]  0.65038023
 [94,]  0.45276689
 [95,]  0.40709332
 [96,]  0.15432109
 [97,]  1.82647742
 [98,] -1.06632710
 [99,] -0.90418405
[100,]  0.29364881
```

the critical value for discrimination $\hat{m}$ is 8.47395886820245e-17 The classification results according to this $\hat{m}$

```
    Truevalue Classified
1      normal     normal
```

2	normal	malignant
3	normal	malignant
4	normal	malignant
5	normal	normal
6	normal	normal
7	normal	normal
8	normal	normal
9	normal	malignant
10	normal	normal
11	normal	normal
12	normal	normal
13	normal	malignant
14	normal	normal
15	normal	normal
16	normal	normal
17	normal	normal
18	normal	malignant
19	normal	normal
20	normal	normal
21	normal	malignant
22	normal	malignant
23	normal	normal
24	normal	normal
25	normal	normal

26	normal	normal
27	normal	malignant
28	normal	normal
29	normal	normal
30	normal	malignant
31	normal	normal
32	normal	malignant
33	normal	normal
34	normal	malignant
35	normal	normal
36	normal	malignant
37	normal	normal
38	normal	normal
39	normal	normal
40	normal	normal
41	normal	normal
42	normal	normal
43	normal	normal
44	normal	normal
45	normal	normal
46	normal	normal
47	normal	normal
48	normal	malignant
49	normal	normal

50	normal	normal
51	malignant	malignant
52	malignant	normal
53	malignant	malignant
54	malignant	normal
55	malignant	malignant
56	malignant	normal
57	malignant	malignant
58	malignant	malignant
59	malignant	normal
60	malignant	malignant
61	malignant	malignant
62	malignant	malignant
63	malignant	malignant
64	malignant	normal
65	malignant	malignant
66	malignant	normal
67	malignant	normal
68	malignant	malignant
69	malignant	malignant
70	malignant	normal
71	malignant	malignant
72	malignant	normal
73	malignant	malignant

74	malignant	normal
75	malignant	malignant
76	malignant	normal
77	malignant	malignant
78	malignant	malignant
79	malignant	malignant
80	malignant	normal
81	malignant	malignant
82	malignant	malignant
83	malignant	malignant
84	malignant	normal
85	malignant	normal
86	malignant	normal
87	malignant	normal
88	malignant	malignant
89	malignant	normal
90	malignant	malignant
91	malignant	malignant
92	malignant	normal
93	malignant	malignant
94	malignant	malignant
95	malignant	malignant
96	malignant	malignant
97	malignant	malignant

```
98  malignant     normal
99  malignant     normal
100 malignant  malignant
```

It is worth mentioning that the final classification was done by leave-one-out cross-validation.