Functional Data Analysis with Applications in Biostatistics

Sarah Jane Ratcliffe

B.Sc.(Hons), University of Technology, Sydney, 1997

Submitted in fulfilment of the requirement for the degree of

Doctor of Philosophy in the Department of Statistics,

Division of Economic and Financial Studies, Macquarie University.

October, 2000

Contents

Summary

Functional data analysis is concerned with the analysis of data for which the observed responses for each subject are continuous curves. In practice, measurements are taken at discrete time points but estimates are required over the entire time interval. Traditional techniques for analysis of multiple curves, such as longitudinal data analysis or time series methods, are unsuitable for this type of data, since there are generally more measurements per subject than subjects and stationarity assumptions do not necessarily hold. With a technology induced growth in data of this kind, research into techniques for functional data analysis has become an emerging area in recent years.

This thesis aims to develop new techniques for functional data analysis, focusing on three problems: logistic regression with a functional regressor, linear and logistic regression for a repeatedly stimulated functional regressor, and a functional mixed-effects type model for joint mean and covariance modelling.

For each of the problems, we develop solutions using a basis function approach, that is, expressing the data for each subject as a linear combination of known basis functions. Using this approach we are able to overcome singularity problems associated with having more measurements than subjects. As well as calculating maximum likelihood or least squares parameter estimates, model diagnostic and smoothing parameter selection issues are addressed.

The techniques developed in this thesis are applied to novel biostatistical data sets: electroencephalographic data and fetal heart rate data. Of main interest is the fetal heart rate data, which motivated the development of the regression techniques for a repeatedly stimulated

functional parameter. It was found that the stimulated fetal heart rates could be used to predict an infant's risk category at birth and psychomotor development at 18 months of age.

Most of the material presented in the thesis is my own work. The exception is:

1. the work described in Section 6.3 is partly due to Victor Solo.

Acknowledgements

I would like to take this opportunity to thank a number of people, without whom this thesis would not have been possible.

Firstly, I wish to thank my supervisors, Professor Victor Solo and Dr Gillian Heller, for all of their help, encouragement and guidance over a number of years. The time spent working as a student under their supervision has been an invaluable learning experience.

Thanks also to Dr Leo Leader, from the School of Obstetrics and Gynaecology, University of New South Wales and the Royal Hospital for Women, for providing the fetal data, feedback on the analyses, medical background and other useful inputs. It has been great working with you throughout this thesis.

Dr Evian Gordon from the Department of Psychiatry, University of Sydney and Westmead Hospital provided the EEG data, while Dr Julian Leslie arranged for its use in this thesis. Thank you.

I am indebted to members of the Statistics Department of Macquarie University for their advice and encouragement throughout my study period. Also, thanks to many friends for their continued moral support.

Finally, I thank my parents, Fred and Beverley, and brother James for all of their support and patience during this degree.

Certificate

The work described in this thesis was carried out in the Department of Statistics, Division of Economic and Financial Studies, Macquarie University, New South Wales, Australia, between March 1997 and October 2000, under the supervision of Professor Victor Solo and Dr Gillian Heller.

This is to certify that the material presented in this thesis is original and, to the best of my knowledge and belief, contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma at a university or other institute of higher learning, except where due acknowledgement is made in the text.

October, 2000.

Sarah Jane Ratcliffe

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, developments in technology have resulted in the growth of collection and storage of data where the observed responses for each subject are continuous curves or functions or time series. In practice, measurements are taken at discrete time points so that the data are long time series but estimates of quantities of interest are required over the entire time interval of observation of the curves. Such data are referred to as *Functional Data*.

Functional data are actually a special type of longitudinal data. Traditionally, for longitudinal data, the number of measurements per subject p is much smaller than the number of subjects n. For functional data, usually the reverse is true; that is, usually $n \ll p$. Thus, if classical longitudinal data analysis techniques (eg. Diggle *et al.*, 1994) were used on functional data, singularity problems would occur. Another approach is to use time series methods (eg. Brillinger, 1981a) which also involve data with $n \ll p$. However, these methods assume that the covariance structure of the data is stationary. Because we have multiple subjects, it is possible to move away from the stationarity assumption by using *nonparametric smoothing methods*. Thus, new methods have been developed to analyse functional data and these methods have come to be known as *Functional Data Analysis*[1] (FDA) (Ramsay and Silverman, 1997).

[1] Traditional nonparametric methods might legitimately be called functional data analysis methods, since they deal with estimation of curves, but for this thesis we use the phrase as in Ramsay and Silverman (1997) where perhaps we might say functional longitudinal data analysis.

Any statistical technique available for the analysis of multivariate or longitudinal data is also desired for functional data but there are also new features. Of particular interest is functional growth curve modelling (e.g. Ramsay, 1982; Kneip and Gasser, 1992) and principal component analysis (e.g. Besse and Ramsay, 1986; Castro *et al.*, 1986; Solo, 1997). The various approaches rely on some form of smoothing and/or constraints to overcome the singularity problem resulting from more unknown parameters than subjects. Common approaches have been basis expansions, smoothing splines, kernel estimators, the addition of a roughness penalty term or some combination of these. In this thesis we aim to develop new methods for analysing functional data: functional logistic regression, joint mean and covariance modelling and regression techniques for repeatedly stimulated functional data.

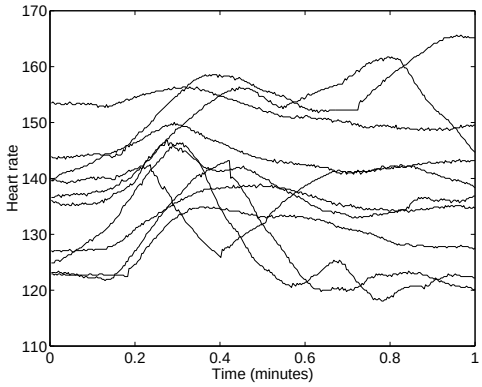
We begin this chapter by outlining some of the issues involved in functional data analysis. This is followed by an overview of each of the research areas listed above. Finally, we indicate the contents of the remaining chapters in Section 1.3.

1.1 Functional Data Analysis

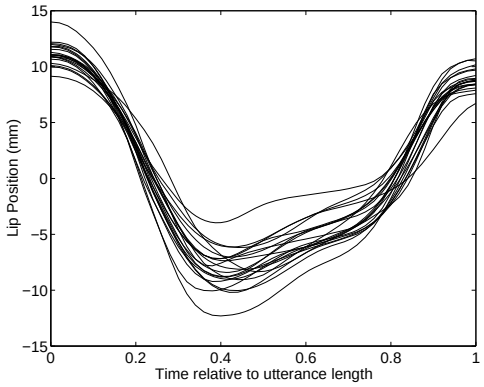
Functional data analysis is concerned with the analysis of multiple, non-stationary curves with $n \ll p$. The data for each subject is correlated but it is assumed that the data from different subjects are independent. Some examples of functional data are shown in Figure 1.1. These include heart rate tracings; recordings of the position of the centre of the lower lip during the utterance of the syllable "bob" (Ramsay *et al.*, 1996; Ramsay and Silverman, 1997); hip angles formed over the gait cycle (one double step) of walking children (Olshen *et al.*, 1989; Rice and Silverman, 1991); and daily smoke levels during the winters of 1958-1971 in London (Shumway *et al.*, 1983).

1.1.1 Data Display

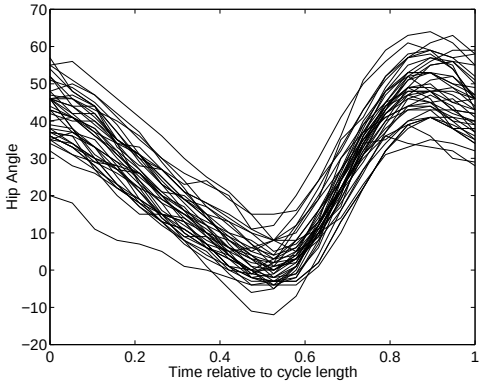
As seen from the examples, one issue for functional data is how the data should be displayed. A simple time plot of the data may not be useful, as shown by the London smoke data



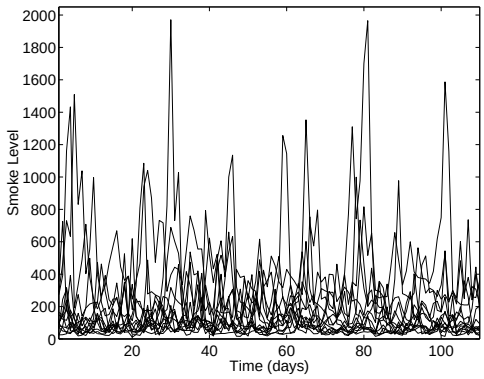
(a) Fetal heart rate tracings over a one minute period for 10 fetuses



(b) Position of the centre of the lower lip while uttering the syllable "bob"



(c) Hip angles of 39 children over a single gait cycle



(d) Daily British smoke levels during winter for 14 years

Figure 1.1: *Examples of functional data.*

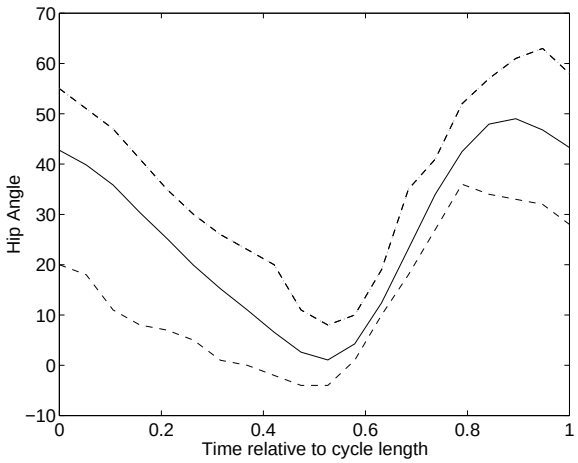


Figure 1.2: *Curves corresponding to an estimated mean curve (solid line) and the highest and lowest (dashed lines) scores on the first principal component for the hip data.*

(Figure 1.1(d)). For other data, eg. the lip data of Figure 1.1(b), this type of plot may be useful for seeing common patterns or trends but it is hard to identify an individual curve. Plotting the data individually or in small groups would allow the curves for each subject to be seen but this makes comparing the curves difficult. So, additional methods are needed. One suggestion, presented by Jones and Rice (1992), is to plot just a few representative curves. These curves were chosen to be the ones corresponding to the median, highest and lowest scores with respect to a particular principal component. (Functional principal component analysis is described in Section 2.5.2). A similar method was suggest by Rice and Silverman (1991) but instead of plotting the curve corresponding to the median value, they plotted the overall sample mean curve. For example, Figure 1.2 shows the mean curve for the hip data plus the individual curves corresponding to the highest and lowest scores for the main (first) principal component. These methods are effective at displaying the main sources of variability in the data. However, this type of plot will not be useful for all data sets. Other plotting ideas are yet to be developed. Thus, data display is a wide research area in itself. However, we do not address this issue further in this thesis.

1.1.2 Curve Registration

A problem encountered in some applications is that of curve registration: the alignment of curves at main features. The idea is to remove the time shifts in the peaks/troughs of the curves before analysis. These shifts are considered to be a nuisance in application since, eg. they may lead to severe bias. For example, Ramsay *et al.* (1995) examined the force exerted during a pinch made by the thumb and forefinger. These pinches commenced at an arbitrary time point for each curve, and masked the effect of interest. Thus, Ramsay *et al.* transformed the time scale for each curve so that the pinches started at the same point and were also aligned at the time of maximum force. Special cases of this problem have long been known in nonparametric time series (eg. cross-spectral estimation: Jenkins and Watts, 1968).

A number of transforms that could be used to align curves are described in Ramsay and Silverman (1997). These include the Procrustes method and time warping. Time warping has an extensive history in the engineering literature (e.g. Sakoe and Chiba, 1978) and adjusts the curves so that important features are aligned with their occurence in the mean function via a warping function. Kneip and Gasser (1992); Wang and Gasser (1997, 1999); Ramsay (1998, 2000), and others, have all used (dynamic) time warping. Another approach, as used by Silverman (1995) in functional principal component analysis, is to incorporate the time shift for each curve directly into the analysis by the addition of a parametric effect.

However, problems still exist with curve registration. In removing the time shifts, one could also be removing variation that is of interest. This may be overcome by aligning derivatives of the curves instead of the actual curves. Curve registration is a whole research area in itself and is not discussed further in this thesis.

1.1.3 Analysis

Having displayed and/or aligned the curves, the next step is to carry out some form of analysis. Classical longitudinal data analysis techniques (eg. Diggle *et al.*, 1994) are inappropriate for functional data. Since $n \ll p$, classical analysis would involve inverting a matrix of

less than full rank, resulting in singularity problems. Time series methods (eg. Brillinger, 1981a; Shumway, 1988) are not preferred, as the stationarity assumption does not necessarily hold. Thus, new techniques have been developed for the analysis of functional data. These techniques are an extension of nonparametric smoothing methods. If the functional data were stationary, then time series methods could be used as they would be more efficient. However, at present, there are no suitable methods for testing for stationarity of functional data. Developing such tests would be an area of future research.

Generally, in order to analyse the curves, the dimensionality of the data is reduced since there are less subjects than measurements per subject. Suppose n subjects have had p measurements of a variable y recorded at equispaced time points $t_1, \ldots, t_p$ on some interval J. Values of y are available at the discrete time points but we wish to estimate y over J. Then $y(t)$ is known as a functional variable and we can model the relationship between y and t for subject i, $y_i(t)$, using a nonparametric smoothing method (Section 2.1) such as basis functions, kernels, local polynomials or smoothing splines.

The basis function approach is the most commonly used method for functional data analysis (Ramsay and Silverman, 1997). In this method, the data for each subject is represented as a linear combination of known basis functions, $\psi_k(t)$, $k = 1, \ldots, m$,

$$y_i(t) = \sum_{k=1}^{m} c_{ik}\psi_k(t)$$

where the basis functions are chosen to reflect the characteristics of the data. The basis approach has been used alone and in conjuction with a penalty term (as also used by smoothing splines). There has been some use of smoothing splines to model each subject's curve. However, kernel and local polynomial methods have not yet received much development. Having modelled the data for each subject, these are then used in what are essentially multivariate data analysis techniques.

1.2 Areas of Research

This thesis aims to develop new techniques for FDA, in particular logistic regression with a functional regressor, linear and logistic regression for a repeatedly stimulated functional regressor, and joint functional mean and covariance modelling. The methods were developed using a basis function approach. Whilst other nonparametric smoothing methods could have been used, they are left for future research. In this section, we present an overview of the research issues.

Functional Logistic Regression

In the past, functional regression and ANOVA have concentrated on the cases of a continuous and a functional response variable, with continuous and/or functional regressors (see Section 2.5.1). The main approach used to calculate estimates has been a basis function approach. However, to the author's knowledge, no methods have been presented for modelling a binary response variable with a functional regressor. In this thesis, we develop such a functional logistic regression with maximum likelihood parameter estimation.

Functional Data with a Repeated Stimulus

In many cases, functional data have special structures that make standard regression techniques unsuitable. One such case is repeatedly stimulated functional data. The functional regressor now consists of two parts: the curve measured within the time frame of each stimulus, and the timing of the stimulus in relation to the other stimuli. Both parts need to be incorporated into the structure of any model.

In this thesis, both functional linear and logistic regression models were needed to analyse data with a repeated stimulus. Thus, we developed basis solutions to both problems. The functional linear regression model builds on the model given by Ramsay and Silverman (1997), while the functional logistic regression model for this type of data extends the functional linear regression model previously developed.

Functional Mean and Covariance Modelling

Principal component analysis (PCA) applied to functional data has been considered by several authors (see Section 2.5.2). These techniques are able to move away from the stationarity assumption inherent in time series modelling of the covariance function. However, they assume that the curves have a zero mean function: that is, $\mu(t) = 0$ for all t. Thus, before functional principal component analysis (fPCA) can be performed, the data needs to be mean adjusted. In most cases, this has been done by subtracting the average value at each discrete time point, though other nonparametric methods could be used.

Having mean adjusted the data, fPCA was developed by expanding the population covariance function in a Karhunen-Loeve expansion (Kanwal, 1971). Eigenfunctions were then estimated using some nonparametric smoothing technique; generally a basis function approach.

However, in most cases, estimating and subtracting the mean function from the data before fPCA is performed is inefficient. The mean and covariance functions need to be estimated together, as one estimate can affect the other. This leads to a functional mixed-effects type model; incorporating fPCA in the covariance estimate. A basis function approach for estimating the parameters in a functional mixed-effects type model is developed in this thesis.

1.3 Plan of Thesis

To recap, in this thesis we develop new methods for analysing functional data, in particular logistic regression with a functional regressor, linear and logistic regression for a repeatedly stimulated functional regressor, and joint functional mean and covariance modelling.

Chapter 2 provides technical background for the methods used in this thesis, as well as showing the development of FDA techniques. In particular, Section 2.1 provides an overview of nonparametric smoothing methods, namely the basis function approach, kernel methods, local polynomials and smoothing splines. A comparision of these methods is also given.

Section 2.2 deals with smoothing parameter selection while Section 2.3 outlines modifications needed for correlated data. Section 2.4 overviews existing methods for handling multiple curves: longitudinal data analysis and time series techniques. Finally, Section 2.5 describes techniques for functional regression and functional principal component analysis.

In Chapter 3, we model a functional regressor with a binary response variable, in what we call functional logistic regression. A basis function approach is used with and without an integrated squared second derivative penalty term. The resulting algorithm is easy to use, with only simple modifications needed to the existing logistic regression algorithm. Model diagnostic issues are also discussed and we use a cross-validated log-likelihood function to select the optimal number of basis functions and/or smoothing parameter. We apply functional logistic regression to electroencephalographic (EEG) data.

In Chapter 4, we derive algorithms for linear and logistic regression for a repeatedly stimulated functional regressor. The basis function approach is again used to generate estimates that are both least-squares and maximum likelihood. The development of the algorithms was motivated by the fetal heart rate data. The functional regressor from this data set consists of fetal heart rate tracings recorded at 0.2 second intervals for 19 minutes, with a noise stimulus applied at regular one minute intervals, giving the regressor a special structure.

We analyse the fetal heart rate data in Chapter 5, using the algorithms developed in Chapter 4. The two response variables presented in this thesis are the infant's risk category at birth (binary) and psychomotor development at 18 months (continuous). Functional models of these data represented a substantial improvement over the best standard linear / logistic models produced using the other covariates. It was found that the heart rates were significant in the prediction of an infant's risk category, and its development at 18 months of age. The need for the stimuli was established through the use of unstimulated control subjects, in the case of the risk category.

In Chapter 6, we present an iterative basis algorithm for the joint modelling of the mean

and covariance functions for functional data. It is essentially a functional mixed-effects type model; incorporating functional principal component analysis in the covariance estimate. Analysis of the algorithm showed that while a unique solution is not guaranteed, it does converge to the unique Moore-Penrose solution for some starting values. A simulation study was used to examine how well the algorithm performed, before applying it to the EEG data from Chapter 3 and human gait data from Rice and Silverman (1991).

Findings from Chapter 2-6 are summarised in Chapter 7, where recommendations for future research are also discussed. Further details of the Bayley scales used to determine variable values in the fetal heart rate data are given in Appendix A. A glossary of abbreviations and notation used in each chapter is provided in Appendix B. Appendix C documents the software developed during the course of this thesis. Finally, Appendix D lists the author's publications in the field of work contained in this thesis.

Chapter 2

Technical Background

This chapter gives a technical background to the development of functional data analysis. Sections 2.1 - 2.3 deal with nonparametric smoothing methods: descriptions of various smoothing methods (which are later used as the basis of models developed for functional data); smoothing parameter selection via cross-validation; and an outline of adjustments for dealing with correlated data. In these sections we keep to traditional notation and use **n for the number of time points rather than p**. Since these smoothing methods are concerned with the estimation of a single function, Section 2.4 then moves towards functional data analysis by describing existing techniques for multiple curves. Finally, we provide details on functional regression and functional principal component analysis in Section 2.5.

2.1 Nonparametric Smoothing Methods

Many nonparametric smoothing methods have been developed for function estimation. Two main areas requiring these techniques are nonparametric regression and density estimation. There is extensive literature on both of these areas. For example, surveys of nonparametric regression techniques are given in Härdle (1990); Hastie and Tibshirani (1990); and Green and Silverman (1994), and nonparametric density estimation in Silverman (1986); and Härdle (1991). We describe the use of nonparametric methods in regression, and where necessary density estimation.

Suppose we are interested in the relation between two continuous variables; t (scaled to the interval [0,1]) and y. If n observations of y are taken at preset values t_i (eg. equispaced at $t_i = i/n$; we don't consider random t_i in this thesis), $i = 1 \ldots n$, then the most common way of testing for a relationship between the resulting data pairs (t_i, y_i), $i = 1, \ldots, n$, has been via linear regression. That is, fitting the model

$$y_i = \alpha + \beta t_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

where ε_i are independent random errors with mean zero and common variance σ^2. However, a straight line model is inappropriate for many data sets. The response variable is better estimated by some sort of curve or function of t, say $g(t)$, giving a model

$$y_i = g(t_i) + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{2.1}$$

In general, a nonparametric estimate of $g(t)$ is just a weighted sum of observed responses,

$$\hat{g}(t) = n^{-1} \sum_{i=1}^{n} w_i(t) y_i$$

The main methods which have been developed to provide these weights, $w_i(t)$, are: basis functions, kernel estimators, local polynomials and smoothing splines and these are discussed in turn below. As well as providing an estimation of the function, these methods can provide estimates of derivatives, which may also be of interest. A special kind of basis method, namely wavelets (Chui, 1992; Antoniadis *et al.*, 1994; Donoho and Johnstone, 1994), is important for irregular functions. However, in this thesis we deal with smooth functions only and so wavelets are not discussed.

We illustrate each of the nonparametric regression techniques using birth rate data (Cook and Weisberg, 1994; Simonoff, 1996). This data consists of monthly birth rates in the United States between 1940 and 1947 (Figure 2.1). As noted by Cook and Weisberg, there appears to be an increase in the birth rate from January 1940 until the peak around August 1942 - September 1943. This peak would correspond to 9 - 14 months after the entry of the United States into World War II. There is a decline after this time, presumably due to the war, until approximately January 1946 (nine months after the end of the war in Europe) when the

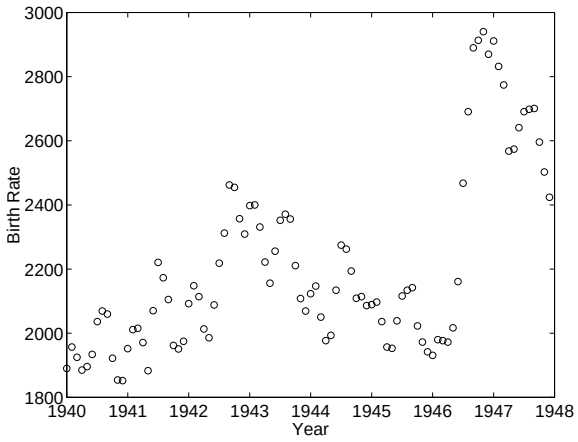


Figure 2.1: *Scatter plot of monthly birth rates in the United States.*

birth rate rises quickly. Nonparametric smoothing methods should provide function estimates which pick up these features in the data.

2.1.1 Basis Functions

The idea behind the basis method is to expand the curve or function to be estimated in terms of a linear combination of known basis functions, $\psi_k(t)$. i.e.

$$g(t) = \sum_{k=0}^{\infty} c_k \psi_k(t), \qquad 0 \leq t \leq 1 \tag{2.2}$$

where the coefficients c_k need to be estimated. The basis functions need not be orthogonal and so the simplest kind are polynomials, $\psi_k(t) = t^k, k = 0, 1, \ldots$. However, these can produce a nearly singular matrix in the calculation of $\hat{c}_k$ via least squares. This can be overcome by using orthogonal basis functions such as Legendre polynomials,

$$\begin{aligned} \psi_0(t) &= 1/\sqrt{2}, \\ \psi_1(t) &= t/\sqrt{2/3}, \\ (k+1)\psi_{k+1}(t) &= (2k+1)t\psi_k(t) - k\psi_{k-1}(t), \qquad k = 1, 2, \ldots \end{aligned}$$

or Fourier basis functions

$$\begin{aligned}\psi_0(t) &= 1,\\ \psi_{2k-1}(t) &= cos2\pi kt,\\ \psi_{2k}(t) &= sin2\pi kt, \qquad k = 1, 2, \ldots\end{aligned}$$

Having chosen the type of basis, the coefficients, and hence the function, can be estimated. In order to do this, the sum needs to be truncated at some point, say m,

$$g_m(t) = \sum_{k=0}^{m} c_k \psi_k(t)$$

This truncation also controls the smoothness of the estimated function, $\hat{g}(t)$. The fewer basis functions, the smoother the estimate. With a Fourier basis, truncation is also equivalent to applying a low-pass filter (as used in time series analysis) to the coefficients c_k.

One approach for calculating the basis estimator is best illustrated by an example. Suppose the unknown regression function $g(t)$ is to be estimated on the interval $J = [0, 1]$ using Fourier basis functions. Using orthogonality

$$\int_0^1 \psi_j(t)\psi_k(t)dt = \delta_{j,k}$$

where $\delta_{j,k}$ is the Kronecker delta function, the coefficients c_k can be found by the covariance calculation

$$c_k = \int_0^1 g(t)\psi_k(t)dt$$

However, this equation involves the unknown function $g(t)$. To develop an estimator for $\hat{c}_k$ over a set of disjoint subintervals $\{J_i\}_{i=1}^n$ spanning J and containing t_i, we first approximate the integral by a sum

$$\begin{aligned}c_k &= \sum_{i=1}^{n} \int_{J_i} g(t)\psi_k(t)dt\\ &\approx \sum_{i=1}^{n} g(t_i) \int_{J_i} \psi_k(t)dt\end{aligned}$$

provided $g(t)$ is smooth and J_i is not too big. It is now natural to estimate $\hat{c}_k$ by replacing $g(t_i)$ by y_i. Thus

$$\hat{c}_k = \sum_{i=1}^{n} y_i \int_{J_i} \psi_k(t)dt \tag{2.3}$$

When the basis is not an orthogonal basis, the coefficients can be estimated using multiple regression techniques (see for example Seber, 1977). Substitute the basis expansion for $g(t)$ into the regression model (2.1) to give

$$y_i = \sum_{k=0}^{m} c_k \psi_k(t_i) + \varepsilon_i$$

The c_k can then be estimated via least squares. When $t_i = i/n$, n is large and the basis is orthogonal, this estimator is nearly the same as (2.3).

The most commonly used non-orthogonal basis is the B-spline basis. B-splines are piecewise polynomials that are joined together in a continuous fashion at values τ_j, $j = 1, \ldots, \kappa$ called *knots*. These knots are chosen so that they span $[0, 1]$, the interval of the data, and satisfy $0 = \tau_1 < \ldots < \tau_\kappa = 1$. The k-th B-spline basis function, $k = 1, \ldots, m$, of degree d (order $d + 1$) can be defined recursively as (Cox, 1972; de Boor, 1972, 1978) $\psi_k(t) = N_{k,d}(t)$ where

$$N_{k,d}(t) = \frac{t - \tau_k}{\tau_{k+d} - \tau_k} N_{k,d-1}(t) + \frac{\tau_{k+d+1} - t}{\tau_{k+d+1} - \tau_{k+1}} N_{k+1,d-1}(t)$$

$$N_{k,0}(t) = \begin{cases} 1, & \text{if } \tau_k \leq t < \tau_{k+1} \text{ or } t = \tau_\kappa \text{ when } k = m, \\ 0, & \text{otherwise.} \end{cases}$$

That is, the k-th B-spline basis function is a polynomial of degree d on the subinterval $[\tau_k, \tau_{k+d+1})$ and zero elsewhere. Also, the polynomials that meet at an interior knot match in the values of a set number of derivatives, usually $d - 1$.

B-splines are widely used as they have some nice properties. They are non-negative, linearly independent functions with compact or local support. That is, $\psi_k(t) = 0$ if $t \notin [\tau_k, \tau_{k+d+1})$. Hence, estimating c_k via least squares is computationally quick since the matrix to be inverted will be a $2d-1$ banded matrix. Also, for an arbitrary knot span $[\tau_j, \tau_{j+1})$, the basis functions

form a partition of unity. That is, $\sum_{j-d}^{j} \psi_k(t) = 1$. All derivatives of $\psi_k(t)$ exist in the interior knot spans, and $\psi_k(t)$ is $d-1$ times continuously differentiable at the knots. Thus, for B-splines the derivatives of the basis estimator $\hat{g}(t)$ can be easily computed. For example, the first derivative of $\hat{g}(t)$ can be found by simply differencing the estimated coefficients, $\hat{c}_k$. Finally, $\psi_k(t)$ attains exactly one maximum value, except for degree zero B-splines. Properties of B-splines with non-distinct knots and algorithms for calculating B-spline basis functions can be found in de Boor (1978); Schumaker (1981); Piegl and Tiller (1997), for example.

One limitation of B-splines is that the solution is sensitive to both the number and placement of the knots. In general, more knots should be placed in areas of high curvature and fewer in areas where $g(t)$ appears relatively smooth. Methods developed for the selection of the knots include Friedman and Silverman (1989) but are not discussed here. Since $t_i = i/n$, we use $\kappa = m + 1 - d$ equispaced knots on $[0, 1]$.

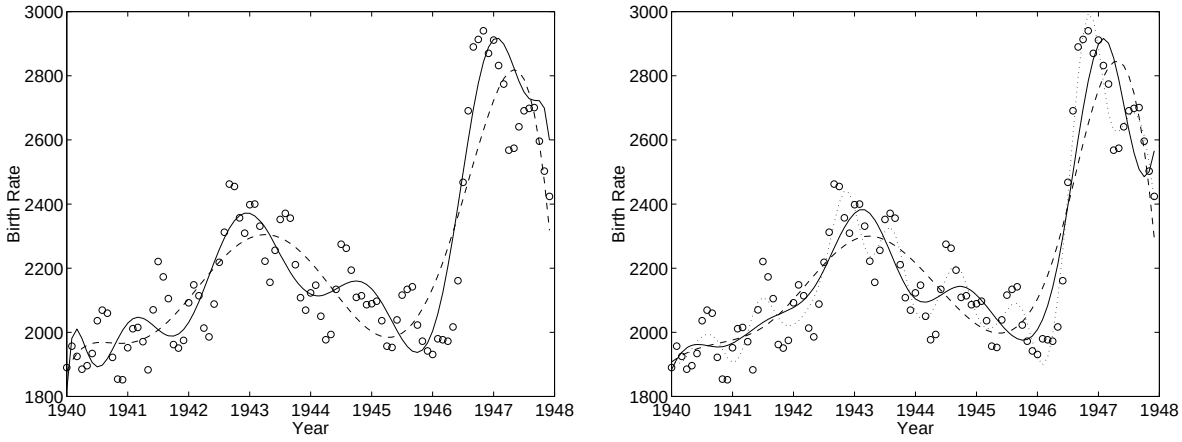
Figure 2.2 illustrates the results of the basis method for the birth rate data. Two different bases, Legendre polynomials and B-splines, and three different m values have been used. As m increases the basis estimators become less smooth, particularly the Legendre polynomial estimator.

In general, basis function estimators are relatively easy to compute and use. Basis estimators are also consistent under mild regularity conditions (Härdle, 1990)

$$\lim_{m \to \infty} P\left(|\hat{g}_m(t) - g(t)| \leq \varepsilon\right) = 1$$

provided as $n, m \to \infty$, $m/n \to 0$ and the t_i are equispaced. Thus, the number of basis functions should grow at a slower rate than the sample size. Consistency is not guaranteed if the t_i are not equispaced but we only deal with equispaced data in this thesis.

The type of basis functions used should be chosen with care. Basis functions that work well with one set of data may be unsuitable for another. This is particularly important as the behaviour of $g(t)$ at the boundaries affects the rate of convergence of the estimator, as defined



(a) Legendre polynomial basis estimators with $m = 8$ (dashed line) and $m = 13$ (solid line) basis functions

(b) B-spline basis estimators with $m = 8$ (dashed line), $m = 13$ (solid line), and $m = 25$ (dotted line) basis functions

Figure 2.2: *Basis estimators for the birth rate data.*

by the pointwise mean square error (MSE). For example, a Fourier basis estimator (Eubank, 1988)

$$\hat{g}_m(t) = \sum_{k=-m}^{m} c_m k e^{2\pi i k t}$$

for regularly spaced, periodic data has (Eubank, 1988)

$$\text{Bias}\left[\hat{g}_m(t)\right] = \sum_{|k|>m} c_k e^{2\pi i k t} - \sum_{|k|\leq m} \left(\sum_{r\neq 0} c_{k+nr} \right) e^{2\pi i k t}$$

$$\text{Var}\left[\hat{g}_m(t)\right] = n^{-1}\sigma^2(2m+1)$$

where

$$c_k = \int_0^1 g(s) e^{-2\pi i k s} ds$$

Hence, the optimal basis size is $m \approx n^{1/5}$ and the optimal pointwise MSE convergence rate is $O\left(n^{-4/5}\right)$. However, if non-periodic data is used with a Fourier basis estimator, the optimal convergence rate is $O\left(n^{-3/4}\right)$. Also, unless m is large, basis estimators cannot exhibit very local features, as seen in Figure 2.2. Kernel, local polynomial, and spline estimators were developed to overcome this problem. Further statistical properties of basis estimators can be found in the above mentioned references, and references found therein.

2.1.2 Kernel Estimators

Kernel estimators were first introduced by Rosenblatt (1956) and Parzen (1962) for density estimation, with statistical properties being calculated by Whittle (1958). A considerable literature has developed since then. A brief outline of their use is given here. For further information see, for example, Wand and Jones (1995); Härdle (1991) and references therein.

We begin with kernel density estimation. The idea is to construct functions centered at each data point and then average across these to produce the density estimate. This results in the kernel density estimator, given by

$$\hat{f}_h(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right)$$

where $K(u)$ is the kernel function and h is the window width, or smoothing parameter, which controls the width of the kernel functions.

Kernels are usually non-negative, symmetric functions. The most commonly used kernels are second order kernels satisfying

$$\int K(u)du = 1 \qquad \int uK(u)du = 0 \qquad \int u^2 K(u)du = k_2 \neq 0$$

Since kernels are themselves density functions, it follows that $\hat{f}_h(y)$ will also be a density function. The pointwise asymptotic bias and variance of the kernel density estimator are (eg. Silverman, 1986)

$$\begin{aligned} \text{Bias}\left[\hat{f}_h(y)\right] &= E\hat{f}_h(y) - f(y) \\ &= \frac{1}{2}h^2 f''(y)k_2 + o\left(h^2\right), \qquad h \to 0 \\ \text{Var}\left[\hat{f}_h(y)\right] &= \frac{1}{nh} f(y) \int K^2(u)du + o\left(\frac{1}{nh}\right), \qquad nh \to \infty \end{aligned}$$

Thus, the bias depends only on the window width h, while the variance depends on both h and n. A small h value will decrease the bias but increase the variance of the estimator. The pointwise mean square error of the estimator is

$$\begin{aligned} \text{MSE}\left[\hat{f}_h(y)\right] = \frac{1}{nh} f(y) \int K^2(u)du + \frac{1}{4}h^4 \left(f''(y)\right)^2 k_2^2 \\ + o\left(\frac{1}{nh}\right) + o\left(h^4\right), \qquad h \to 0,\ nh \to \infty \end{aligned} \tag{2.4}$$

Kernel	$\boldsymbol{K(u)}$	$\boldsymbol{k_2 = \int u^2 K(u) du}$	$\boldsymbol{\int K^2(u) du}$
Uniform	$\frac{1}{2} I(\lvert u\rvert \leq 1)$	1/3	1/2
Triangular	$(1 - \lvert u\rvert)\, I(\lvert u\rvert \leq 1)$	1/6	2/3
Epanechnikov	$\frac{3}{4}(1 - u^2)\, I(\lvert u\rvert \leq 1)$	1/5	3/5
Biweight	$\frac{15}{16}(1 - u^2)^2\, I(\lvert u\rvert \leq 1)$	1/7	5/7
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$	1	0.2821

Table 2.1: *Some common kernel functions.*

Thus, provided $h \to 0$ and $nh \to \infty$, the MSE will converge to zero, and hence, the estimator will be consistent. From (2.4), the optimal rate of convergence for the MSE is $O\left(n^{-4/5}\right)$, using an optimal window width of $h \approx n^{-1/5}$.

Nadaraya (1964) and Watson (1964) used the kernel density estimator to develop a kernel regression estimator as follows,

$$\hat{g}_h(t) = \frac{1}{nh} \sum_{i=1}^{n} \frac{K(\frac{t-t_i}{h})}{\hat{f}_h(t)} y_i \tag{2.5}$$

It is again a weighted average with the divisor $\hat{f}_h(t)$ ensuring the weights sum to unity. Other kernel estimators have been developed; including the Priestly-Chao estimator (Priestly and Chao, 1972) and the Gasser-Müller estimator (Gasser and Müller, 1979). These also estimate $g_h(t)$ using kernel functions but with a different weight function, $w(t)$, to that used by the Nadaraya-Watson estimator.

Some common kernel functions are given in Table 2.1 The smoothness of the estimated functions are controlled by the window width, h. As $h \to 0$, the estimated regression function with exhibit more local behaviour. This is illustrated in Figure 2.3. As h becomes smaller, the estimator becomes less smooth. In particular, a very small value of h results in an estimator which simply interpolates the data. So, we can regard h as a measure of the size of the smallest feature or "bump" that can be estimated in the function.

Kernel estimators, in general, are easily computed and have desirable statistical properties.

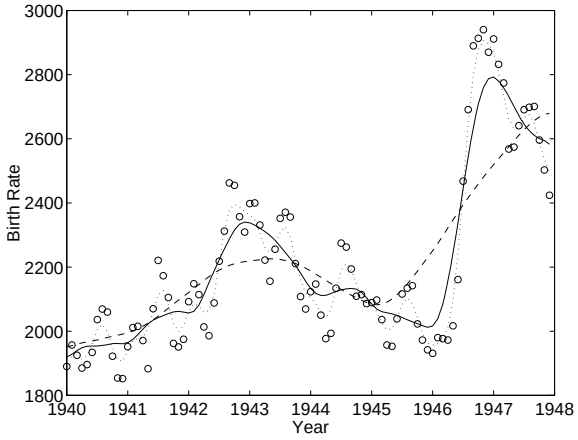


Figure 2.3: *Kernel estimators for the birth rate data with* $h = 0.2$ *(dashed line),* $h = 0.07$ *(solid line), and* $h = 0.03$ *(dotted line).*

Let $\hat{g}_h(t) = \hat{r}_h(t)/\hat{f}_h(t)$, $\hat{r}_h(t) = \frac{1}{nh}\sum K(\frac{t-t_i}{h})y_i$, then for the Nadaraya-Watson estimator, the pointwise asymptotic bias, variance and MSE of $\hat{r}_h(t)$ is the same as that for $\hat{f}_h(t)$. That is, (Härdle, 1991)

$$\begin{aligned}
\text{Bias}\left[\hat{r}_h(t)\right] &= \frac{1}{2}h^2 r''(t)k_2 + o\left(h^2\right), \qquad h \to 0 \\
\text{Var}\left[\hat{r}_h(t)\right] &= \frac{1}{nh}r(t)\int K^2(u)du + o\left(\frac{1}{nh}\right), \qquad nh \to \infty \\
\text{MSE}\left[\hat{r}_h(t)\right] &= \frac{1}{nh}r(t)\int K^2(u)du + \frac{1}{4}h^4\left(r''(t)\right)^2 k_2^2 \\
&\quad + o\left(\frac{1}{nh}\right) + o\left(h^4\right), \qquad h \to 0,\ nh \to \infty
\end{aligned}$$

Thus, $\hat{r}_h(t)$ is also a consistent estimator. Combining the results for $\hat{r}_h(t)$ and $\hat{f}_h(t)$, the approximate MSE for the Nadaraya-Watson estimator is (Härdle, 1991)

$$\begin{aligned}
\text{MSE}\left[\hat{g}_h(t)\right] \approx \frac{1}{nh}\frac{\sigma^2(t)}{f(t)}\int K^2(u)du + \frac{1}{4}h^4\left(g''(t) + 2\frac{g'(t)f'(t)}{f(t)}\right)^2 k_2^2 \\
+ o\left(nh^{-1}\right) + o\left(h^4\right), \qquad h \to 0,\ nh \to \infty
\end{aligned}$$

where $\sigma^2(t)$ is the conditional variance. Hence, if $h \to 0$ and $nh \to \infty$, the Nadaraya-Watson kernel regression estimator $\hat{g}_h(t)$ is also pointwise consistent. The optimal trade-off between the bias and variance is achieved when $h \approx n^{-1/5}$. Using this h value, the optimal MSE rate of convergence for the estimator is $O\left(n^{-4/5}\right)$. Note that, comparing the above expressions with those for the basis estimator shows that $1/m$ may be regarded as a window width.

However, traditional kernel estimators have a boundary bias problem (see Wand and Jones, 1995, and references therein). The bias has a different order of magnitude at boundary points as compared to interior points. For example, the Nadaraya-Watson estimator has boundary bias of $O(h)$. This is due to part of the kernel window having no data at boundary points. This boundary bias results in an inflated optimal MSE convergence rate. For example, the rate for the Nadaraya-Watson estimator is $O\left(n^{-2/3}\right)$ near the boundaries. Kernels that are modified near the boundary have been used to correct this asymptotic discrepancy (Gasser and Müller, 1979). Another advantage of kernel estimators is that asymptotic confidence intervals for $\hat{g}_h(t)$ can be calculated (Härdle, 1990, Section 4.2); as can derivatives of $\hat{g}_h(t)$.

2.1.3 Local Polynomial Regression

Local polynomial regression finds the estimate of the regression function at any point, t, by fitting a degree m polynomial to the data nearby t via weighted least squares. The weights are chosen so that data points close to t have large weights, with the weights decreasing as the points become further away from t. Thus, the local polynomial estimator at t, $\hat{g}_h(t)$, is the intercept term $\hat{c}_0(t)$ found by minimising

$$\sum_i w_h(t_i - t)\left(y_i - \sum_0^m c_k(t_i - t)^k\right)^2 \tag{2.6}$$

where $w_h(\cdot)$ is a hump shaped weight function centered at 0. Common weight functions are nearest neighbour weights and kernel weights found using kernel functions. Nearest neighbour weight functions are symmetric, non-increasing functions for $t \geq 0$. At each t, the weight function is centered and scaled so that at the r-th nearest neighbour (data point) of t the weight function is zero. Weight functions satisfying these properties are known as r-th nearest neighbour estimators. Figure 2.4 illustrates local linear ($m = 1$) polynomial estimation on the birth rate data using Epanechnikov kernel weights.

Like the kernel estimator, the local polynomial estimator is linear in the data. This can be more clearly seen by expressing the estimator using an *equivalent kernel*. An equivalent

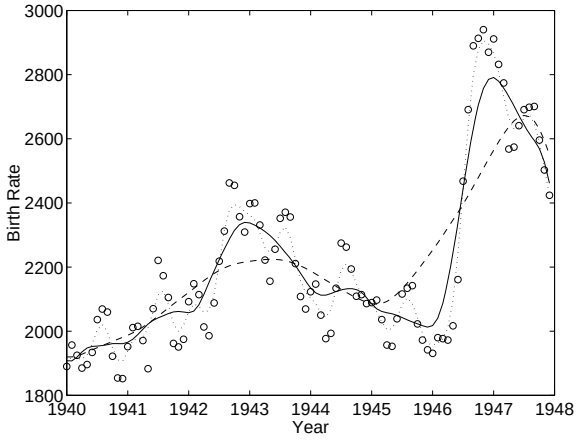


Figure 2.4: *Local linear polynomial estimators for the birth rate data with* $h = 0.03$ *(dashed line),* $h = 0.07$ *(solid line), and* $h = 0.2$ *(dotted line).*

kernel is defined as (Gasser *et al.*, 1985)

$$K_\nu^* = \text{the } (\nu+1)\text{th element of } \; S^{-1}\,[1\; t\; \ldots\; t^m]^T\, K(t)$$

where

$$S = [\mu_{j+l}]_{0\le j,l\le m}$$
$$\mu_j = \int u^j K(u)du$$

and $K(u)$ is a kernel function. For example, if an Epanechnikov kernel function was used to generate the weights the equivalent kernel would be (see Figure 2.5)

$$K_\nu^*(u) = \sum_{j=0}^{m+1} \lambda_j u^j$$

where for a local linear curve estimation ($m = 1$, $\nu = 0$)

$$\lambda_j = \begin{cases} 0 & \text{if } j+2 \text{ odd,} \\ \frac{3\,(-1)^{j/2}(2+j)!}{8\,j!\,(j+1)\,(\frac{2-j}{2})!\,(\frac{2+j}{2})!} & \text{if } j+2 \text{ even.} \end{cases}$$

For a local cubic ($m = 3$) curve estimation

$$\lambda_j = \begin{cases} 0 & \text{if } j+4 \text{ odd,} \\ \frac{15\,(-1)^{j/2}(4+j)!}{64\,j!\,(j+1)\,(\frac{4-j}{2})!\,(\frac{4+j}{2})!} & \text{if } j+4 \text{ even.} \end{cases}$$

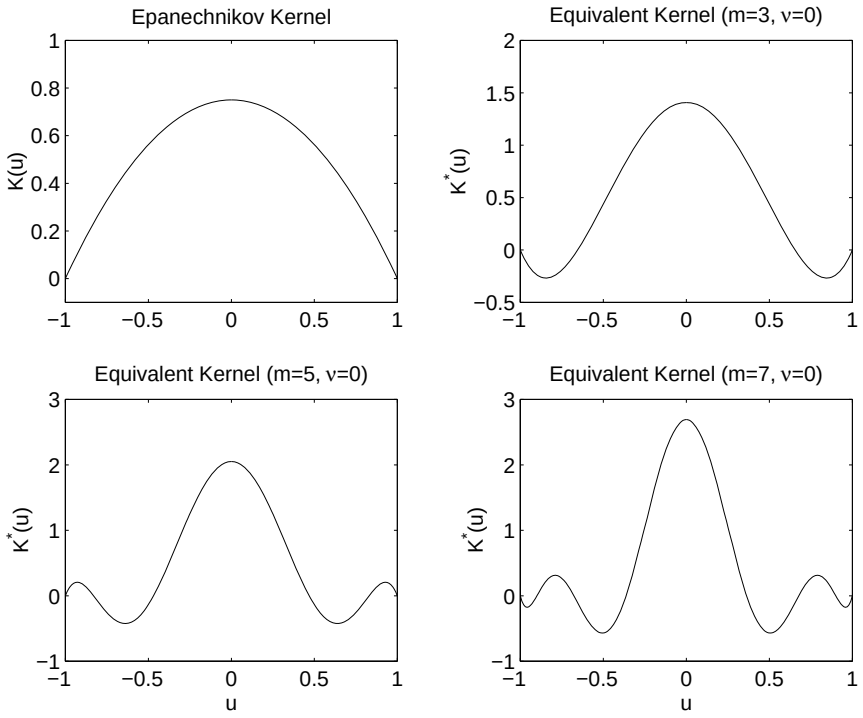


Figure 2.5: *The Epanechnikov kernel and its equivalent kernel for local polynomial curve estimation, for some values of* m.

Thus, we can rewrite the local polynomial estimator (or any derivative of the estimator) as (Fan and Gijbels, 1996)

$$\hat{g}_h^{(\nu)}(t) = \hat{c}_\nu(t) = \frac{1}{nh^{\nu+1}f(t)} \sum_1^n K_\nu^* \left(\frac{t_i - t}{h} \right) y_i \left(1 + o\left(1\right)\right)$$

A special case of local polynomial regression is the Nadaraya-Watson kernel estimator (2.5). It corresponds to fitting degree zero polynomials with a kernel weight function. In fact, local polynomial regression is closely related to kernel estimation in general (Müller, 1987). The degree of the polynomial plays the role of the order of the kernel. It is also similar to the basis estimator. Instead of finding the coefficients for $g(t)$ via least squares, it finds them via weighted least squares. However, local polynomial estimators have superior properties, especially when the t_i are random. For example, kernel estimators result in either an increased bias (eg. Nadaraya-Watson estimator) or variance (eg. Gasser-Müller estimator) when the t_i

are random. Local polynomials adapt to random t_i's without increasing the bias or variance. They also have the advantage of no boundary bias problem. That is, the bias doesn't increase near the boundaries, as for other estimators, but remains constant over the entire data interval.

The conditional asymptotic bias and variance of $\hat{g}_h(t)$ are given by (Simonoff, 1996, and references therein)

$$\text{Bias}\left[\hat{g}_h(t)|t_1,\ldots,t_n\right] = \begin{cases} \frac{h^{m+1}g^{(m+1)}(t)\mu_{m+1}(K_{(m)})}{(m+1)!} + o\left(h^{m+1}\right), & \text{if } m \text{ is odd} \\ h^{m+2}\left[\frac{g^{(m+1)}(t)f'(t)}{f(t)(m+1)!} + \frac{g^{(m+2)}(t)}{(m+2)!}\right] & \\ \qquad \times\mu_{m+2}(K_{(m)}) + o\left(h^{m+2}\right), & \text{if } m \text{ is even} \end{cases}$$

$$\text{Var}\left[\hat{g}_h(t)|t_1,\ldots,t_n\right] = \frac{\sigma^2(t)\int K_{(m)}^2(u)du}{nhf(t)} + o\left(\frac{1}{nh}\right)$$

where $\mu_q(K_{(m)}) = \int u^q K_{(m)}(u)du$ and $K_{(m)}$ is a $(m+1)$th order kernel function when m is odd and a $(m+2)$th order kernel function when m is even. So for example, a local linear polynomial estimator $(m=1)$ would have asymptotic conditional bias of $O\left(h^2\right)$ and conditional MSE of

$$\begin{aligned}\text{MSE}\left[\hat{g}_h(t)|t_1,\ldots,t_n\right] = &\frac{1}{4}h^4\left(g''(t)\mu_2(K_{(1)})\right)^2 \\ &+ \frac{\sigma^2(t)\int K_{(1)}^2(u)du}{nhf(t)} + o\left(h^4\right) + o\left(\frac{1}{nh}\right)\end{aligned}$$

A local quadratic polynomial estimator $(m=2)$ would have conditional bias of $O\left(h^4\right)$ and

$$\begin{aligned}\text{MSE}\left[\hat{g}_h(t)|t_1,\ldots,t_n\right] = &h^8\left(\frac{g^{(3)}(t)f'(t)}{6f(t)} + \frac{g^{(4)}(t)}{24}\right)^2\mu_4(K_{(2)})^2 \\ &+ \frac{\sigma^2(t)\int K_{(2)}^2(u)du}{nhf(t)} + o\left(h^8\right) + o\left(\frac{1}{nh}\right)\end{aligned}$$

A local cubic estimator would also have conditional asymptotic bias of $O\left(h^4\right)$ but the bias, and hence the MSE, has a simpler expression

$$\begin{aligned}\text{MSE}\left[\hat{g}_h(t)|t_1,\ldots,t_n\right] = &\frac{1}{24}h^8\left(g^{(3)}(t)\mu_4(K_{(3)})\right)^2 \\ &+ \frac{\sigma^2(t)\int K_{(3)}^2(u)du}{nhf(t)} + o\left(h^8\right) + o\left(\frac{1}{nh}\right)\end{aligned}$$

So, the degree of the fitted polynomials determines the order of the bias for the estimator. Generally, even degree polynomial ($m = 0, 2, \dots$) fits are not recommended as their asymptotic bias involves $f'(t)$ and they are outperformed by polynomial fits with an odd degree ($m = 1, 3, \dots$).

The optimal asymptotic conditional MSE is achieved using the Epanechnikov kernel to generate the weights (Fan *et al.*, 1995). The optimal window width is then $h \approx n^{-1/(2m+3)}$ for odd m and $h \approx n^{-1/(2m+5)}$ for even m. These result in optimal MSE convergence rates of $O\left(n^{-(2m+2)/(2m+3)}\right)$ and $O\left(n^{-(2m+4)/(2m+5)}\right)$ respectively. For example, local linear, quadratic and cubic polynomial estimators would have optimal MSE convergence rates of $O\left(n^{-4/5}\right)$, $O\left(n^{-8/9}\right)$ and $O\left(n^{-8/9}\right)$, respectively.

However, local polynomial estimators can be sensitive to outliers. A robust extension was proposed by Cleveland (1979) for smoothing scatter plots which overcomes this problem. His idea was to iteratively fit the estimator using nearest neighbour weights. The weights are updated at each iteration according to the residuals of the previous fit; small residuals give large weights and large residuals small weights. His estimator is known as Locally Weighted Scatter plot Smoothing or *Loess* (or lowess). When the t_i are equispaced, the loess estimator corresponds exactly to (2.6) (except at the boundaries).

Local polynomial fitting has been used for many years to smooth time series data (Macauley, 1931). Other important references for local polynomials include Stone (1977); Müller (1987); Cleveland and Devlin (1988); Fan (1992, 1993), and Hamming (1989); Kendall and Ord (1990) in the time series domain. Fan in particular illustrated the MSE properties of $\hat{g}_h(t)$ and that local polynomials with kernel weights have certain minimax optimality properties. A description of local polynomial regression and its properties can also be found in Fan and Gijbels (1996); Simonoff (1996), for example.

2.1.4 Penalty Method - Smoothing Splines

A spline is a piecewise polynomial. The polynomials are joined at points, called knots, such that they form a continuous function with a specified number of continuous derivatives. A spline of order r has piecewise polynomials of degree $r-1$, $r-2$ continuous derivatives and an $(r-1)$st derivative that is a step function with jumps at the knots. In addition, a natural spline on $[0,1]$ of order $r = 2d$ satisfies the boundary conditions $g^{(j)}(0) = g^{(j)}(1) = 0$, $j = d, \ldots, r-1$.

Smoothing splines were derived in nonparametric regression as a solution to the following minimisation problem. Find $g(t)$ to minimise

$$n^{-1} \sum (y_i - g(t_i))^2 + h^{2d} \int_0^1 \left(g^{(d)}(t)\right)^2 dt \tag{2.7}$$

where $h > 0$ is the smoothing parameter. The penalty term is used to explicity control the smoothness of the estimated function. If $h = 0$, the estimator would simply interpolate the data, passing through each of the y_i. As $h \to \infty$, the smoothing spline estimator approaches the least squares regression line. Thus, the penalty term gives a tradeoff between the fit to the data and the smoothness of the estimator.

Schoenberg (1964), and independently Reinsch (1967), showed that the minimiser of (2.7) is a natural polynomial spline, although the basic idea of splines is attributed to Whittaker (1923). In particular, they showed that the natural cubic spline ($r = 4$) with knots at the t_i's is the unique solution to the minimisation of

$$n^{-1} \sum (y_i - g(t_i))^2 + h^4 \int_0^1 (g''(t))^2 dt$$

with $g''(0) = g''(1) = g^{(3)}(0) = g^{(3)}(1) = 0$. This is the most commonly used and widely studied smoothing spline (eg. Wahba, 1990). A simple derivation of the smoothing spline using basic matrix analysis is given in Solo (2000). Figure 2.6 shows the natural cubic smoothing spline estimator for the birth rate data.

A smoothing spline estimator of $g(t)$ can be written in general form, as the weighted sum

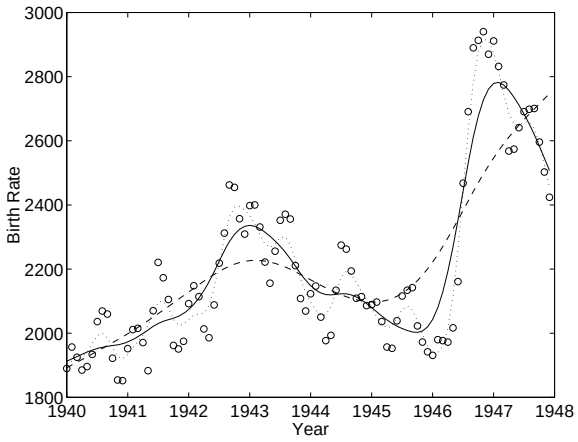


Figure 2.6: *Cubic smoothing spline estimators for the birth rate data with* $h = 0.5$ *(dashed line),* $h = 0.13$ *(solid line), and* $h = 0.0005$ *(dotted line).*

of the responses. The weight function for splines is actually equivalent in form to a kernel estimator. This is most easily understood in the case where $t_i = i/n$. In this case, changing to a Fourier series by Parseval's theorem, (and taking $[a, b] = [0, 1]$),

$$\begin{aligned} & n^{-1} \sum \left(y_i - g(i/n)\right)^2 + h^4 \int_0^1 \left(g''(t)\right)^2 dt \\ & \approx \sum \left(\tilde{y}_k - \tilde{g}_k\right)^2 + h^4 \sum k^4 \tilde{g}_k^2 \end{aligned}$$

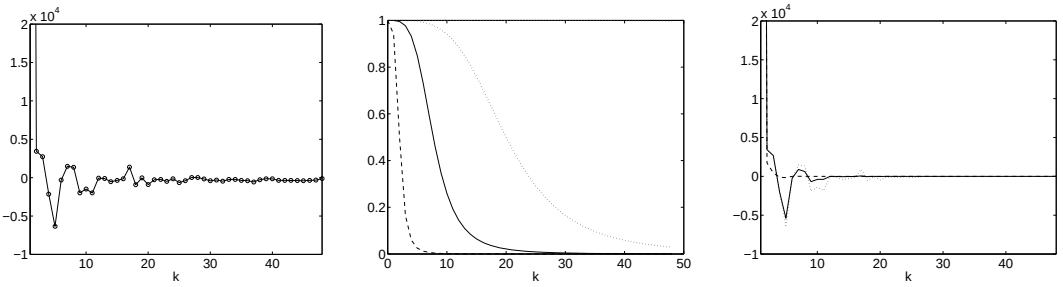
where $\tilde{g}_k = \sum g(i/n) \exp(-j2\pi k \frac{i}{n}) \approx \int g(t) \exp(-j2\pi kt) dt$ are the Fourier coefficients of $g(t)$ and similarly $\tilde{y}_k = \sum y_i \exp(-j2\pi k \frac{i}{n})$. Differentiating with respect to $\tilde{g}_k$ gives

$$\begin{aligned} & -\left(\tilde{y}_k - \tilde{g}_k\right) + h^4 k^4 \tilde{g}_k = 0 \\ & \Rightarrow \ \tilde{g}_k = \frac{\tilde{y}_k}{1 + (hk)^4} \end{aligned}$$

This exhibits the Fourier coefficients of the function estimator as a product of the data Fourier coefficients and a low pass filter. Taking inverse Fourier series gives $\hat{g}(t)$ as a convolution

$$\begin{aligned} \Rightarrow \ \hat{g}(t) &= \sum \frac{1}{h} K\left(\frac{t - i/n}{h}\right) y_i \\ K(u) &= \text{ inverse Fourier transform of } \frac{1}{1 + k^4} \\ &\approx \frac{1}{2} \exp\left(-|u|\right) \sin\left(|u| + \frac{\pi}{4}\right) \end{aligned}$$

For example, Figure 2.7 shows the Fourier coefficients of the birth rate data, the low pass filter, $\left(1 + (hk)^4\right)^{-1}$, and the resulting Fourier coefficients of the estimator.



(a) Fourier coefficients for the birth rate data, $\tilde{y_k}$.

(b) The lower pass filter, $\left(1+(hk)^4\right)^{-1}$.

(c) Fourier coefficients for the smoothing spline estimator, $\tilde{g_k}$.

Figure 2.7: *Fourier coefficients for the birth rate data, the low pass filter and the resulting Fourier coefficients for the smoothing spline estimator with* $h = 0.5$ *(dashed line),* $h = 0.13$ *(solid line), and* $h = 0.05$ *(dotted line).*

Continuing with the Fourier analysis, we can derive approximate pointwise asymptotic properties of the natural smoothing spline, thus;

$$\begin{aligned}
\hat{\tilde{g}}_k &= \frac{\tilde{y}_k}{1+(hk)^4} \\
\Rightarrow\ E(\hat{\tilde{g}}_k) &\approx \frac{\tilde{g}_k}{1+(hk)^4} \\
\Rightarrow\ E(\hat{\tilde{g}}_k) - \tilde{g}_k &= \frac{-(hk)^4}{1+(hk)^4}\tilde{g}_k \\
\Rightarrow\ \text{Bias}\left[\hat{g}(t)\right] &\approx h^4 g^{(iv)}(t) \\
\text{Var}\left[\hat{\tilde{g}}_k\right] &= \frac{\sigma^2/n}{\left(1+(hk)^4\right)^2} \\
\hat{g}(t) &= \sum \hat{\tilde{g}}_k \exp(-j2\pi kt) \\
\Rightarrow\ \text{Var}\left[\hat{g}(t)\right] &\approx \sum \text{Var}\left[\hat{\tilde{g}}_k\right] \\
&\approx \frac{\sigma^2}{n}\sum \frac{1}{\left(1+(hk)^4\right)^2} \\
&\approx \frac{\sigma^2}{nh}\int \frac{1}{\left(1+\omega^4\right)^2} d\omega
\end{aligned}$$

So, assuming $h \to 0$ and $nh \to \infty$, the pointwise asymptotic bias is of $O(h^4)$ and the variance of $O\left(1/(nh)\right)$. The asymptotic MSE is minimised when $h \approx n^{-4/9}$, giving an

optimal convergence rate of $O\left(n^{-8/9}\right)$.

Splines that do not satisfy the boundary conditions have a boundary bias problem (Rice and Rosenblatt, 1983; Utreras, 1988). For example, a cubic spline satisfing $g''(0) = g''(1) = 0$ but not $g^{(3)}(0) = g^{(3)}(1) = 0$ has a bias of $O\left(h^{3/4}\right)$ near the boundaries and if neither of these conditions is satisfied, a bias of $O\left(h^{1/2}\right)$. Smoothing splines have the advantage of uniqueness. That is, there is only one function $\hat{g}_h(t)$ which minimises (2.7). However, splines can be severely affected by outliers. Further properties of the smoothing spline can be found in de Boor (1978); Eubank (1988); Wahba (1990).

2.1.5 Comparison of Methods

In practice, choosing the nonparametric smoothing method to use will partially depend on the context of the problem. No method is "best" in all situations. Instead, the best method for a particular dataset will depend on many factors including the complexity of the relationship and the sample size (eg. Banks *et al.*, 1995; Marron, 1996). We outline below some of the strengths and weaknesses of the methods in different situations.

The basis method is relatively simple to use and understand. It is particularly good when the relationship is of a known, relatively smooth form (eg. periodic). In other situations, using B-spline basis functions allows for greater flexibility in the model, through piecewise polynomials, while maintaining the simplicity of the calculations and understanding. However, a large number of basis functions (m) may be needed in order for the estimator to exhibit very local behaviour. A large m will reduce the bias but at the expense of a large variance. Thus, basis estimators are not recommended for estimating functions with a large amount of local behaviour.

Kernel estimators can model local behaviour with overparameterisation. They are still easy to understand since they simply fit a local constant to the data around each point at which the estimator is required. Weighted least squares can be used to find each estimate. Kernel

estimators suffer from a boundary bias problem as the kernel window at boundary points is missing data. This problem will occur in any area with only a small amount of data. Hence, kernel estimators are not recommended for sparse datasets.

Local polynomial estimators retain the advantages of kernel estimators but without the boundary bias problem. Sparse data problems can be overcome to some extent by fitting nearest neighbour weights rather than kernel weights. The resulting estimates are even more flexible than kernel estimates as they fit a local polynomial at each point instead of a local constant. They are particularly superior when the t_i's are random as they adapt without increasing the bias or variance. However, local polynomial estimators are sensitive to outliers; as are kernel estimators. This problem can be overcome though by using a robust local polynomial method such as loess.

Another approach which corrects the boundary bias problem of kernels while still allowing the estimator to have differing degrees of smoothness is the natural smoothing spline. Instead of fitting local polynomials of low degree possessing all derivatives, the smoothing spline fits piecewise polynomials with discontinuities at the knots in lower order derivatives. However, if the natural boundary conditions are not satisfied, then a smoothing spline will still have a boundary bias problem. Smoothing spline estimators also differ in optimising a *penalised* likelihood; although a penalty term can also be easily incorporated into the basis method. They are also sensitive to outliers since they are a least squares method. Like the local polynomial method, smoothing spline estimators behave well for random t_i's. However, the problem of choosing the number and position of the knots is a difficult one.

In terms of asymptotic pointwise MSE, the kernel, local linear polynomial, and Fourier basis methods all have an optimal convergence rate of $O\left(n^{-4/5}\right)$. Local quadratic or cubic polynomial estimators and natural smoothing splines have a slightly better convergence rate of $O\left(n^{-8/9}\right)$. Since the methods have similar asymptotic rates for equispaced t_i's (which are of interest), this is not a large consideration in determining the "best" method.

2.2 Controlling Smoothness

With any of the nonparametric methods discussed, the smoothness of the estimated function needs to be controlled. Smoothing is actually a trade-off between the goodness-of-fit and the "roughness" of the estimated function. The amount of smoothing is controlled by the window width or smoothing parameter, h, and/or $1/m$ (as seen by comparing the asymptotic MSE formulae) for basis functions. In the following we will refer to any of these as the smoothing parameter h.

Choosing the smoothing parameter is an extremely important part of any nonparametric method. Oversmooth and local features will be missed (not a good fit to the data). Undersmooth and the function will closely follow the data, resulting in an extremely "bumpy" estimate. Thus, there have been extensive investigations into automated techniques for choosing the smoothing parameter. These include Cross-Validation (CV) (Stone, 1974; Green and Silverman, 1994), Akaike's Information Criterion (AIC) (Akaike, 1970, 1973; Bozdogan, 1987), and Bayesian Information Criterion (BIC) (Akaike, 1978; Schwarz, 1978). The most commonly used method is cross-validation, which we discuss below.

2.2.1 Cross-Validation

Cross-validation is a data driven method for choosing the smoothing parameter. It is originally due to Allen (1974) and Stone (1974), and was motivated by the idea of prediction. Suppose we take a new observation y_{new} at the point t_{new}. Then a good estimator, $\hat{g}(t)$, would give a small error value. However, generally in practice no new data is available. This is where the "leave one out" idea of cross-validation comes in. The idea is to omit each subject in the data set in turn, generate estimates using the remaining data, and then minimise the error over all the omitted ("new") data to find the optimal smoothing parameter value.

In practice, the error criterion should relate to the problem being studied. Such an analysis is outside the scope of this thesis and we follow standard practice and deal with mean integrated

squared error (MISE). Also we assume deterministic t_i thoughout.

$$\begin{aligned}\text{MISE} &= E\int (g(t) - \hat{g}_h(t))^2\, dt \\ &= E\|g(t) - \hat{g}_h(t)\|^2\end{aligned}$$

where $\hat{g}_h(t)$ is the estimated regression function found using any of the methods from Section 2.1 with a smoothing parameter value of h. Since $g(t)$ is unknown, we estimate the MISE using cross-validation. Let $\hat{g}_h^{(-i)}(t)$ be the estimated function when the i-th subject is omitted from the data. Then the cross-validation function is

$$\text{CV}(h) = n^{-1}\sum_1^n \left(y_i - \hat{g}_h^{(-i)}(t_i)\right)^2 \tag{2.8}$$

Since $E(\text{CV}(h)) \approx \text{MISE}$ (eg. Stone, 1974), the optimal smoothing parameter is then the value of h which minimises $\text{CV}(h)$. A grid search is usually used to find h as $\text{CV}(h)$ may not have a unique minimum. Wahba and Wold (1975) were the first to use cross-validation to choose h for smoothing splines, while Clark (1975) used it in the context of kernel smoothing. Extensive work has also be carried out in the field of density estimation (eg. Rudemo, 1982; Bowman, 1984).

Direct calculation of $\text{CV}(h)$ for a number of h values can be time consuming as it requires the estimation of n curves $\hat{g}_h^{(-i)}(t_i)$, for each value of h. The calculation of the CV function can be sped up using the hat matrix, $A(h)$. Since the nonparametric estimators discussed above are linear in t_i there will exist a matrix $A(h)$ such that

$$\begin{bmatrix}\hat{g}_h^{(-i)}(t_1)\\ \vdots \\ \hat{g}_h^{(-i)}(t_n)\end{bmatrix} = A(h)\begin{bmatrix}y_1\\ \vdots \\ y_n\end{bmatrix}$$

The CV function can then be calculated in $O(n)$ computations using the estimated regression function for the full data set

$$\text{CV}(h) = n^{-1}\sum_1^n \frac{(y_i - \hat{g}_h(t_i))^2}{(1 - a_{ii}(h))^2} \tag{2.9}$$

For further details on the calculation of $A(h)$ for different nonparametric methods see Eubank (1988), for example.

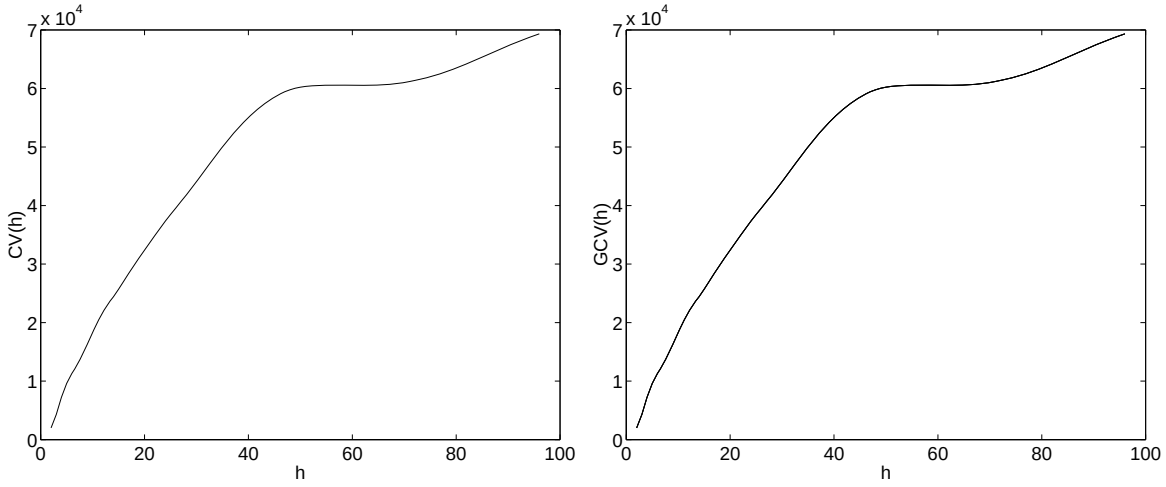


Figure 2.8: *Cross-validation (left) and generalised cross-validation (right) plots for a kernel estimator of the birth rate data.*

A related criterion is Generalised Cross-Validation (GCV). It was first proposed by Craven and Wahba (1979) and is essentially a weighted version of CV(h). Instead of dividing the residuals by $1 - a_{ii}(h)$ as in CV, GCV divides by the average of these values, $1 - n^{-1} \operatorname{tr} A(h)$.

$$\mathrm{GCV}(h) = n^{-1} \frac{\sum_1^n \left(y_i - \hat{g}_h^{(-i)}(t_i) \right)^2}{\left(1 - n^{-1} \operatorname{tr} A(h)\right)^2} \tag{2.10}$$

GCV has the advantage of being computationally quicker as $\operatorname{tr} A(h)$ can be computed without finding the $a_{ii}(h)$'s. It is also less sensitive to extreme values. As noted in Green and Silverman (1994), when $t_i = i/n$ and periodic boundary conditions are satisfied, then $a_{ii}(h)$ are equal for all i and GCV is identical to CV. This can be seen in the birth rate example (Figure 2.8). Since the t_i's are equispaced, the CV and GCV plots for a kernel estimator are identical. However, the $CV(h)$ values took longer to compute. If the t_i are random, then GCV outperforms CV (Kohn *et al.*, 1991).

In practice, smoothing parameters chosen using CV generally tend to result in an under-smoothed estimator (Chiu, 1990). That is, the curves are too "bumpy". CV is intended to be an unbiased estimator of the MISE, resulting in an asymptotically optimal choice of h, but it has high variability. Also, the estimated smoothing parameter $\hat{h}$ from CV converges to

the optimal value $\hat{h}_{opt}$, which minimises the MISE for the given data set, at a slow rate. For example, Härdle *et al.* (1988) showed this rate to be of $O\left(n^{-1/10}\right)$ for kernel estimators. That is, $n^{1/10}(\hat{h} - \hat{h}_{opt})/\hat{h}_{opt}$ tends to an asymptotic normal distribution. However, the estimated function found using $\hat{h}$, $\hat{g}_{\hat{h}}(t)$, is asymptotically consistent in that as $n \to \infty$

$$L(\hat{h}) = n^{-1}\|g(t) - \hat{g}_{\hat{h}(t)}\|^2 \to 0$$

in probability, for all t, where $g(t)$ is the true function. Li (1984) proved this for the case of local polynomials with nearest neighbour weights. Thus, cross-validation is asymptotically optimal under regularity conditions in that (Li, 1987)

$$\frac{L(\hat{h})}{\inf_h L(h)} \to 1$$

in probability.

2.3 Methods for Correlated Data

The nonparametric smoothing methods as decribed in Section 2.1 relied on the assumption of independent errors. However, in many datasets this assumption is false. Of particular interest is data sampled over time, such as the US birth rate data. Obviously, the data at adjacent time points are not independent. This autocorrelation can effect the asymptotic properties of the estimator and the behaviour of data-based smoothing parameter selectors, such as cross-validation. Thus, some modification of the techniques are needed for correlated data. Since nonparametric smoothing methods for correlated data are beyond the scope of this thesis, we present a short outline of the two main approaches only. Further information can be found in (Härdle, 1990, Chapter 7) and (Simonoff, 1996, Section 5.5), for example.

The first method is to assume that the errors are independent and use one of the standard smoothing techniques to estimate the function, $g(t)$. In this case, cross-validation (Section 2.2) tends to oversmooth data with negatively autocorrelated errors and undersmooth data with positively autocorrelated errors (eg. Diggle and Hutchinson, 1989). Thus, the smoothing parameter needs to be adjusted due to the correlation. For example, Chiu (1989); Alt-

man (1990), and Hart (1991) all looked at modifications of cross-validation (and generalised cross-validation) for autocorrelated errors with kernel smoothing. These methods require the correlation function to be estimated. The standard approach assumes $\text{Cov}(\varepsilon_i, \varepsilon_j) = \gamma(|i-j|)$, for some known function $\gamma(u)$.

The second method for modelling correlated data is to modify the estimator to allow for correlated noise. For example, Kohn *et al.* (1992) developed a state space solution for a spline estimator with stationary noise modelled by an ARMA (Box and Jenkins, 1976) model. This approach also requires a modification of the cross-validation procedure using an estimated correlation function. Thus, this approach is more complex than the previous method.

2.4 Towards Functional Data Analysis

The previous smoothing techniques were concerned with the estimation of a single function or curve. We now move towards functional data by briefly examining techniques for the analysis of multiple curves. We now have, for each of the n subjects, measurements of the dependent variable y taken repeatedly, usually through time. Thus, $y_i(t_j)$ is the measurement of the dependent variable taken at time t_j for subject i. We look at the simplest case: equispaced time points with measurements taken at the same times for each subject. That is, $t_j = j/p$, where p is the number of measurements on each subject.

2.4.1 Classical Longitudinal Data Analysis

Classical longitudinal data analysis assumes that the individual subjects are independent while the measurements for each subject are correlated. This correlation is assumed to be consistent for all subjects. However, the number of measurements per subject must be much smaller than the number of subjects: that is, $p \ll n$. Numerous books and papers have been written on longitudinal data analysis; see, for example, Goldstein (1979); Cox and Oakes (1984); Jones (1993); Lindsey (1993); Diggle *et al.* (1994); Vonesh and Chinchilli (1997) and references therein. We briefly outline the various approaches below.

All multivariate statistical techniques have been adapted to longitudinal data. These adaptations take into account the correlation within subjects. However, in order to model the correlation, assumptions about its structure are required; eg. stationarity.

Three of the main approaches to longitudinal data analysis are marginal models, random effects models, and transition models. Marginal models (Liang *et al.*, 1992) separately model the mean (plus predictor effects) and covariance. For example, in a regression model, the within-subject correlation is modelled separately from the mean and the effect of the independent variables. A marginal model assumes that the variance at time t_j depends on the mean at t_j through a known variance function, and the correlation between measurements at times t_j and t_k depend on the mean at these times through another known function.

Random effects models (Laird and Ware, 1982) allow the regression coefficients to vary from one subject to another, according to a known distribution. This allows for heterogeneity due to unmeasured factors. They model the conditional expectation of the response, $y_i(t_j)$, given the subject specific coefficients, β_i, as $E(y_i(t_j)|\beta_i) = x_i^T(t_j)\beta_i$, where $x_i(t_j)$ is the vector of independent variables for subject i at time t_j, and the $\beta_i = \beta + U_i$ follow a known distribution with mean β and U_i is a mean zero random vector. Random effects models are particularly useful when we are interested in inference about individual subjects rather than population averages, which are of interest with marginal models.

Transition models (eg. Muenz and Rubinstein, 1985) more closely resemble time series models. They assume the present observation for a subject depends on the past observations, as well as any predictor variables. A special case of these types of models are Markov chains (eg. Billingsley, 1961; Feller, 1968). Time series models, such as autoregressive processes, which have been adapted to longitudinal data (eg. Kenward, 1987) also fall into this category.

For any of these model, methods for estimating the parameters include weighted least squares, restricted maximum likelihood (REML) (Patterson and Thompson, 1971), generalised estimating equations (GEE) (Zeger *et al.*, 1988), best linear unbiased predictor (BLUP) (Robin-

son, 1991), and Gibbs sampling (Zeger and Karim, 1991).

2.4.2 Functional Data Analysis based on Stationarity

When the number of measurements per subject is greater than the number of subjects ($p > n$) then classical longitudinal data analysis techniques are generally no longer suitable. Using these methods would result in singularity problems. Time series analysis can be used to overcome this problem by relying on stationarity. In fact, longitudinal data analysis via time series analysis can be thought of as *stationary* functional data analysis.

Most of the time series techniques have been developed in the frequency domain, wherein the data is modelled as a sum of periodic sine and cosine waves of different periods or frequencies. This approach relies on the fact that the ordinates of Fourier frequencies of finite Fourier transform of the data are nearly independent and approximately normally distributed with variance equal to the spectrum at that frequency.

Techniques developed for the analysis of **multiple** time series include ANOVA, regression analysis, and principal component analysis. Further information on these techniques can be found in Shumway (1988) and Brillinger (1981b), and references therein.

Time series ANOVA (Shumway, 1970; Brillinger, 1973) can be performed on ordinary stationary time series or on components of a stationary point process. It can be thought of as a special kind of functional data analysis but with a stationarity assumption. Both fixed and random effects models have been developed. For example,

$$y_{ij}(t) = \mu_{ij} + \gamma(t) + \beta_j(t) + \varepsilon_{ij}(t)$$

where i is the subject index, μ_{ij} are constant, $\gamma(t)$ is constant in the fixed effect model or a stationary series in the random effects models, and $\beta_j(t)$ and $\varepsilon_{ij}(t)$ are independent realisations of stationary series. Parameter estimates are found by a complex version of standard ANOVA carried out on the discrete Fourier transforms of the observed series $y_{ij}(t)$.

Regression analysis is also carried out in the frequency domain. Let $x_{ij}(t)$ be the jth regressor time series for the ith subject. Assuming $x_{ij}(t)$ is a stationary, zero-mean normal process, then

$$y_i(t) = \alpha(t) + \sum_{j=1}^{q} \sum_{s=-\infty}^{\infty} \beta_j(s) x_{ij}(t-s) + \varepsilon_i(t)$$

The Fourier transform of the parameters can be found via least squares type calculations on the cross-spectrum vector between the regressors and response and the cross-spectral matrix of the regressor series; both of which can be easily found from the Fourier transforms of the covariances and autocovariances of the observed time series (eg. Shumway, 1970). The frequency domain approach has the advantage of only requiring a $q \times q$ matrix to be inverted. Also, the regressor series are allowed to be cross-correlated.

Principal component analysis for stationary time series (Hannan, 1961; Brillinger, 1969, for example) is concerned with approximating a set of time series by a filtered version of itself. However, the linear filter is constrained to be of a rank less than the set of series. Direct calculation of the covariance matrix would result in singularity problems, since $n \ll p$. The stationarity assumption allows the covariance matrix to be estimated using the finite Fourier transforms of all the series in the set. Hence, time series principal components can be found using a complex version of multivariate principal component analysis.

2.5 Functional Data Analysis

So far we have discussed function estimation for a single curve and techniques for multiple curves when $n > p$ or the curves are stationary. We now turn our attention to functional data analysis techniques. Classical longitudinal data analysis techniques (Section 2.4.1) are inappropriate for functional data since $n \ll p$. The time series methods described in Section 2.4.2 are not preferred as the stationarity assumption does not necessarily hold. Thus, new techniques have been developed for the analysis of functional data.

In this section we describe existing methods for functional data in two areas; regression and

principal component analysis. Whilst other areas are also important, these methods form the background to the techniques developed in this thesis.

2.5.1 Functional ANOVA and Regression

Linear regression and ANOVA are common statistical techniques for exploring the relationship between variables. In the functional setting, either the response variable, regressor (predictor) variable or both can be functional.

Functional Response

When only the response is functional, modifications to standard ANOVA have been developed. Ramsay *et al.* (1996) proposed a "naive" functional ANOVA (FANOVA) model to analyse lip motions produced from the utterance of different syllables (treatments). They did this by extending multivariate ANOVA (MANOVA) by simply replacing the discrete variable index in MANOVA with a continuous time variable and using standard ANOVA techniques to find the estimates at each t. Thus, their model was

$$y_{ij}(t) = \mu(t) + \alpha_j(t) + \varepsilon_{ij}(t)$$

subject to

$$\sum_j \alpha_j(t) = 0 \qquad \text{for all } t$$

and estimates at t were found by minimising the least squares criterion (or alternatively a penalised criterion to account for the constraints). Continuous estimates were then generated by interpolating between the fitted t values. FANOVA was also presented in Ramsay and Silverman (1997) and used to model temperature data by climate zones and the effect of shoeing conditions on horses. While the "naive" approach is relatively simple, it does not take into account the continuous nature of the parameters. That is, the fact that the parameters are correlated across the time frame of the data is ignored.

Faraway (1997) proposed a functional ANOVA which used nonparametric smoothing. The

first step was to generate estimates of smooth curves for each subject. Any of the methods from Section 2.1 could be used but loess was recommended because of its robustness properties. Once curve estimates were found at specified values of t, pointwise estimates of the parameters were found using traditional ANOVA techniques. While Faraway presented some statistical inference for the model, it is only tentative and further theoretical investigation of its properties is needed. This model is better than the naive approach but it still requires estimates to be generated separately at each time point t. Fan and Lin (1998) used Fourier transforms of the functional response to develop a high-dimensional ANOVA (HANOVA) to determine if there was a significant difference in the curves between groups. However, they assumed that the errors were stationary time series with zero means.

A fully nonparametric functional ANOVA was presented by Ramsay and Silverman (1997). They modelled both the response and the parameter functions via basis expansions. Let Z be a design matrix for the model and $\beta(t) = [\mu(t),\ \alpha_1(t),\ \alpha_2(t), \ldots]^T$ be a vector of the parameters to be estimated, then the model can be written as $Y(t) = Z\beta(t) + \varepsilon(t)$, where $Y(t) = [y_1(t)\ \ldots\ y_n(t)]^T$, $\varepsilon(t) = [\varepsilon_1(t)\ \ldots\ \varepsilon_n(t)]^T$. The estimates were then found by minimising the least squares criterion $\int \|Y(t) - Z\beta(t)\|^2 dt$, subject to any constraints, and substituting basis expansions for $Y(t)$ and $\beta(t)$. Let $L\beta(t) = 0$, for all t, represent the constraints, and assume ρ is given. Then, this leads to solving

$$\left(Z^T Z + \rho L^T L\right) B = Z^T C$$

where

$$\begin{aligned} y_i(t) &= \sum c_{ik}\psi_k(t) = c_i^T \psi(t) \\ \Rightarrow\ Y(t) &= C\psi(t) \\ \beta(t) &= B\psi(t) \\ \psi(t) &= [\psi_1(t)\ \ldots\ \psi_m(t)]^T \\ C &= [c_1\ \ldots\ c_n]^T \end{aligned}$$

Ramsay and Silverman recommended using a value of $\rho = 1$. A further model was also outlined which controlled the smoothness of the estimated $\hat{\beta}(t)$ explicitly through the addition

of a roughness penalty term to the least squares criterion. One of the advantages of Ramsay and Silverman model is that the correlation across t is taken into account for both the response and the parameters. Also a single step is needed to generate the estimates, rather than separate smoothing and estimation steps.

Functional Regressor

When data containing a functional regressor (or covariate) is to be modelled, functional linear modelling techniques have been developed. For these models, either just a predictor or both the response and a predictor can be functional. Of most interest is the case of a continuous (not functional) response variable y and a functional predictor variable x. In this case, the functional linear model is given by

$$y_i = \alpha + \int_J x_i(t)\beta(t)\,dt + \varepsilon_i, \qquad i = 1, \ldots n$$

where α is a constant parameter and $\beta(t)$ is the functional parameter; both of which need to be estimated. A solution analgous to ridge regression was found by Goutis (1998) by minimising the penalised sum of squares. Ramsay and Silverman (1997) presented a basis solution to this problem. Both the functional predictor and parameter were modelled as linear combinations of the basis functions, $\psi_k(t)$.

$$x_i(t) = \sum_{k=1}^{m} c_{ik}\psi_k(t) = c_i^T\psi(t)$$
$$\beta(t) = \sum_{k=1}^{m} b_k\psi_k(t) = b^T\psi(t)$$

Substituting these basis expansions into the linear model gives

$$\begin{aligned} y_i &= \alpha + \int_J c_i^T\psi(t)\psi^T(t)b\,dt + \varepsilon_i \\ &= \alpha + c_i^T W b + \varepsilon_i \end{aligned}$$

where $W = \int_J \psi(t)\psi^T(t)dt$. If a Fourier basis is used, then $W = I_m$, the order m identity matrix. Otherwise, W must be approximated using a method of quadrature or calculated analytically. The parameter estimates $\hat{\alpha}$, $\hat{b}$ can then be found using standard linear regression techniques. Ramsay and Silverman also modified this model to include a roughness penalty term; as well as giving computational details on generating the estimates.

Functional Response and Regressor

Functional linear models with both a functional response and a predictor have been presented in Ramsay and Dalzell (1991). They investigated the relationship between precipitation and temperature values over a year at different weather stations. They did this by finding smoothing spline representations of the data for each station, and then relating the variables via harmonic weight functions. A basis solution to the same problem was given in Ramsay and Silverman (1997). This method is simply a combination of the two basis methods outlined above. Ramsay and Silverman also outlined a smoothing spline solution for functional linear modelling.

The functional linear models are more general than time series methods for regression. In time series models the function linking the regressors and response must be time invariant. In functional models, this does not have to hold. Thus, models of the form

$$y_i(t) = \alpha + \int \beta(t,s)x_i(s)ds$$

are possible in the functional setting.

2.5.2 Functional Principal Component Analysis

Principal component analysis of time series data with $n \ll p$ first received attention in the time series domain in the 1970s; see Shumway (1970); Brillinger (1973, 1980). The curve for each subject was modelled as

$$y_i(t) = \mu(t) + \varepsilon_i(t)$$

where $\mu(t)$ represents the mean at time t and $\varepsilon_i(t)$ the residual variation. The idea was that removing the mean from y_i would result in the modelling of a stationary process for the residual variation. However, the assumption of a stationary residual covariance structure may not hold for functional data. Several papers have considered principal component analysis applied to functional data without the stationarity assumptions.

However, before functional Principal Component Analysis (fPCA) can be performed, the

data needs to be mean adjusted. In most cases, this has been done by simple averaging; finding the average value at each discrete time point. Rice and Silverman (1991) extended this by smoothing the pointwise averages using penalized least squares to produce a cubic spline estimate of $\mu(t)$. Other nonparametric methods could also be used; for example, Hart and Wehrly (1986) use a kernel regression estimator. Having mean adjusted the data, the eigenfunctions of the residual variation can be found. We assume in the following that $y_i(t)$ has been mean adjusted, so $Ey_i(t) = 0$ for all t.

To develop a principal component analysis we expand the population covariance function in a Karhunen-Loeve expansion (Kanwal, 1971)

$$\mathrm{Cov}(Y(s), Y(t)) = \Gamma(s,t) = \sum_{1}^{\infty} \lambda_u \phi_u(s) \phi_u(t)$$

Each eigenfunction, $\phi_u(t)$, and its corresponding eigenvalue, λ_u, can be found by solving the eigenproblem

$$\int \Gamma(s,t)\phi_u(t)dt = \lambda_u \phi_u(s), \quad u = 1, 2, \ldots \tag{2.11}$$

subject to $\|\phi_u\|^2 = 1$ and $\langle \phi_u, \phi_v \rangle = 0, u \neq v$. A number of methods have been developed to find these eigenfunctions from finite data estimates.

Castro *et al.* (1986) looked at extending traditional multivariate PCA to smooth curves. They took into account the spacing of the time points to develop an algorithm capable of handling both equispaced and random t_i's. However, they calculate the sample covariance function, Γ_n, directly from the data using the standard estimator.

$$\hat{\Gamma}_n(s,t) = \frac{1}{n} \sum_{i=1}^{n} y_i(s) y_i(t) \tag{2.12}$$

This estimator results in singularity problems if used with functional data since $n < p$.

The most commonly used approach to overcome this singularity problem is a basis approach. The basic idea is to model the curve for each subject using known basis functions, $\psi_k(t)$,

$k = 1, \ldots, m$, as in functional regression. Thus,

$$y_i(t) = \sum c_{ik}\psi_k(t) = c_i^T\psi(t)$$
$$\Rightarrow\ Y(t) = C\psi(t)$$

However, the eigenfunctions, ϕ_u, are also modelled using these basis functions.

$$\phi_u(t) = \sum f_u\psi_k(t)$$
$$= \psi^T(t)f_u$$

Using the sample covariance function (2.12) as the estimate for the population covariance function along with the basis expansions, the eigenproblem (2.11) becomes

$$n^{-1}\int \psi^T(s)C^TC\psi(t)\psi^T(t)f_u\,dt = \lambda_u\psi^T(s)f_u$$

Since this equation must hold for all s we deduce

$$n^{-1}C^TCWf_u = \lambda_u f_u$$

Further, the constraint $\|\phi_u\|^2 = 1$ implies that $f_u^TWf_u = 1$

Ramsay (1982); Besse and Ramsay (1986); Ramsay and Dalzell (1991) used a basis approach. They calculated smooth principal components by extending traditional multivariate PCA using Hilbert spaces. The resulting equations were solved using a basis expansion with reproducing kernels, to produce a weighted PCA. Ramsay *et al.* (1994, 1996) used these techniques to analysis human growth data and lip motion data, respectively.

Rice and Silverman (1991); Ramsay and Silverman (1997) also used a basis approach to find the eigenfunctions. However, they imposed smoothing via the addition of a roughness penalty to the basis expansion. This penalty was the integrated squared second derivative of the eigenfunction, $P(\phi) = \int D^2\phi D^2\phi$, where D is d/dt, although other penalties could be used. Assuming $D^2\phi$ and $D^3\phi$ satisfy either natural or periodic boundary conditions, the modified eigenequation became

$$WC^TCWf_u = \lambda_u(W + \alpha P(\psi))f_u$$

subject to

$$f_u^T W f_u = 1 \qquad \text{and}$$

$$f_u^T W f_v + \alpha f_u^T P(\psi) f_v = 0, \quad u \neq v$$

This method has the advantage that the smoothing can be controlled more precisely and the basis can be truncated, for computational purposes, without substantially altering the results. Pezzulli and Silverman (1993) looked at some of the theoretical properties of Rice and Silverman's procedure. They showed that the estimates were consistent providing the smoothing parameter converged to zero as $n \to \infty$, and that smoothing improved the estimated eigenfunctions compared to the eigenfunctions calculated using the raw empirical covariance kernel.

A different form of smoothing was used by Silverman (1996). The method is similar to that of Rice and Silverman (1991) but Silverman incorporated the roughness penalty into the orthonormality constraint instead of incorporating it into the eigenanalysis directly.

Kneip (1994) estimated the smooth eigenfunctions of the data via cubic B-splines. However, he assumed that $\text{var}(\varepsilon_i(t)) = 1$, or could be transformed to equal one using the estimated variances. The eigenfunctions were then used to give a low dimensional model of the curve for each subject. The technique was applied to the U.K. family expenditure survey data from 1968 to 1983.

Another approach to fPCA was presented by Solo (1997). Here, the estimates were found exactly using a roughness penalty but without the use of a basis expansion or discretisation. The eigenfunctions were solved using 'continuous-discrete' variational calculus.

In most cases, estimating and subtracting the mean function from the data before fPCA is performed is inadequate. The mean and covariance functions need to be estimated together, as one estimate can affect the other. This leads to a functional mixed-effects type model; incorporating fPCA in the covariance estimate as developed in chapter 6.

Chapter 3

Functional Logistic Regression

3.1 Introduction

In recent years, techniques have been presented for the linear modelling of data where one covariate is a function rather than a single measurement. However, as outlined in Section 1.1, existing regression techniques deal with either a continuous or functional response. Presented in this chapter is a method for FDA with a binary response, which we call *functional logistic regression.*

We develop a basis solution to functional logistic regression by extending existing functional linear models, using standard generalized linear modelling (glm) techniques (McCullagh and Nelder, 1989), to the case of a binary response variable. This results in a functional logistic regression with maximum likelihood parameter estimates. The methods presented in Chapters 3 - 4 and the results from Chapter 5 also appear in Ratcliffe *et al.* (2000a) and Ratcliffe *et al.* (2000b).

3.2 Basis Method

For each of the n subjects, the observed response, y_i, is assumed to come from a Bernoulli distribution with probability of success π_i. As with standard glm, the data are modelled

using the logit or logistic link function, giving the functional logistic regression model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \sum_{j=1}^{r} \alpha_j z_{i,j} + \int x_i(t)\beta(t)\,dt, \qquad i = 1, \ldots, n \tag{3.1}$$

where

$$\begin{aligned}
\alpha_0 &= \text{ the constant parameter,}\\
\alpha_j &= \text{ the parameter for the } j^{th} \text{ scalar covariate,}\\
z_{i,j} &= \text{ the } j^{th} \text{ scalar covariate for the } i^{th} \text{ subject,}\\
x_i(t) &= \text{ the functional covariate for the } i^{th} \text{ subject at time } t,\\
\beta(t) &= \text{ the functional parameter at time } t,\\
i &= 1, \ldots, n \qquad j = 1, \ldots, r
\end{aligned}$$

As mentioned earlier, solving the above problem directly requires the inversion of a singular matrix. Instead we model the functional covariate and parameter nonparametrically by basis functions, $\psi_k(t)$, $t = t_1, \ldots, t_p$. Clearly, the basis functions should be chosen to reflect the characteristics of the data. An inappropriate choice of basis functions would require a larger number of functions (m) to adequately model the data and parameter. For example, if the data for each subject came from a straight line, say $x_i(t) = a_{i0} + a_{i1}t$, then the polynomial basis functions $\psi_k(t) = t^{k-1}$, $k = 1, \ldots, m$, would model $x_i(t)$ perfectly with only $m = 2$ functions. However, Fourier basis functions, $\psi_1(t) = 1$, $\psi_{2k}(t) = \cos 2\pi kt$, $\psi_{2k+1}(t) = \sin 2\pi kt$, $k = 1, \ldots, m/2$, would require a large m in order to adequately model $x_i(t)$.

Having chosen the basis functions, the data and functional parameter are modelled as

$$x_i(t) = \sum_{k=1}^{m} c_{ik}\psi_k(t) = c_i^T\psi(t) \tag{3.2}$$

$$\beta(t) = \sum_{k=1}^{m} b_k\psi_k(t) = b^T\psi(t) \tag{3.3}$$

where b and c_i are vectors of basis coefficients which need to be estimated. Using

$$\begin{aligned} x(t) &= [x_1(t) \ \dots \ x_n(t)]^T \\ X_{(n\times p)} &= [x(t_1) \ \dots \ x(t_p)] \\ C_{(n\times m)} &= [c_1 \ \dots \ c_n]^T \\ \Psi_{(m\times p)} &= [\psi(t_1) \ \dots \ \psi(t_p)] \end{aligned}$$

estimates for C can be found easily via least squares.

$$\begin{aligned} x_i(t) &= c_i^T \psi(t) \\ \Rightarrow \ X &= C\Psi \\ \Rightarrow \ \hat{C} &= X\Psi^T \left(\Psi\Psi^T\right)^{-1} \end{aligned}$$

Substituting the basis expansions (3.2), (3.3), and using $z_i = [1 \ z_{i1} \ \dots \ z_{ir}]^T$, $\alpha = [\alpha_0 \ \alpha_1 \ \dots \ \alpha_r]^T$, the regression model (3.1) becomes

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= z_i^T\alpha + \int c_i^T\psi(t)\psi^T(t)b \ dt \\ &= z_i^T\alpha + c_i^T W b \\ \Rightarrow \ \log\left(\frac{\pi}{1-\pi}\right) &= Z\alpha + CWb \end{aligned} \tag{3.4}$$

where $W = \int \psi(s)\psi^T(s)\,ds$, $\pi = [\pi_1 \ \dots \ \pi_n]^T$, $Z = [z_1 \ \dots \ z_n]^T$, and $\log(u) = [\log(u_1) \ \dots \ \log(u_n)]^T$.

3.2.1 Estimating Parameters

Model (3.4) can be written as

$$\log\left(\frac{\pi}{1-\pi}\right) = [Z \ CW] \begin{bmatrix} \alpha \\ \\ b \end{bmatrix}$$

which is similar to standard logistic regression models. Hence, maximum likelihood parameter estimates can be found using the Fisher scoring method (Fisher, 1925).

Since the observed responses, y_i, are assumed to be Bernoulli(π_i), the log-likelihood function can be expressed as

$$l(\pi; y) = \sum_{i=1}^{n} \left[y_i \ \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right] \tag{3.5}$$

Using

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i - \pi_i}{\pi_i(1-\pi_i)}$$
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = z_i^T\alpha + c_i^T W b$$

the derivative of the log-likelihood with respect to α_j is

$$\begin{aligned}\frac{\partial l}{\partial \alpha_j} &= \sum_{i=1}^{n} \frac{\partial l}{\partial \pi_i}\frac{d\pi_i}{d\eta_i}\frac{\partial \eta_i}{\partial \alpha_j} \\ &= \sum_{i=1}^{n}(y_i - \pi_i)z_{i,j}^T\end{aligned} \tag{3.6}$$

and the Fisher information for α is given by

$$\begin{aligned}-E\left(\frac{\partial^2 l}{\partial\alpha_j\partial\alpha_k}\right) &= \sum_{i=1}^{n}\pi_i(1-\pi_i)z_{i,j}^T z_{i,k}^T \\ &= \{Z^T w^* Z\}_{jk}\end{aligned}$$

where w^* is the diagonal weights matrix with $\pi_i(1-\pi_i)$ as its diagonal elements. Thus, by Fisher's scoring method, the new estimate of α, $\hat{\alpha}^{new}$, is given by

$$\begin{aligned}\hat{\alpha}^{new} &= \alpha - \left(E\left(\frac{\partial^2 l}{\partial\alpha\partial\alpha^T}\right)\right)^{-1}\left(\frac{\partial l}{\partial\alpha}\right) \\ &= \alpha + \left(Z^T w^* Z\right)^{-1}\left(Z^T(y-\hat{\pi})\right) \\ &= \left(Z^T w^* Z\right)^{-1}\left(Z^T w^* Z\alpha + Z^T(y-\hat{\pi})\right) \\ &= \left(Z^T w^* Z\right)^{-1} Z^T w^*(\nu - CW\hat{b})\end{aligned} \tag{3.7}$$

where $\nu_i = \hat{\eta}_i + (y_i - \hat{\pi}_i)/w_{ii}^*$ is analogous to the local dependent variable z_i in standard glm. Similarly,

$$\hat{b} = \left(W^T C^T w^* CW\right)^{-1} W^T C^T w^*(\nu - Z\hat{\alpha}) \tag{3.8}$$

Combining equations (3.7), (3.8) gives

$$\begin{bmatrix}\hat{\alpha}\\ \hat{b}\end{bmatrix} = \begin{bmatrix} Z^T w^* Z & Z^T w^* CW \\ (CW)^T w^* Z & (CW)^T w^* CW\end{bmatrix}^{-1}\begin{bmatrix} Z^T \\ (CW)^T\end{bmatrix} w^*\nu \tag{3.9}$$

which corresponds to a weighted least squares estimate.

Since ν depends on the parameter estimates and these estimates in turn depend on ν, an iterative algorithm (as given in McCullagh and Nelder, 1989) is used to estimate the parameters in practice.

STEP 1: Given initial estimates $\hat{\alpha}_0$ and $\hat{\beta}_0$, calculate $w^*_{ii} = \pi_i(1-\pi_i)$ and hence $\hat{\nu}_i = \hat{\eta}_i + (y_i - \hat{\pi}_i)/w^*_{ii}$.

STEP 2: Given $\hat{\nu}$, calculate new estimates $\hat{\alpha}_1$ and $\hat{\beta}_1$ using (3.9).

Generally, initial estimtes $\hat{\alpha}_0$, $\hat{\beta}_0$ are found using standard least squares with $\hat{\nu}_0 = (y + \frac{1}{2})/2$. This adjustment of the y_i's is necessary to prevent problems associated with a response of zero, namely the evalution of $\log(0)$ as the starting value for ν.

Also, note that the final predicted probabilities are found using the basis expansion, $C\Psi$, and not the raw functional data, X, since the basis expansion is a smoothed estimate of X. Using the raw data could result in unstable predictions, especially if the raw data are not smooth.

3.2.2 Choosing the Basis Dimension

For the basis method, the number of basis functions, m, is selected via the cross-validated log-likelihood. We have

$$\mathrm{CV}(m) = -2\sum_i \left(y_i \log \hat{\pi}_{i,-i} + (1-y_i)\log(1-\hat{\pi}_{i,-i})\right) \tag{3.10}$$

where $\hat{\pi}_{i,-i}$ is the predicted probability of a success for subject i when the parameter estimates are found without subject i. The optimal m value is the one which maximises the CV score.

3.3 Truncated Basis Expansion plus Penalty

As with functional regression, the smoothness of the functional parameter $\beta(t)$ can also be contolled via the addition of a roughness penalty (Rice and Silverman, 1991; Green and Silverman, 1994; Ramsay and Silverman, 1997). This has the advantage that the smoothing

can be controlled more precisely and the basis expansions can be truncated, for computational purposes, without substantially altering the results.

One common choice for the penalty is the integrated squared second derivative of the functional parameter: $P(\beta) = \int \beta''(t)^2 dt$. The parameter estimates are then found by maximising the penalised log-likelihood function

$$l(\pi; y) - \frac{1}{2} h \int \beta''(t)^2 dt \tag{3.11}$$

Under the penalised log-likelihood, the estimate for α remains unchanged but not $\beta(t)$.

Using the basis expansion for $\beta(t)$, the penalty term becomes

$$\begin{aligned} P(\beta) &= \int b^T \psi''(t)(\psi''(t))^T b \; dt \\ &= b^T P_\psi b \end{aligned}$$

where $P_\psi = \int \psi''(t)(\psi''(t))^T dt$. The derivative of the penalised log-likelihood with respect to b is

$$\frac{\partial l}{\partial b} = (CW)^T (y - \hat{\pi}) - h P_\psi b$$

and the Fisher information for b is given by

$$-E\left(\frac{\partial^2 l}{\partial b \partial b^T}\right) = (CW)^T w^* CW + h P_\psi$$

Thus, using Fisher scoring, the estimate for b becomes

$$\begin{aligned} b^{new} &= b - \left(E\left(\frac{\partial^2 l}{\partial b \partial b^T}\right)\right)^{-1} \left(\frac{\partial l}{\partial b}\right) \\ &= b + \left((CW)^T w^* CW + h P_\psi\right)^{-1} \left((CW)^T (y - \hat{\pi}) - h P_\psi b\right) \\ &= \left((CW)^T w^* CW + h P_\psi\right)^{-1} \left((CW)^T w^* CW b + h P_\psi b + (CW)^T (y - \hat{\pi}) - h P_\psi b\right) \\ &= \left((CW)^T w^* CW + h P_\psi\right)^{-1} (CW)^T w^* (\nu - Z\hat{\alpha}) \end{aligned} \tag{3.12}$$

Thus combining equations (3.7) and (3.12), the parameter estimates are

$$\begin{bmatrix} \hat{\alpha} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} Z^T w^* Z & Z^T w^* CW \\ (CW)^T w^* Z & (CW)^T w^* CW + h P_\psi \end{bmatrix}^{-1} \begin{bmatrix} Z^T \\ (CW)^T \end{bmatrix} w^* \nu$$

under a penalised functional logistic regresion model. Once again, an iterative algorithm (McCullagh and Nelder, 1989) is used to find the estimates in practice.

3.3.1 Choosing m and h

As in the basis method, the number of basis functions and the smoothing parameter can be chosen using the cross-validated log-likelihood. However, the log-likelihood is now being maximised over m and h.

$$CV(m,h) = -2\sum_i \left(y_i \log \hat{\pi}^h_{i,-i} + (1-y_i)\log\left(1-\hat{\pi}^h_{i,-i}\right)\right) \tag{3.13}$$

where $\hat{\pi}^h_{i,-i}$ is the predicted probability of success for subject i when the parameter estimates are found without subject i and using a smoothing parameter of h.

3.4 Model Diagnostics

To compare functional logistic regression models, the residual deviance can be used. For functional logistic regression, this will reduce to (see McCullagh and Nelder, 1989, page 121 for details)

$$D(y;\hat{\pi}) = -2\hat{\eta}^T\hat{\pi} - 2\sum_i \log(1-\hat{\pi}_i) \tag{3.14}$$

For standard logistic regression, the deviance function is not uniquely defined: it depends on whether the data are grouped or ungrouped. However, with functional logistic regression we will always have ungrouped data since each subject has a unique functional covariate.

As with standard glm, the importance of a covariate(s) can be examined by looking at the change in deviance between the model with and the model without the covariate(s). The change in deviance between these two nested models, A and B say,

$$D(y;\hat{\pi}_A) - D(y;\hat{\pi}_B)$$

can be approximated by a χ^2 distribution.

Another common measure of the goodness of fit of a model is the Pearson X^2 test statistic. However, as described in McCullagh and Nelder (1989), the extreme sparseness of the

functional data reduces this statistic to

$$X^2 = \sum \frac{(y_i - \bar{y})^2}{\bar{y}(1-\bar{y})} = n$$

which is not a useful measure of the goodness of fit of a functional model.

Classification tables can be used as another indicator of a model's goodness-of-fit. Models can be compared by examining the percentages correctly classified within the two observed response groups. However, these tables depend upon the cut-off probability, or cut value, used. The cut value is the probability at which the classification of a subject changes from failure to success. For example, with a cut-off probability of 0.6, subjects with a predicted probability of success, $\hat{\pi}_i$, greater than 0.6 would be classified as having a success, while the remaining subjects would be classified as having a failure.

For any model, the sensitivity is defined to be the probability of predicting a success when the observed response is a success, and the false positive rate is the probability of predicting a success when the observed response is a failure. These two measures are used to find the best cut-off probability for the model. This value gives the best trade-off between the measures; we want the highest sensitivity possible without the corresponding false positive rate being too high, or over some acceptable limit.

Receiver Operating Characteristic (ROC) curves (Hanley, 1989) use the sensitivity and false positive rates to compare models. An ROC curve is produced by plotting these two measures; found for a range of cut-off probabilities. An example of an ROC curve is given in Figure 3.1. Generally, if the curve is below the "sensitivity = false positive rate" line, the model is a poor predictor; if it's above the line, the model is a useful predictor. Thus, when comparing models, the "higher" the curve the better the model.

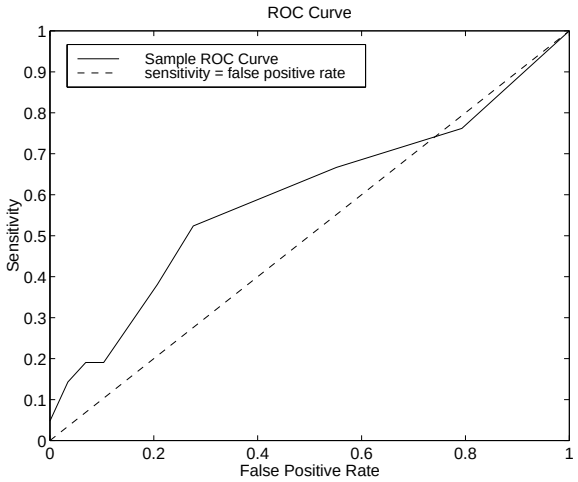


Figure 3.1: *Example of an ROC Curve.*

3.5 Application to EEG Data

These techniques were applied to electroencephalographic (EEG) recordings from human subjects. These recordings are part of a continuing larger study at the Department of Psychiatry, University of Sydney and Westmead Hospital. The data have been kindly provided by Dr Evian Gordon. The data (Figure 3.2) consist of evoked response potentials from the centre of the Frontal Lobe position of the brain for 91 subjects. For each subject, equispaced measurements were taken every four milliseconds over an eight second period; 50 measurements were taken before a noise stimulus was applied, one measurement at the stimulus, and 149 measurements after the stimulus. The EEG recordings were adjusted to remove the effect of normal brain wave activities, such as blinking. The response variable is the sex of the subject: the aim is to investigate if males and females process the noise stimulus differently, not to predict the sex of the patient.

Figure 3.3 shows EEG tracings for males and females separately, as well as the simple average EEG tracings for each sex. Both sexes follow the same general pattern with a rise followed by a fall in the EEG value after the noise stimulus before approximately returning to the baseline value. However, there are two main differences between the sexes: the average EEG tracings for females has a larger amplitude after the noise stimulus than males, and the EEG

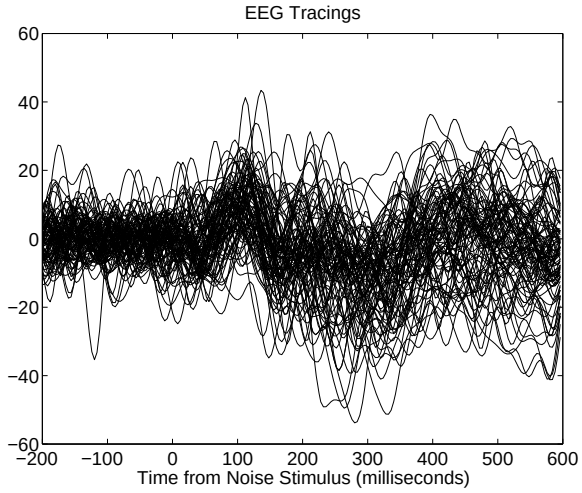


Figure 3.2: *EEG recordings for 91 subjects from the Frontal Lobe position of the brain.*

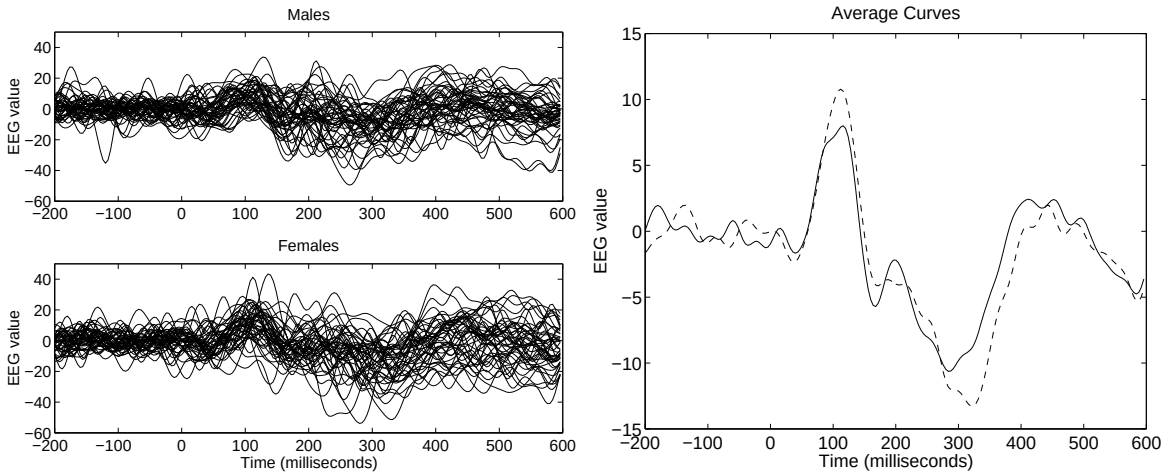


Figure 3.3: *Left: EEG tracings for males (top) and females (bottom). Right: Simple average EEG tracings for males (solid) and females (dashed).*

tracings before the stimulus would indicate that the sexes process random environmental influences differently. Thus, there may be a difference in the way males and females process a noise stimulus.

Using Fourier basis functions to model the traces, cross-validation (Figure 3.4) gave an optimal basis size of $m = 15$ and smoothing parameter of $h = 0$. Thus, the penalty term is not

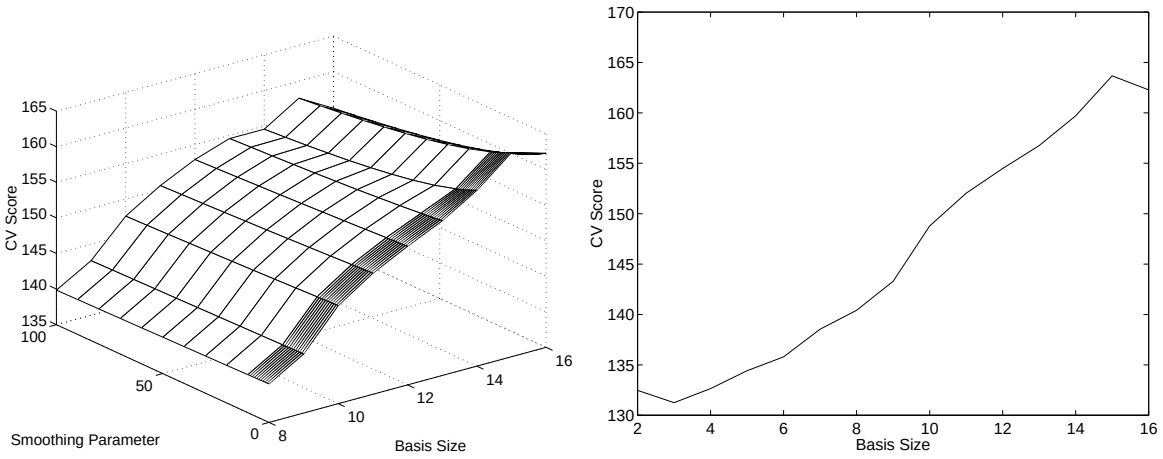


Figure 3.4: *Cross-validation plots for functional logistic regression of the EEG data.*

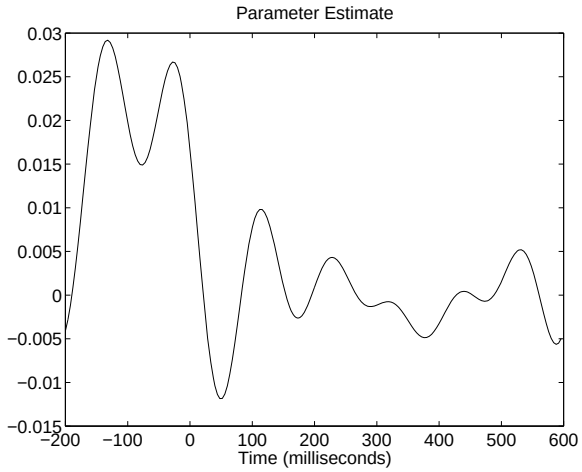


Figure 3.5: *Estimated functional parameter* $\beta(t)$ *using* $m = 15$ *Fourier basis functions for the EEG data.*

needed to control the smoothness for these data and the straight basis method can be used.

Assigning female as a "success" and male as a "failure", the resulting functional logistic model was

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -1.204 + \int x_i(t)\beta(t)\,dt$$

with the estimated functional parameter $\hat{\beta}(t)$ given in Figure 3.5. This model had a deviance of 104.567 on 75 degrees of freedom ($p = 0.014$). Other variables, such as age, had no

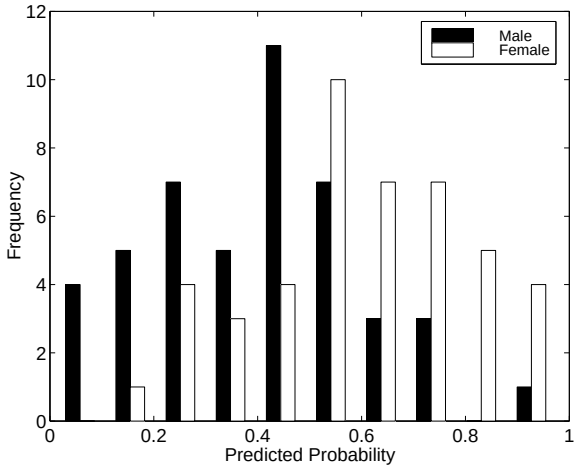


Figure 3.6: *Predicted probabilities of being female, split by known sex, in the EEG data.*

significant effect.

Figure 3.6 shows the distribution of the predicted probabilities, split by sex. There appears to be some distinction in the probabilities for males and females. The best cut-off probability was found to be $\hat{\pi} = 0.5$. Using this value, approximately three-quarters of the males and females were correctly classified (see Table 3.1). Thus, it appears there is a slight but significant difference in the EEG tracings between the sexes. The functional parameter appears to mainly use differences in the pre-stimulus tracings and the height of the initial reaction to the stimulus to differentiate between the sexes. Similar results could have also been obtained using 17 B-spline basis functions. Thus, for this EEG data the choice of basis functions was not too important.

In standard logistic regression, the choice of the starting value (β_0) is also not too important (McCullagh and Nelder, 1989), although a bad value can result in the algorithm diverging. We investigated the effect of different β_0's on the results for functional logistic regression using the EGG data. Under a range of starting values, the functional logistic regression algorithm converged to the same solution. The only effect was in the number of iterations to convergence. This ranged from six to eight iterations, with the proposed starting value from Section 2.2.1 achieving the optimal six iterations. So, different starting values only give

Classification Table for EEG Data			
	Observed Level		
Predicted	Male	Female	Total
Male	34	12	
Female	12	33	
% Correct	73.9%	73.3%	73.6%

Table 3.1: *Summary of classifications for sex using the basis method for functional logistic regression with a cut-off of* $\hat{\pi} = 0.5$.

a small change in the number of iterations to convergence (if convergence is achieved). This result was also evident in other data sets tested.

3.6 Discussion

In this chapter, we have presented a basis solution to functional logistic regression assuming a Bernoulli distribution for the observed responses. The maximum likelihood parameter estimates can be found easily using a slight modification of the existing glm algorithm. Like standard glm, the choice of the starting value was not too important, with the number of iterations being reduced by one or two. In the data sets tested, the optimal number of iterations (and convergence) was always achieved using $\nu = (y + \frac{1}{2})/2$. The number of basis functions, and the optimal smoothing parameter value in the penalised model, was chosen using a cross-validated log-likelihood function.

The EEG traces taken at the Frontal lobe (Fz) position of the brain were used to illustrate functional logistic regression. There appears to be a difference in the tracings of males and females. They appear to process their environment differently and females record more of a reaction to the stimulus in their tracing than males. For this data set it was found that the choice of basis functions was also not too important.

Chapter 4

Functional Data with a Repeated Stimulus

4.1 Introduction

In this chapter we extend the methods for both functional regression (Section 2.5.1) and functional logistic regression (Chapter 3) to the situation where there is a special structure in the functional data, viz a repeated stimulus. These models have been motivated by the fetal heart rate data set which is described in detail in Chapter 5.

The functional covariate now consists of two parts: the curve measured within the time frame of each stimulus, and the timing of the stimulus in relation to the other stimuli. The curves are functions of time although measurements are only taken at p points in time. Both of these parts need to be incorporated into the structure of the models.

4.2 Continuous Response

We begin by looking at the model for a continuous response. The proposed model is a modified version of the functional linear regression model given by Ramsay and Silverman (1997), incorporating the special repeated stimulus structure. Scalar (categorical or continuous)

covariates are also included. The regression model is

$$y_i = \alpha_0 + \sum_{j=1}^{r} \alpha_j z_{i,j} + \sum_{s=1}^{q} \gamma_s \int x_{i,s}(t)\beta(t)dt + \varepsilon_i, \qquad i = 1, \ldots, n \tag{4.1}$$

where

$y_i =$ the scalar response for the i^{th} subject,

$\alpha_0 =$ the constant parameter,

$\alpha_j =$ the parameter for the j^{th} scalar covariate,

$z_{i,j} =$ the j^{th} scalar covariate for the i^{th} subject,

$\gamma_s =$ the parameter for the s^{th} stimulus,

$x_{i,s}(t) =$ the functional covariate for the i^{th} subject, measured at time t within the s^{th} stimulus

$\beta(t) =$ the functional parameter for time within a stimulus, measured at time t

$\varepsilon_i =$ the error associated with subject i, $\sim NID(0, \sigma_\varepsilon^2)$

$i = 1, \ldots, n \qquad j = 1, \ldots, r \qquad s = 1, \ldots, q$

As before, the normal equations cannot be solved directly. There will always be more unknown parameters than subjects in the estimation of $\beta(t)$, resulting in an infinite number of solutions with a perfect but meaningless fit. Once again, we present a basis solution to this problem. Thus, the functional time parameter, $\beta(t)$, is modelled as

$$\beta(t) = \sum_{1}^{m} b_k \psi_k(t) = b^T \psi(t) \tag{4.2}$$

The functional covariate was measured at each (s, t) combination; resulting in $q \times p$ measurements for each subject. A basis expansion was also used to model the time measurements

within each stimulus

$$\begin{aligned} x_{i,s}(t) &= c_{i,s}^T \psi(t) \\ x_i &= [x_{i,1}(t) \ \ldots \ x_{i,q}(t)]^T \\ X_{i,(q\times p)} &= [x_i(t_1) \ \ldots \ x_i(t_p)]^T \\ C_{i,(q\times m)} &= [c_{i,1} \ \ldots \ c_{i,q}]^T \\ \Psi_{(m\times p)} &= [\psi(t_1) \ \ldots \ \psi(t_p)]^T \\ \Rightarrow \ X_i &= C_i \Psi \end{aligned} \tag{4.3}$$

Estimates for C_i were found using least squares; $C_i = X_i \Psi^T \left(\Psi \Psi^T\right)^{-1}$.

Substituting (4.2) and (4.3) into the regression model (4.1) gives

$$\begin{aligned} y_i &= \alpha_0 + \sum_{j=1}^{r} \alpha_j z_{i,j} + \sum_{s=1}^{q} \gamma_s \int c_{i,s}^T \psi(t) \psi^T(t) b \, dt + \varepsilon_i \\ &= z_i^T \alpha + \gamma^T C_i W b + \varepsilon_i \end{aligned}$$

where $\alpha = [\alpha_0 \ \alpha_1 \ \ldots \ \alpha_r]^T$, $\gamma = [\gamma_1 \ \ldots \ \gamma_q]^T$, $z_i = [1 \ z_{i,1} \ \ldots \ z_{i,r}]^T$ and $W = \int \psi(s)\psi^T(s)ds$ $= \Psi \Psi^T$. Thus, the parameter estimates, $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$, are found by solving

$$y_i = z_i^T \hat{\alpha} + \hat{\gamma}^T C_i W \hat{b} \tag{4.4}$$

and using equation (4.2) to estimate $\hat{\beta}$. The predicted responses, $\hat{y}_i$, are calculated using these estimates.

$$\hat{y}_i = z_i^T \hat{\alpha} + \hat{\gamma}^T C_i W \hat{b} \tag{4.5}$$

As in Section 3.2.1, the predicted values are found using $C_i W \hat{b}$ and not $\int x_{i,s}(t)\hat{\beta}(t)dt$, as the raw data, $x_{i,s}(t)$, may not be smooth. Any roughness, or irregularities, in $x_{i,s}(t)$ would result in extremely inaccurate (unstable) predictions.

4.2.1 Estimating Parameters

Least squares parameter estimates were found by minimising the error, or residual, sum of squares (RSS).

$$\begin{aligned} RSS &= \sum_i (y_i - \hat{y}_i)^2 \\ &= \sum_i \left[y_i - \left(z_i^T \hat{\alpha} + \hat{\gamma}^T C_i W \hat{b} \right) \right]^2 \end{aligned}$$

For each parameter, RSS is minimised by equating the appropriate partial derivative to zero, i.e. by solving $\partial RSS/\partial\hat{\alpha} = 0$, $\partial RSS/\partial\hat{b} = 0$ and $\partial RSS/\partial\hat{\gamma} = 0$ simultaneously for $\hat{\alpha}$, $\hat{b}$ and $\hat{\gamma}$. Taking partial derivatives gives

$$\begin{aligned} \frac{\partial RSS}{\partial\hat{\alpha}} &= -2 \sum_i z_i \left[y_i - \left(z_i^T \hat{\alpha} + \hat{\gamma}^T C_i W \hat{b} \right) \right] \\ &= -2 \left[\sum_i z_i y_i - \sum_i z_i z_i^T \hat{\alpha} - \sum_i z_i \hat{\gamma}^T C_i W \hat{b} \right] = 0 \\ \frac{\partial RSS}{\partial\hat{b}} &= -2 \sum_i (\hat{\gamma}^T C_i W)^T \left[y_i - \left(z_i^T \hat{\alpha} + \hat{\gamma}^T C_i W \hat{b} \right) \right] \\ &= -2 \left[\sum_i W^T C_i^T \hat{\gamma} y_i - \sum_i W^T C_i^T \hat{\gamma} z_i^T \hat{\alpha} - \sum_i W^T C_i^T \hat{\gamma} \hat{\gamma}^T C_i W \hat{b} \right] = 0 \\ \frac{\partial RSS}{\partial\hat{\gamma}} &= -2 \sum_i \left[y_i - \left(z_i^T \hat{\alpha} + \hat{\gamma}^T C_i W \hat{b} \right) \right] (C_i W \hat{b})^T \\ &= -2 \left[\sum_i y_i \hat{b}^T W^T C_i^T - \sum_i z_i^T \hat{\alpha} \hat{b}^T W^T C_i^T - \sum_i \hat{\gamma}^T C_i W \hat{b} \hat{b}^T W^T C_i^T \right] = 0 \end{aligned}$$

Thus,

$$\begin{aligned} \hat{\alpha} &= \left(\sum_i z_i z_i^T \right)^{-1} \left(\sum_i z_i y_i - \sum_i z_i \hat{\gamma}^T C_i W \hat{b} \right) \\ \hat{b} &= \left(\sum_i W^T C_i^T \hat{\gamma} \hat{\gamma}^T C_i W \right)^{-1} \left(\sum_i W^T C_i^T \hat{\gamma} y_i - \sum_i W^T C_i^T \hat{\gamma} z_i^T \hat{\alpha} \right) \\ \hat{\gamma}^T &= \left(\sum_i y_i \hat{b}^T W^T C_i^T - \sum_i z_i^T \hat{\alpha} \hat{b}^T W^T C_i^T \right) \left(\sum_i C_i W \hat{b} \hat{b}^T W^T C_i^T \right)^{-1} \end{aligned}$$

Substituting for $\hat{\alpha}$ and $\hat{b}$, results in parameter estimates given by

$$\hat{\alpha} = \left(\sum_i z_i z_i^T - \left(\sum_i z_i \hat{\gamma}^T C_i W\right)\left(\sum_i W^T C_i^T \hat{\gamma}\hat{\gamma}^T C_i W\right)^{-1}\left(\sum_i W^T C_i^T \hat{\gamma} z_i^T\right)\right)^{-1}$$
$$\times \left(\sum_i z_i y_i - \left(\sum_i z_i \hat{\gamma}^T C_i W\right)\left(\sum_i W^T C_i^T \hat{\gamma}\hat{\gamma}^T C_i W\right)^{-1}\left(\sum_i W^T C_i^T \hat{\gamma} y_i\right)\right)$$

$$\hat{b} = \left(\sum_i W^T C_i^T \hat{\gamma}\hat{\gamma}^T C_i W - \left(\sum_i W^T C_i^T \hat{\gamma} z_i^T\right)\left(\sum_i z_i z_i^T\right)^{-1}\left(\sum_i z_i \hat{\gamma}^T C_i W\right)\right)^{-1}$$
$$\times \left(\sum_i W^T C_i^T \hat{\gamma} y_i - \left(\sum_i W^T C_i^T \hat{\gamma} z_i^T\right)\left(\sum_i z_i z_i^T\right)^{-1}\left(\sum_i z_i y_i\right)\right)$$

$$\hat{\gamma}^T = \left(\sum_i y_i \hat{b}^T W^T C_i^T - \sum_i z_i^T \hat{\alpha}\hat{b}^T W^T C_i^T\right)\left(\sum_i C_i W \hat{b}\hat{b}^T W^T C_i^T\right)^{-1}$$

Now, let

$$D = \begin{bmatrix} \hat{\gamma}^T C_1 W \\ \vdots \\ \hat{\gamma}^T C_n W \end{bmatrix}$$
$$E = \left[(C_1 W \hat{b}) \ \ldots \ (C_n W \hat{b})\right]^T$$
$$y = [y_1 \ \ldots \ y_n]^T$$
$$Z = [z_1 \ \ldots \ z_n]^T$$

Then, summing over the subjects, the parameter estimates become

$$\hat{\alpha} = \left(Z^T Z - Z^T D(D^T D)^{-1} D^T Z\right)^{-1}\left(Z^T y - Z^T D(D^T D)^{-1} D^T y\right)$$
$$\hat{b} = \left(D^T D - D^T Z(Z^T Z)^{-1} Z^T D\right)^{-1}\left(D^T y - D^T Z(Z^T Z)^{-1} Z^T y\right)$$
$$\hat{\gamma}^T = \left(y^T E - (Z\hat{\alpha})^T E\right)\left(E^T E\right)^{-1}$$

which reduce to

$$\hat{\alpha} = \left(Z^T\left(I - D(D^T D)^{-1} D^T\right) Z\right)^{-1} Z^T\left(I - D(D^T D)^{-1} D^T\right) y \tag{4.6}$$
$$\hat{b} = \left(D^T\left(I - Z(Z^T Z)^{-1} Z^T\right) D\right)^{-1} D^T\left(I - Z(Z^T Z)^{-1} Z^T\right) y \tag{4.7}$$
$$\hat{\gamma}^T = (y - Z\hat{\alpha})^T E\left(E^T E\right)^{-1} \tag{4.8}$$

Thus, both $\hat{\alpha}$ and $\hat{b}$ depend on $\hat{\gamma}$ and $\hat{\gamma}$ depends on $\hat{\alpha}$ and $\hat{b}$. Note that equations (4.6) and (4.7) can be combined and written as

$$\begin{bmatrix} \hat{\alpha} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} Z^T Z & Z^T D \\ D^T Z & D^T D \end{bmatrix}^{-1} \begin{bmatrix} Z^T \\ D^T \end{bmatrix} y \tag{4.9}$$

These estimates lend themselves to an iterative solution. The algorithm used to find the estimates is described below. In semi-parametric regression, this type of problem can be solved without iteration (Speckman, 1988). However, developing a non-iterative approach in a FDA setting is a future research topic.

Algorithm

In practice, the following iterative algorithm is used to estimate the parameters.

STEP 0: Find initial estimates for α and b.

These estimates are found by setting $\hat{\gamma}_i = 1$ for all i. These γ values correspond to the assumption that the timing of the stimulus has no effect. The initial estimates, $\hat{\alpha}_{(0)}$ and $\hat{b}_{(0)}$, are then found using equation (4.9),

$$\begin{bmatrix} \hat{\alpha}_{(0)} \\ \hat{b}_{(0)} \end{bmatrix} = \begin{bmatrix} Z^T Z & Z^T D_{(0)} \\ D_{(0)}^T Z & D_{(0)}^T D_{(0)} \end{bmatrix}^{-1} \begin{bmatrix} Z^T \\ D_{(0)}^T \end{bmatrix} y$$

where $D_{(0)}$ is calculated using $\hat{\gamma}_{(0)} = \mathbf{1}$, a $(q \times 1)$ vector of ones.

At iteration k, the parameter estimates, $\hat{\alpha}_{(k)}$, $\hat{b}_{(k)}$ and $\hat{\gamma}_{(k)}$, are given by:

STEP 1: Estimate $\gamma_{(k)}$, using $\hat{\alpha}_{(k-1)}$ and $\hat{b}_{(k-1)}$.

$$\hat{\gamma}_{(k)}^T = \left(y - Z\hat{\alpha}_{(k-1)}\right)^T E_{(k-1)} \left(E_{(k-1)}^T E_{(k-1)}\right)^{-1}$$

where $E_{(k-1)}$ is calculated using $\hat{b}_{(k-1)}$.

STEP 2: Estimate α and b, using $\hat{\gamma}_{(k)}$.

$$\begin{bmatrix} \hat{\alpha}_{(k)} \\ \hat{b}_{(k)} \end{bmatrix} = \begin{bmatrix} Z^T Z & Z^T D_{(k)} \\ D_{(k)}^T Z & D_{(k)}^T D_{(k)} \end{bmatrix}^{-1} \begin{bmatrix} Z^T \\ D_{(k)}^T \end{bmatrix} y$$

where $D_{(k)}$ is calculated using $\hat{\gamma}_{(k)}$.

STEP 3: Check for convergence.

If the algorithm has converged, then stop;

else repeat from Step 1.

Convergence is determined by examining the relative change in the parameter estimates from one iteration to the next. If the changes are sufficiently small, then the algorithm is assumed to have converged.

4.2.2 Model Diagnostics

The assumptions of the model can be tested by examining the residuals, $\hat{\varepsilon}_i = y_i - \hat{y}_i$, as in standard linear regression. For example, the assumption of normally distributed residuals could be checked via a normal probability plot of the residuals. The overall fit of any model can be evaluated using the usual coefficient of determination R^2.

As in standard linear regression, the importance of any scalar covariate can be determined using a t-test on $n-(m+q+r)$ degrees of freedom. The null hypothesis is that the covariate and the response are independent and the alternate hypothesis is that the two are related. The standard error of $\hat{\alpha}_j$ is given by $\hat{\sigma}_\varepsilon\sqrt{s_{jj}}$, where

$$\hat{\sigma^2}_\varepsilon = \frac{RSS}{n-m-q-r}$$

and s_{jj} is the j-th diagonal element of

$$S = \begin{bmatrix} Z^TZ & Z^TD \\ D^TZ & D^TD \end{bmatrix}^{-1}$$

The importance of the functional covariate can be determined via a partial F test. If the functional covariate and the response are independent, then either one or both of the functional parameters will be zero since the effect of the two parameters is multiplicative. Thus the null and alternate hypotheses for the test are:

$$H_0 : \gamma = 0 \text{ and/or } \beta(t) = 0$$

$$H_1 : \gamma \neq 0 \text{ and } \beta(t) \neq 0$$

The partial F statistic is calculated by fitting two models. A model containing only the r scalar covariates plus a constant is first fitted to the data to obtain a residual sum of squares (RSS_0). The full model, with the $m+q$ functional parameters, is then fitted to obtain a second residual sum of squares (RSS_f). The partial F statistic is then

$$F = \frac{(RSS_f - RSS_0)\,/\,(m+q)}{RSS_f\,/\,(n-m-q-r)}$$

which has the $F_{m+q,n-m-q-r}$ distribution under H_0. If $F > F_{m+q,n-m-q-r,\alpha}$, then the null hypothesis is rejected and the functional covariate is judged to be a significant predictor of the response.

4.2.3 Choosing m

The number of basis functions, m, used to model both the functional time parameter, $\beta(t)$, and the raw data, $x_{i,s}(t)$, must be chosen. Cross-validation (Rice and Silverman, 1991) has been used for this purpose. Cross-validation involves omitting each subject in turn, estimating the parameters with the remaining subjects, and then using these estimates to predict the response for the missing subject. For functional regression, the number of basis functions is found by minimising the cross-validated residual sum of squares

$$\begin{aligned} \mathrm{CV}(m) &= n^{-1}\sum_i \hat{\varepsilon}^2_{i,-i}(m) \\ &= y_i - z_i^T\hat{\alpha}_{-i} - \hat{\gamma}_{-i}^T C_i W \hat{b}_{-i} \end{aligned} \tag{4.10}$$

where $\hat{\varepsilon}_{i,-i}$ is the error associated with the predicted value for the i^{th} subject when the parameter estimates are found omitting the i^{th} subject; and where C_i, W, and $\hat{b}_{-i}$ are all found using m basis functions.

The choice of m also depends on the scalar parameters in the model, since these parameter estimates depend on β. Thus, a two step procedure is used to find the best model for each response. Firstly, cross-validation is used to find the best m value for the functional model containing no scalar covariates. This m value is then used to find the significant scalar covariates. Once the parameters that are important have been determined, cross-validation is re-run to get the optimal number of basis functions for the final model.

Using a non-iterative procedure for estimating the parameters, eg. analogous to that developed by Speckman (1988) in semiparametric regression, would only require cross-validation to be performed once. However, as previously mentioned, this type of approach is yet to be developed.

4.3 Binary Response

We now propose a model for a binary response with a repeatedly stimulated functional covariate, using a modification of functional logistic regression (Chapter 3). Assuming that the observed response, y_i, comes from a Bernoulli distribution with probability of success π_i, the functional logistic regression model incorporating the repeated stimulus is

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \sum_{j=1}^{r} \alpha_j z_{i,j} + \sum_{s=1}^{q} \gamma_s \int x_{i,s}(t)\beta(t)dt, \qquad i = 1, \ldots, n \tag{4.11}$$

where α_0, α_j, $z_{i,j}$, γ_s, $x_{i,s}(t)$, $\beta(t)$ are defined in Section 4.2. Once again, we model the functional parameter and covariate using basis expansions, as given in (4.2), (4.3). Using these, the logistic regression model (4.11) becomes

$$\begin{aligned} log\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha_0 + \sum_{j=1}^{r} \alpha_j z_{i,j} + \sum_{s=1}^{q} \gamma_s \int x_{i,s}(t)\beta(t)dt, \qquad i = 1, \ldots, n \\ &= z_i^T \alpha + \gamma^T C_i W b \end{aligned}$$

where again $W = \int \psi(s)\psi^T(s)ds = \Psi\Psi^T$.

4.3.1 Estimating Parameters

Maximum likelihood parameter estimates are found using a combination of the generalized linear modelling (glm) algorithm (McCullagh and Nelder, 1989) and the algorithm for functional regression with a repeated stimulus, given in section 4.2.1.

As with standard glms, the actual responses, y_i, are assumed to come from a Bernoulli(π_i)

distribution. Thus, the log-likelihood function can be expressed as

$$l(\pi; y) = \sum_{i=1}^{n} \left[y_i \, log \left(\frac{\pi_i}{1-\pi_i} \right) + log(1-\pi_i) \right] \tag{4.12}$$

Parameter estimates can be found by maximising equation (4.12) with respect to α, b and γ. This is done using the Fisher scoring method (Fisher, 1925).

Firstly, the derivative of the log-likelihood function with respect to π_i is

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i - \pi_i}{\pi_i(1-\pi_i)}$$

Using

$$log \left(\frac{\pi_i}{1-\pi_i} \right) = \eta_i = z_i^T \alpha + \gamma^T C_i W b$$

the derivative of the log-likelihood with respect to α_k is

$$\begin{aligned} \frac{\partial l}{\partial \alpha_k} &= \sum_{i=1}^{n} \frac{\partial l}{\partial \pi_i} \frac{d\pi_i}{d\eta_i} \frac{\partial \eta_i}{\partial \alpha_k} \\ &= \sum_{i=1}^{n} \frac{y_i - \pi_i}{\pi_i(1-\pi_i)} \pi_i(1-\pi_i) z_{i,k}^T \\ &= \sum_{i=1}^{n} (y_i - \pi_i) z_{i,k}^T \end{aligned}$$

The Fisher information for α is given by

$$\begin{aligned} -E \left(\frac{\partial^2 l}{\partial \alpha_j \partial \alpha_k} \right) &= \sum_{i=1}^{n} \frac{1}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \alpha_j} \frac{\partial \pi_i}{\partial \alpha_k} \\ &= \sum_{i=1}^{n} \frac{1}{\pi_i(1-\pi_i)} \pi_i(1-\pi_i) z_{i,j}^T \pi_i(1-\pi_i) z_{i,k}^T \\ &= \sum_{i=1}^{n} \pi_i(1-\pi_i) z_{i,j}^T z_{i,k}^T \\ &= \{ Z^T w^* Z \}_{jk} \end{aligned}$$

where w^* is the diagonal weights matrix with $w_i^* = \hat{\pi}_i(1-\hat{\pi}_i)$. By Fisher's scoring method, the new estimate of α, $\hat{\alpha}^{new}$, is given by

$$-E \left(\frac{\partial^2 l}{\partial \alpha^2} \right) \hat{\alpha}^{new} = -E \left(\frac{\partial^2 l}{\partial \alpha^2} \right) \hat{\alpha} + \frac{\partial l}{\partial \alpha}$$

Now

$$\begin{aligned} -E \left(\frac{\partial^2 l}{\partial \alpha^2} \right) \hat{\alpha} &= Z^T w^* Z \hat{\alpha} \\ &= Z^T w^* (\hat{\eta} - D\hat{b}) \end{aligned}$$

Thus, the new estimate is found by solving

$$\begin{aligned} Z^T w^* Z\hat{\alpha}^{new} &= Z^T w^*(\hat{\eta} - D\hat{b}) + Z^T(y - \hat{\pi}) \\ &= Z^T w^*(\nu - D\hat{b}) \\ \Rightarrow \qquad \hat{\alpha} &= \left(Z^T w^* Z\right)^{-1} Z^T w^*(\nu - D\hat{b}) \\ \nu &= \hat{\eta} + (w^*)^{-1}(y - \hat{\pi}) \end{aligned} \tag{4.13}$$

Similarly,

$$\hat{b} = \left(D^T w^* D\right)^{-1} D^T w^*(\nu - Z\hat{\alpha}) \tag{4.14}$$

$$\hat{\gamma}^T = (\nu - Z\hat{\alpha})^T w^* E \left(E^T w^* E\right)^{-1} \tag{4.15}$$

Note that equations (4.13) and (4.14) for $\hat{\alpha}$ and $\hat{b}$ can be combined and written as

$$\begin{bmatrix} \hat{\alpha} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} Z^T w^* Z & Z^T w^* D \\ D^T w^* Z & D^T w^* D \end{bmatrix}^{-1} \begin{bmatrix} Z^T \\ D^T \end{bmatrix} w^* \nu \tag{4.16}$$

As with functional regression with a repeated stimulus, these estimates lend themselves to an iterative solution.

Algorithm

STEP 0: Find an initial estimate for η_i, $\tilde{\eta}_i$.

$$\tilde{\eta}_i = (y_i + 0.5)/2$$

Since y_i is either 0 or 1, we need to adjust the response by 0.5 (see McCullagh and Nelder, 1989). Using $\tilde{\eta}_i$, find estimates for the parameters α, γ, and b. These estimates are found by applying the functional regression algorithm from section 4.2.1 with $\tilde{\eta} = [\tilde{\eta}_1 \ \ldots \ \tilde{\eta}_n]^T$ as the response.

$$\begin{bmatrix} \hat{\alpha} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} Z^T Z & Z^T D \\ D^T Z & D^T D \end{bmatrix}^{-1} \begin{bmatrix} Z^T \\ D^T \end{bmatrix} \tilde{\eta}$$

$$\hat{\gamma}^T = (\tilde{\eta} - Z\hat{\alpha})^T E \left(E^T E\right)^{-1}$$

STEP 1: Find the linear predictors $\hat{\eta}_i$ for the regression using the parameter estimates.

$$\hat{\eta}_i = z_i^T \hat{\alpha} + \hat{\gamma}^T C_i W \hat{b}$$

STEP 2: Calculate the probability of a success for each subject.

$$\hat{\pi}_i = \frac{exp(\hat{\eta}_i)}{1 + exp(\hat{\eta}_i)}$$

STEP 3: Calculate the adjusted dependent variable ν_i and weights w_i.

$$\begin{aligned} w_i &= \hat{\pi}_i(1 - \hat{\pi}_i) \\ \nu_i &= \hat{\eta}_i + \frac{y_i - \hat{\pi}_i}{w_i} \end{aligned}$$

STEP 4: Re-estimate the parameter values via a weighted version of the functional regression algorithm in section 4.2.1, with ν_i as the response variable.

$$\nu_i = z_i^T \alpha + \gamma^T C_i W b$$

where the parameter estimates in the algorithm are found using equations (4.16) and (4.15), with w^* as the diagonal weights matrix and ν the vector of adjusted dependent variables.

$$\begin{bmatrix} \hat{\alpha} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} Z^T w^* Z & Z^T w^* D \\ D^T w^* Z & D^T w^* D \end{bmatrix}^{-1} \begin{bmatrix} Z^T \\ D^T \end{bmatrix} w^* \nu$$

$$\hat{\gamma}^T = (\nu - Z\hat{\alpha})^T w^* E \left(E^T w^* E\right)^{-1}$$

STEP 5: Check for convergence.

If the algorithm has converged, then STOP;

else REPEAT from STEP 1.

Convergence is determined by looking at the relative change in the parameter estimates from one iteration to the next. If the changes are sufficiently small, the algorithm is assumed to have converged.

4.3.2 Choosing m

The number of basis functions, m, in any model can be chosen using the cross-validation techniques given for the functional logistic regression model (Section 3.2.2). Similarly, the goodness-of-fit of any model can be determined using techniques such as residual deviance and classification tables, given in section 3.4.

Chapter 5

The Fetal Heart Rate Data

5.1 Introduction

The fetal heart rate data[1] consists of measurements from periodically stimulated fetal heart rate tracings. Approximately 6000 measurements per subject were recorded over a 19 minute period for 73 subjects. The aim of the study was to determine if the fetal heart rate responses to the stimulus are predictive of birth outcomes, and the infant's development at 18 and 36 months of age.

In the past, researchers have analysed this response to stimulus data using the concept of habituation, which is defined as the decrease leading to cessation of a behavioural response that occurs when an initially novel stimulus is presented repeatedly (Thompson and Glansman, 1966). It has been shown that the habituation pattern reflects the processes of the central nervous system. A normal habituation pattern is evidence of an intact and fully functioning central nervous system (Jeffrey and Cohen, 1971; Lewis, 1971; Madison *et al.*, 1986) while an impaired habituation pattern may indicate some form of brain damage (Holloway and Parsons, 1971). For example, hyperactive (Hutt and Hutt, 1964; Tizard, 1968) and autistic children (Hutt *et al.*, 1965) have been found to have impaired habituation, as have high risk (Eisenberg *et al.*, 1966) and traumatized newborn infants (Bronstein *et al.*, 1968). Over-

[1]Supplied by Dr Leo Leader, School of Obstetrics and Gynaecology, University of New South Wales, Sydney, Australia

all, it has been concluded that habituation in infants can more accurately predict cognitive development than traditional tests (Bornstein and Sigman, 1986; McCall and Carriger, 1993).

However, the notion of habituation is partially subjective due to the way in which it is defined. Three main definitions have been used: a response less than some percentage of the initial reponse (Ferguson *et al.*, 1978); a response less than some amount for all subjects (Buckwald and Humphrey, 1973); and a lack of response for a specified number of successive stimuli (Brackbill *et al.*, 1974; Leader *et al.*, 1982). For these criteria, the researcher chooses the percentage, the response amount or what qualifies as a lack of response and the number of successive stimuli.

In order to remove the subjective definition of habituation, we have used the entire stimulated fetal heart rate tracings over the 19 minute period instead of determining habituation. Since there are many more heart rate measurements per fetus than there are fetuses in the study, functional data analysis techniques have been used to analyse the data.

5.2 Study Description

The fetal heart rate measurements were taken on 73 pregnant women, in late pregnancy. For each subject, the fetal heart rate was recorded over a 19 minute period, 14 days or less before birth. The fetus was stimulated by placing a vibroacoustic stimulator on the mother's abdomen over the fetal head. A one-second stimulus was given every minute for a total of 19 stimuli. It is expected that this stimulus initially results in an increase in the fetal heart rate, with the response decreasing with repetition of the stimulus. The heart rates were measured every 0.2 seconds from 5 seconds before the first stimulus, until 55 seconds after the 19th stimulus.

Of the 73 fetuses, 10 were removed from the study: 1 since the fetus failed to respond to the stimulus until approximately the 15th minute, 1 due to the heart rate monitor continually

dropping out, 3 fetuses who had their heart rate measured for less than 19 minutes and 5 premature babies, since the heart rate was not taken during late pregnancy. These exclusions resulted in a final sample of 63 subjects.

Besides the heart rates, other covariates were included in the study. These included the gestational age of the fetus at the time of the heart rate measurements, and the sex of the infant. Maternal measurements included parity (number of previous births), age at delivery (in years), and average smoking, drinking and caffeine intake levels. Parental influences upon the infant's development were accounted for through two other measurements: a global measure of the socio-ecomonic factors, which includes the highest level of education and occupation of both parents; and the Home Observation for Measurement of the Environment (HOME) (Caldwell *et al.*, 1967), which examines the level of emotional support and parental involvement available to the infant at home.

As previously stated, the aim was to determine if the fetal heart rate responses to the stimulus are predictive of birth outcomes, and of the infant's development at 18 and 36 months of age, as defined using the Bayley Scales of Infant Development (BSID) (Bayley, 1993) (see Appendix A for further details). Two of the outcomes are presented here: the infant's risk category at birth, and psychomotor development at 18 months of age.

In addition, a further 10 unstimulated subjects were included in the study. For these subjects, the fetal heart rate measurements were recorded every 0.2 seconds for 16-19 minutes *without* a stimulus being applied. These subjects were used to determine if stimulation was necessary for the heart rates to be predictive. However, only the birth outcomes were available for these subjects; the 18 month results were not collected in six cases and in four cases the infant had not reached 18 months of age at the time of writing. Validation of the method using the controls was therefore only possible for the birth outcome.

To begin with, the outcomes were modelled using standard linear or logistic regression without the functional covariate. Functional linear or logistic regression for a repeated stimulus

(Chapter 4) was then used to find the best model containing the functional heart rate covariate. Thus, comparisons between the best models with and without the heart rate could be made.

5.2.1 Description of Covariates

For the fetal data, the functional covariate is the fetal heart rates, or pulse. We assume that the functional time parameter within a stimulus $\beta(t)$ is the same for all of the stimuli; that is, $\beta(t)$ is independent of the timing of the stimuli. Since measurements were taken every 0.2 seconds for 5 seconds before the stimulus, at the stimulus, and every 0.2 seconds for 55 seconds after the stimulus, there were $p = 25 + 1 + 275 = 301$ measurements taken within each stimulus. Although each stimulus actually contained one minute and 0.2 seconds worth of data, for simplicity we refer to this as a minute.

For the analyses, the raw heart rate data and the time parameter, β, were modelled using Fourier basis functions since the heart rates were roughly periodic. This is due to each fetus having an initial base heart rate to which they approximately returned after each stimulus, but before the next stimulus was applied. The equations for the parameter estimates simplify slightly using Fourier basis functions since

$$W = \int \psi(s)\psi^T(s)ds = I_m$$

where I_m is the order m identity matrix.

As well as the pulse covariate, three scalar covariates were found to be important in the modelling of the outcomes:

- *agegp35*

  The mother's age at delivery was divided into two groups; < 35 years of age [level 0] and ≥ 35 years of age [level 1]. Of the 63 subjects in the study, 50 (79.4%) of the mothers were younger than 35 years of age.

- *parity*

  Parity was also divided into two groups: no previous births [level 0], and one or more previous births [level 1]. For this sample, 37 (58.7%) of the mothers had borne no previous children.

- *sex*

  The sex variable was used to indicate the sex of the child; 0 for male and 1 for female. Thirty-four (54.0%) of the infants were male.

Since the heart rate tests were conducted in late pregnancy, the average gestational age at the tests was 38.5 weeks, with a standard deviation of 1.63. The maximum gestational age was 42 weeks and the minimum was 35 weeks. Within this range, the actual gestational age was not significant in any model. Other covariates were also used but were found to be insignificant.

5.3 Risk Category

For the risk category, subjects were divided into two groups depending on their antenatal course and other birth outcomes: normal and high risk. The high risk group included medicated hypertension, fetal distress, intra uterine growth retardation, and poor doppler. In this study, 46 of the 63 pregnancies (73.0%) were considered to be normal [level 0], while 17 (27.0%) were considered high risk [level 1].

None of the covariates was significant in a logistic regression model using only the scalar covariates. However, for comparison purposes, the best of these insignificant models is given below. Included in this model were the mother's age group and her parity. Mothers who have never borne a child are 2.48 (= 1/0.403) times more likely to have a high risk birth outcome than mothers who have borne previous children. Also, mothers over the age of 35 are 2.57 times more likely to have a high risk outcome than younger mothers (see Table 5.1). Under this model, using a classification cut-off of $\hat{\pi} = 0.5$, only 3/17 (17.6%) of the high risk pregnancies were correctly identified (see Table 5.2). The cut-off of $\hat{\pi} = 0.5$ was chosen as

Logistic Regression Analysis for Risk Category				
MODEL: parity + age group				
Covariate	**Coef**	**StDev**	**P**	**Exp(coef)**
Constant	-0.875	0.373	0.019	
Parity (1)	-0.909	0.648	0.161	0.403
Agegp35 (1)	0.945	0.706	0.181	2.573
D = 70.314				

Table 5.1: *Logistic Regression Summary for Risk Category.*

Classification Table for Risk Category			
MODEL: parity + age group			
	Observed Level		
Predicted	Normal	High Risk	Total
Normal	44	14	
High Risk	2	3	
% Correct	95.7%	17.6%	74.6%

Table 5.2: *Logistic Regression - Summary of Classifications for Risk Category with cut-off of* $\hat{\pi} = 0.5$.

it gave the best result in terms of balancing the false positive rate and the sensitivity of the model.

Using functional logistic regression with a repeated stimulus, the best model contained only the functional pulse covariate; all other scalar covariates were not significant, as expected. The functional model was found to be

$$log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -0.910 + \sum_{s=1}^{19} \hat{\gamma}_s \int x_{i,s}(t)\hat{\beta}(t)dt \tag{5.1}$$

with a deviance of 84.497 on 38 degrees of freedom. The functional time parameter, $\beta(t)$, was modelled using $m = 5$ basis functions. The classification table for this model is given in Table

Classification Table for Risk Category MODEL: pulse			
	Observed Level		
Predicted	Normal	High Risk	Total
Normal	45	1	
High Risk	1	16	
% Correct	97.8%	94.1%	96.8%

Table 5.3: *Functional Logistic Regression - Summary of Classifications for Risk Category with cut-off of* $\hat{\pi} = 0.6$.

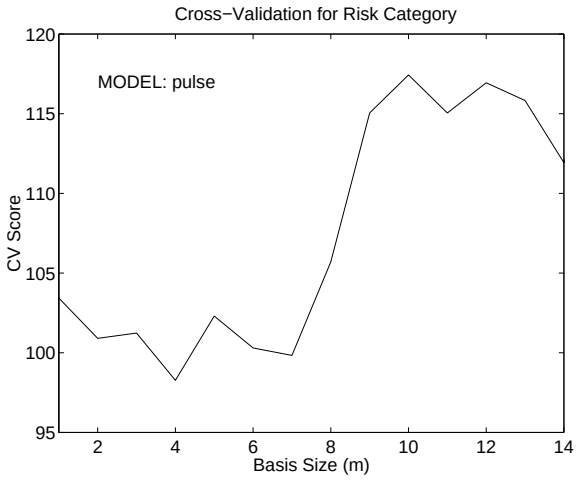
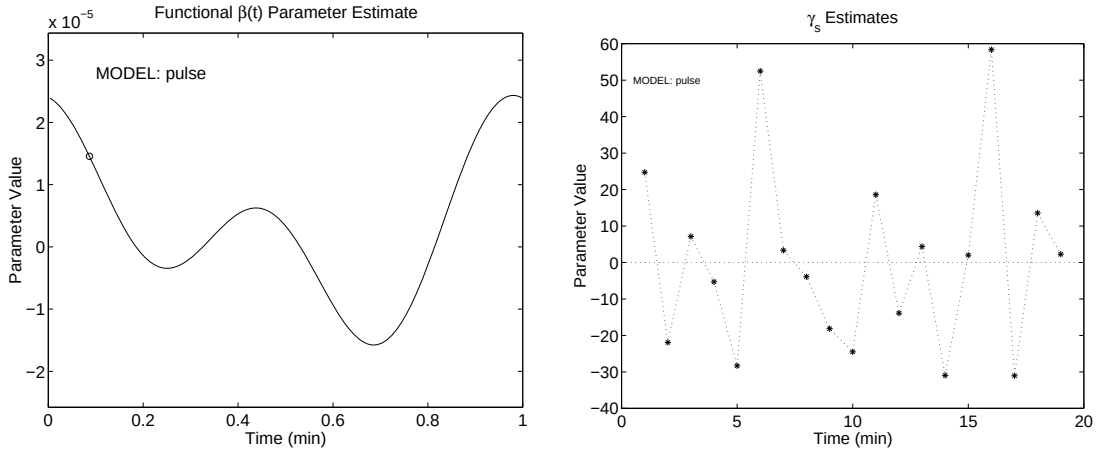


Figure 5.1: *Cross-Validation plot for basis size selection for Risk Category.*

5.3. Overall, 96.8% of the subjects were correctly classsified. Using the best classification cut-off of $\hat{\pi} = 0.6$, only one normal and one high risk birth were incorrectly classified.

The number of basis functions was found using cross-validation, see Figure 5.1. Since in this case we are using CV to maximise the likelihood, we wish to maximise the cross-validation score. The optimal number of basis functions is $m = 10$; however, from a clinical viewpoint, the functional parameter is difficult to interpret when 10 basis functions are used. Another local maximum occurs at $m = 5$: using this value, the results produced were similar (only one



(a) Functional time parameter estimate

(b) Time (min) parameter estimate

Figure 5.2: *Parameter estimates from Functional logistic regression for Risk Category. Within each minute, the stimulus occurs after 5 seconds, indicated by a circle on the time parameter estimate.*

subject was classified differently). The parameter was also more easily interpreted, in terms of seeing if the functional model would relate to the clinical idea of habituation. Another advantage of using the $m = 5$ model is that five fewer parameters needed to be estimated. Thus, five basis functions were used.

The estimated functional time parameter, $\hat{\beta}(t)$, and minute parameters, $\hat{\gamma}_s$, are shown in Figure 5.2. The timing of the stimulus within each minute is indicated by a circle on the time parameter estimate. The functional parameter appears to place more emphasis on the heart rate in the later stages of each minute. This ties in with the clinical idea of habituation. "Normal" fetuses should have stopped reacting to any stimulus well before the end of each minute.

The histogram of the predicted probabilities is given in Figure 5.3. This clearly shows a differentiation in the predicted probabilities between the normal and high risk groups.

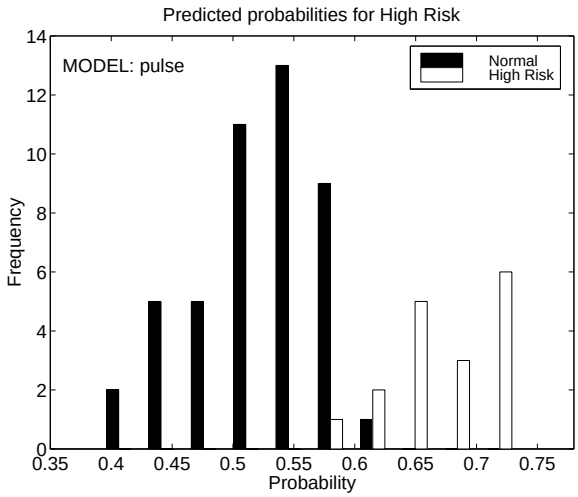


Figure 5.3: *Histogram of probabilities of a High Risk Birth split by observed response.*

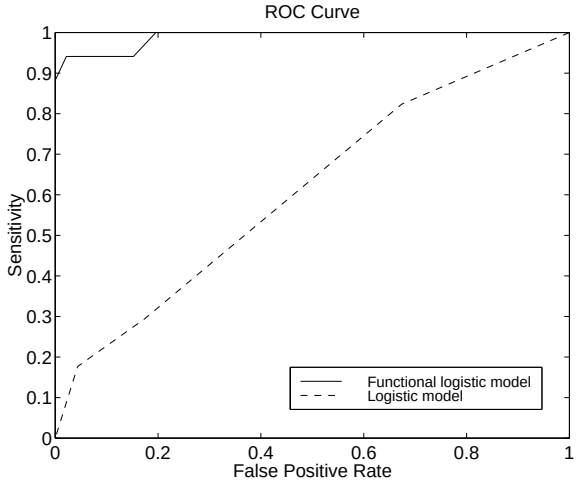


Figure 5.4: *ROC Curves for the functional and simple logistic regression models for Risk Category.*

Comparing the logistic and functional logistic models, we see that the functional model is significantly better than the simple logistic model. The functional model correctly detected 96.8% of the high risk pregnancies, while the logistic model only correctly detected 21.4%. The ROC curves (Figure 5.4) also show that the functional logistic model is superior to the logistic model. However, this improvement does come at the cost of more degrees of freedom being used in the model.

Classification Table for Risk Category MODEL: pulse			
	Observed Level		
Predicted	Normal	High Risk	Total
Normal	4	2	
High Risk	3	1	
% Correct	57.1%	33.3%	50.0%

Table 5.4: *Summary of Classifications of Risk Category for the Controls, cut-off probability* $\hat{\pi} = 0.6$.

5.3.1 Controls

The results of the stimulated heart rates were validated using the 10 unstimulated subjects as controls. Since the models were based on 19 minutes worth of heart rate measurements, the subjects with less than 19 minutes had the start of their measurements replicated at the end to produce a full 19 minutes worth of data. The predictors and heart rate measurements were then placed into the model (5.1), found using the stimulated heart rates, in order to predict the risk category for the controls.

Of the 10 controls, 7 of the fetuses were normal and 3 were high risk. Using the functional logistic model, the classification summary for the controls is given in Table 5.4. It appears that there is no relationship between the true and predicted outcomes. Thus, without the stimuli being applied, the heart rate does not give a useful prediction of the infant's risk category.

5.4 PDI at 18 Months

We now turn our attention to the continuous responses at 18 and 36 months of age. Since the techniques are the same for each of the possible responses, the modelling of only one of

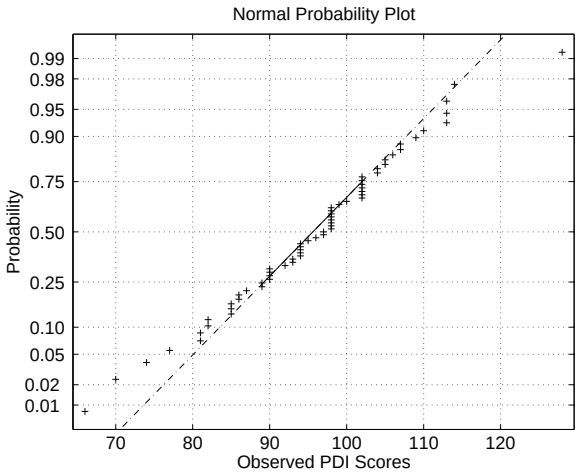


Figure 5.5: *Normal probability plot for the PDI scores at 18 months.*

the variables, PDI at 18 months, is given.

The Psychomotor Development Index (PDI) is used to measure the child's fine- and gross-motor development. The PDI scores were found to be approximately normally distributed (Figure 5.5) with a mean of 95.76 points and a standard deviation of 11.16. The highest score was 128 and the minimum was 66. We are modelling the raw PDI scores but, once found, the PDI scores are generally classified into 4 groups:

PDI Score	Class
0 - 69	1
70 - 84	2
85 - 114	3
115+	4

with class 4 being the highest (best) category.

A standard linear regression model, without the functional covariate, was fitted first. In the best linear regession model, the only significant covariate was found to be the child's sex (p=0.034). Girls have, on average, a 5.9 points higher PDI score than boys. The R^2 for this model was 7.10% (Table 5.5).

Regression Analysis for PDI, 18 Months				
Covariate	**Coef**	**StDev**	**T**	**P**
Constant	93.029	1.860	50.03	0.000
Sex	5.936	2.741	2.17	0.034
$R^2 = 7.10\%$				

Table 5.5: *Regression summary for PDI at 18 months, using only scalar covariates.*

Functional Regression Analysis for PDI, 18 Months					
Covariate		**Coef**	**StDev**	**T**	**P**
Constant		93.043	14.917	6.24	0.000
Sex		6.035	2.111	2.86	0.007
ANOVA					
Source	**df**	**SS**	**MS**	**F**	**P**
Scalar Cov.	1	551.5	551.5	8.23	0.007
Pulse	25	4760.0	190.4	2.84	0.002
(γ)	(19)	(3565.8)	(187.7)		
($\beta(t)$)	(6)	(1194.2)	(199.0)		
Residual	36	2411.9	67.0		
Total	62	7723.4			
$R^2 = 68.77\%$					

Table 5.6: *Functional regression summary for PDI at 18 months.*

Using functional regression, the best model (Table 5.6) for PDI at 18 months had an R^2 of 68.8%; a substantial improvement over the simple linear regression model. The child's sex was still significant (p=0.007), with girls having an average increase of 6.0 points over boys. The functional heart rate covariate was also found to be highly significant (p=0.002).

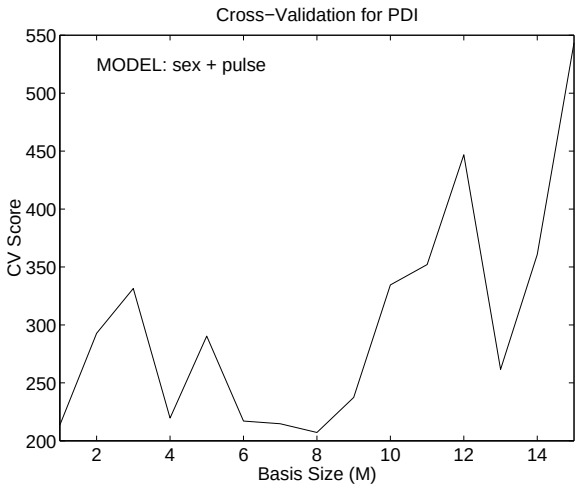


Figure 5.6: *Cross-Validation plot for basis size selection for PDI, 18 months.*

Cross-validation of the RSS (Figure 5.6) was used to estimate the number of basis functions needed to model both the raw heart rate data and $\beta(t)$. The absolute minimum was found to be $m = 8$; however, six basis functions were used as $m = 6$ produced similar results to $m = 8$ but had a more interpretable functional time parameter and the CV values at $m = 6$ and $m = 8$ are close.

The estimated functional time parameter, $\hat{\beta}(t)$, is shown in Figure 5.7. In Figures 5.7 (a) - (d), four sample heart rates from within one minute have been shown and their integral, $\int x(t)\hat{\beta}(t)dt$, calculated using this estimated parameter, is given. The integral value is used to show the relationship between the heart rate within *one* minute, ignoring the stimulus effect, and the PDI score. Assuming the infants are of the same sex and the subjects have the same heart rate pattern for all of the 19 minutes, then a subject with the pattern in Figure 5.7(a) would have a higher PDI score than one with the pattern in Figure (d), with Figures (b) and (c) somewhere in between. Thus, subjects with a higher PDI score are more likely to have a heart rate that does react to the stimulus but this reaction stops fairly quickly. However, in practice, it is hard to determine by inspection which subject will have a higher PDI score, as a subject will have a combination of these, and other, patterns over the 19 minutes and the minute effect also needs to be taken into account.

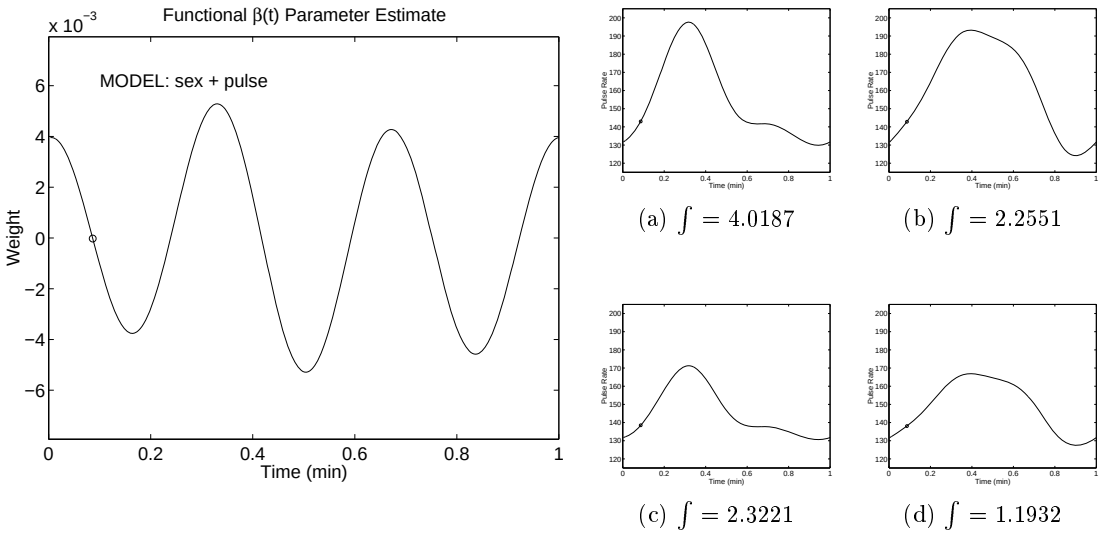


Figure 5.7: *Functional time parameter for PDI, 18 months, and its effect on sample heart rates. ($\int = \int x(t)\hat{\beta}(t)dt$ - using smoothed $x(t)$). The time at which the stimulus occurred is indicated by a circle/dot, near the start of each function.*

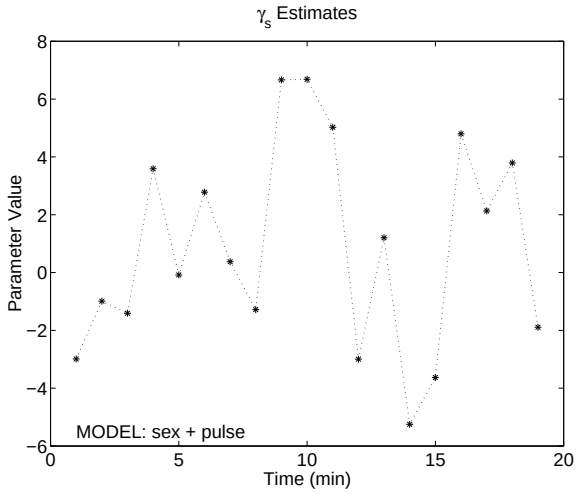
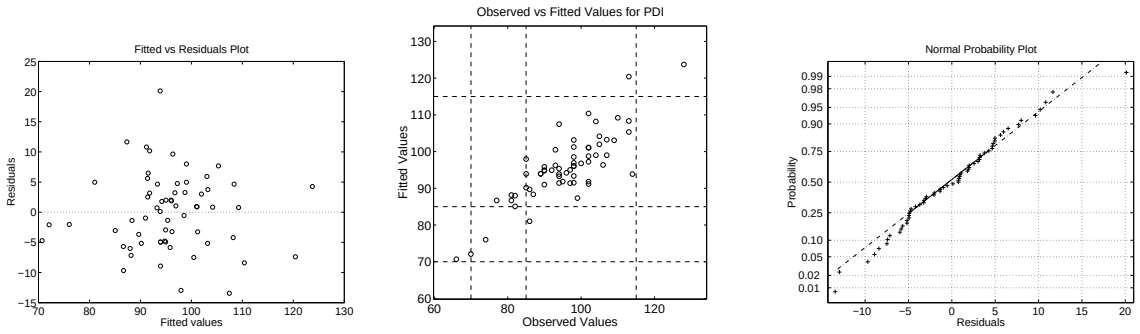


Figure 5.8: *Time parameter estimates, $\hat{\gamma}_s$, for PDI, 18 months.*

The estimated stimuli parameters, $\hat{\gamma}_s$, produced a weighted average over the 19 minutes of the heart rate (Figure 5.8). Ignoring the functional time effect, within each minute, the most important minutes are the 9^{th}, 10^{th}, 11^{th}, and 16^{th} in the the positive direction, and the 14^{th} and 15^{th} in the negative direction. The negative values for $\hat{\gamma}_s$ mean that subjects who have stopped reacting to the stimulus in these minutes will have higher PDI scores.



(a) Fitted vs Residuals

(b) Observed vs Fitted values

(c) Residual Normal Plot

Figure 5.9: *Functional regression model diagnostics for PDI, 18 months.*

The diagnostics for this model are given in Figure 5.9. The fitted values and residuals are plotted in Figure 5.9(a). There does not appear to be any pattern in the plot. Figure 5.9(b) gives the observed versus fitted values. The dashed lines indicate the class subdivisions for PDI. Only seven of the subjects have been incorrectly classified under this model. However, these subjects, in particular those in the lower classes, are close to their true classes. Note, there is only a small number of subjects not in class 3. This may be the reason why the fitted values for the bottom two classes are higher than the observed values. The assumption of normally distributed errors was checked via a normal probability plot of the residuals, Figure 5.9(c). The residuals appear to be normally distributed, with one large residual corresponding to the maximum PDI score of 128.

5.5 Discussion

The stimulated fetal heart rates were found to be a good predictor of the infant's risk category and psychomotor development at 18 months. For both responses, the functional models represented a substantial improvement over the standard linear/logistic models. These models also have the advantage that no notion of habituation needed to be incorporated into the methodology, and hence there were no subjective definitions.

The functional model correctly predicted 94.1% of the high risk pregnancies, compared to 17.6% for the best standard logistic regression model. However, it should be borne in mind that these percentages are optimistic, as the parameter estimates were based on the same data on which predictions were made.

The importance of the stimulus was established through the use of control subjects. It was found that the stimuli were necessary for the heart rates to be predictive. Without the repeated stimlus, the heart rates were no more predictive of the infant's risk category at birth than chance.

Similarly, it was shown that the changes in fetal heart rate after stimulation, as well as the infant's sex, are important predictors of a child's psychomotor development at 18 months. As with risk category, the predictions are optimistic as the observed PDI scores were used to predict themselves. Jackknifing could be used to overcome this problem. However, for the fetal heart rate data this method does not produce accurate predictions for the extreme classes. Of the PDI scores, 87% of the data were in class 3. Thus, when a score from class 1, 2 or 4 was omitted, there were not enough data remaining in the class to provide adequate information for the prediction; the predictions were drawn towards class 3 values. This was especially true for classes 1 and 4, with predictions for the omitted scores being made outside the range of the data generating the parameter estimates.

The results for PDI showed there was a significant gender effect with females outperforming males. This has important ramifications for future developmental studies. It suggests that all developmental data needs to be analysed by gender to ensure that differences in outcomes are not concealed by gender.

A disadvantage of the functional models is that they use substantially more degrees of freedom compared to the standard linear and logistic models. Thus, the more complex the functional covariate structure, the more subjects are needed in the study.

Also of interest was that parental influences had no effect on the outcomes. However, the PDI scores at 36 months were also modelled using the functional techniques. At this age, HOME (as well as sex) had become a significant predictor (p=0.006) of PDI; the more parental involvement and emotional support provided, the higher the PDI score. Thus, while parental influences have no significant effect at 18 months, they are important in a child's development at 36 months of age.

The study found that smoking, drinking and caffeine intake had no effect on the outcomes measured. However in the sample population, these rates were particularly low. It may be that a higher preponderance of these factors in a sample would lead to a better assessment of their impact on the newborn infant's health and development.

The findings from this study may have important clinical implications. Using the functional model after stimulation as a clinical test may help alert the obstetrician to a high risk birth, and to determine the optimal time to deliver a fetus in a complicated pregnancy. It may also have a useful role in alerting parents about possible developmental delay, so that early intervention could be undertaken.

Chapter 6

Functional Mean and Covariance Modelling

In this chapter we examine the problem of joint mean and covariance modelling. Unlike functional principal component analysis or growth curve modelling, the mean function is not considered separately to the covariance function; the two are estimated together. This chapter presents a basis solution to the joint modelling, as well as discussing properties of the resulting algorithm.

6.1 Previous Approaches

Suppose n subjects have had measurements of a variable y recorded at a number of time points. Each subject can have a different number of measurements taken at different times. Thus, $y_{il} = y_i(t_{il})$ is the measurement on the i-th subject taken at the l-th time for subject i, t_{il}, $i = 1 \ldots n$, $l = 1 \ldots p_i$. Estimates are now needed for the mean function, $\mu(t)$, and for the covariance function, $\Gamma(s,t)$, of $y_i(t)$.

Although there is a considerable literature on FDA (see references in Ramsay and Silverman, 1997) there is not much work on joint modelling of mean and covariance. Hart and Wehrly (1986) used a kernel method to estimate the mean but the covariance structure was assumed to be stationary and was estimated from the averages of the time series.

A kernel method was also used by Staniswalis and Lee (1998) to estimate the mean and co-

variance functions of longitudinal data. The mean was estimated using scatterplot smoothers while the random effects were modelled using a basis expansion, with the normalised eigenfunctions of the random effects covariance function used as the bases. A boundary corrected kernel at each time point was used to estimate both the basis coefficients and the covariance function of the data. Their method assumes a small number of measurements on each subject at different times and that the times at which these measurements occur provide a dense coverage of the time interval as n increases. For large data sets with many measurements per subject this approach will be computationally prohibitive.

Fan and Zhang (1998) developed a two step procedure for modelling the mean and covariance functions. The first step involves finding raw estimates of the mean parameters at each time point via least squares on all those subjects with measurements taken at the same time point. The covariance function is modelled using the residual and hat matrix from these least squares fits. These raw parameter estimates are smoothed in the second step. Fan and Zhang used local polynomial smoothing but splines, kernel or local linear regression could have been used. Given there are T distinct time points across the subjects, Fan and Zhang's method overcomes the problem of inverting a matrix of approximate size $(nT) \times (nT)$ found in other methods (eg. Brumback and Rice (1998) who modelled the curves for each subject using cubic smoothing splines, assuming a particular form for the covariance kernel. Estimates were found using mixed-effects methodology.). However, their smoothing step still requires the inversion of a $T \times T$ matrix. For functional data, where T is large, this inversion will be computationally intensive.

Anderson and Jones (1995) used polynomials to model the mean function but the random effect for each subject was approximated via a smoothing spline. They used a state-space representation of the model to generate the estimates for the parameters. The mean parameters were estimated using maximum likelihood. Since the random effect smoothing splines plus the measurement error could be thought of as an integrated random walk with error, the likelihood for the covariance functions was maximised recursively using a Kalman filter. Once estimated, Anderson and Jones produce the splines using a smoothing algorithm.

Wang (1998) reversed this and used smoothing splines to model the mean while the random effects were modelled parametrically via mixed effects methodology. He assumed that the covariance matrices for both the random effects and measurement error depended on a parsimonious set of parameters. The mean function was then estimated using penalised maximum likelihood. The smoothing parameter for the mean was estimated with the random effect parameters via generalised maximum likelihood.

Rice and Wu (1999) presented a mixed effects solution to the modelling of the mean and covariance functions for functional data. Random effects terms were used to allow the subjects to vary about the mean curve. Their technique extends classical linear mixed effects methods by modelling both the mean and random effects nonparametrically using spline basis functions. Uncorrelated measurement errors $\epsilon_{il} \sim iid(0, \sigma^2)$ were also included in the model

$$y_{il} = \sum_{1}^{m} \beta_r \bar{B}_r(t_{il}) + \sum_{1}^{q} \gamma_r^i B_r(t_{il}) + \epsilon_{il}$$

where the $\gamma_i = \{\gamma_r^i\}$ are assumed to be random coefficients with mean 0 and covariance Γ, and $\bar{B}_r(t)$ and $B_r(t)$ are the mean and random effect B-spline basis functions, respectively. The covariance function of the data is thus

$$V_{s,t} = \text{cov}(y_i(s), y_j(t)) = \left(\sum_{1}^{q} \sum_{1}^{q} \Gamma_{ru} B_r(s) B_u(t) + \sigma^2 \delta_{s,t} \right) \delta_{i,j}$$

where δ is the Kronecker delta function.

Rice and Wu then proceed to estimate the model parameters using standard mixed effects methodology. The EM algorithm (Dempster *et al.*, 1977) is used to estimate the β_r's, σ^2 and Γ, while the γ_r^i's are found using the BLUP estimator (Robinson, 1991). As noted by Rice and Wu, the estimate for each individual curve will be shrunk towards the population mean, and there is no simple connection between Γ and the eigenanalysis of V. However, the first term of V can be approximated using the estimated $\hat{\Gamma}$ and eigenanalysis can be performed on the resulting matrix.

James and Hastie (1999) used a reduced rank form of a mixed effects model to estimate

the data. Their concern is with data different to standard functional data in that only a small number of measurements are available from the curve of each subject; instead of the usual large number of measurements. Both the mean and random effects are represented using spline basis functions. However, instead of estimating the entire covariance matrix, and its principal components, they restrict the rank of the covariance function and estimate the principal components directly. The only difference between their reduced rank model and standard mixed effects is in the fitting procedure.

All these methods use either basis expansions or smoothing splines. The basis expansion technique is quite flexible but it does have the weakness that the regularising parameters $1/m$, $1/q$ can only take discrete values. This can be overcome by using a joint basis-penalty method as in Ramsay and Silverman (1997). However, that technique does not produce orthogonal eigenfunctions. We develop a basis method that is computationally much simplier than Ramsay and Silverman (1997) and preserves eigenfunction orthogonality. Our procedure bears some relation to that of James and Hastie (1999), however our work is independent.[1]

The remainder of the chapter is organised as follows. In Section 6.2 we present the basis algorithm with an intuitive approach that views it as a natural extension of functional PCA. We examine this algorithm, and its convergence, in more detail in Section 6.3. Section 6.4 contains a simulation study while real data are analysed in Section 6.5. Conclusions are offered in section 6.6. In the following, we assume equispaced sampling at times $t_l = l\Delta$, $l = 1 \dots p$.

6.2 The Basis Method

For the ith subject, we represent the time series as a mean plus a noise,

$$y_i(t) = \mu(t) + \nu_i(t)$$

[1] Our algorithm was presented at the 1999 Joint Statistical Meetings (Ratcliffe and Solo, 1999) where we also became aware of the work of James and Hastie.

where $\mu(t)$ is the mean at time t and the noise, $\nu_i(t)$, consists of random effects and a residual. Both the mean and the noise are modelled nonparametrically by basis expansions but the basis elements need not be the same. Thus

$$\mu(t) = \sum_1^m \chi_k(t)\beta_k = x^T(t)\beta \tag{6.1}$$

$$\nu_i(t) = \sum_1^q \zeta_k(t)b_k^i = z^T(t)b_i \tag{6.2}$$

and m, $q \ll n$. By collecting the various time series into vectors

$$Y_i = [y_i(t_1) \ \ldots \ y_i(t_p)]^T$$

$$X_{(p\times m)} = [x(t_1) \ \ldots \ x(t_p)]^T$$

$$Z_{(p\times q)} = [z(t_1) \ \ldots \ z(t_p)]^T$$

we can express the time series for the ith subject as

$$Y_i = X\beta + Zb_i$$

We write the basis expansions and vectors in a different form from the previous chapters for ease of use in future calculations. Also, in this form the expression for Y_i resembles standard mixed effects notation.

We now estimate the model by an iterative algorithm which we call cyclic estimating equations (c.f. cyclic descent). The two steps in this algorithm are:

R-STEP - getting the random effects

> Given β do a functional principal component analysis (Ramsay and Silverman (1997)) (fPCA) to get the random effects.

F-STEP - getting the fixed effects

> Given the random effects estimate β by least squares.

We begin by describing the R-step. Suppose a value β_0 is given. We calculate the random coefficients b_i by

$$\begin{aligned} b_i &= \left\langle z, z^T \right\rangle^{-1} \left\langle z, y_i - x^T\beta_0 \right\rangle \\ &= \left(Z^TZ\right)^{-1} Z^T\left(Y_i - X\beta_0\right) \end{aligned} \tag{6.3}$$

The sample covariance matrix for Y is defined as

$$\Gamma_n(t,s) = n^{-1}\sum_1^n \nu_i(t)\nu_i(s)$$

Substituting the basis expansions for $\nu_i(t)$ gives

$$\Gamma_n(t,s) = z^T(t)S_b z(s) \tag{6.4}$$

with $S_b = n^{-1}\sum_1^n b_i b_i^T$. The fPCA of this sample covariance matrix generates eigenfunctions $\phi_k(t)$ and corresponding eigenvalues λ_k (ordered from the largest down) satisfying the eigenequations

$$\int \Gamma_n(t,s)\phi_k(s)ds = \lambda_k\phi_k(t)$$
$$\int \phi_k(t)\phi_u(t)dt = \langle\phi_k,\phi_u\rangle = \delta_{k,u}$$

Substituting (6.4) and a corresponding basis expansion for the eigenfunctions,

$$\phi_k(t) = z^T(t)f_k$$

into the eigenequations gives

$$z^T(t)\int S_b z(s)z^T(s)f_k = \lambda_k z^T(t)f_k$$
$$\int f_k^T z(t)z^T(t)f_u = \delta_{k,u}$$

This is equivalent to a weighted matrix PCA

$$S_b W f_k = \lambda_k f_k, \qquad k = 1\ldots q \tag{6.5}$$

$$f_k^T W f_u = \delta_{k,u} \tag{6.6}$$

$$W = \int z(s)z^T(s)ds = Z^T Z\Delta$$

Having found f_k (and hence $\phi_k(t)$) from (6.5), (6.6), we estimate the random effect component

of the noise from the dominant eigenfunctions corresponding to the r ($< q$) largest eigenvalues.

$$\begin{aligned}
\hat{\nu}_i(t) &= \sum_1^r \phi_k(t) \langle \phi_k, \nu_i \rangle \\
&= \phi^T(t) \langle \phi, \nu_i \rangle \\
\phi^T(t) &= [\phi_1(t) \ \ldots \ \phi_r(t)] \\
\Rightarrow \ \hat{\nu}_i(t) &= \phi^T(t) \left\langle \phi, y_i - x^T \beta_0 \right\rangle \\
\Rightarrow \ \hat{N}_i &= \begin{bmatrix} \hat{\nu}_i(t_1) \\ \vdots \\ \hat{\nu}_i(t_p) \end{bmatrix} \\
&= \Phi \Phi^T (Y_i - X\beta_0) \Delta \qquad (6.7) \\
\Phi &= \begin{bmatrix} \phi^T(t_1) \\ \vdots \\ \phi^T(t_p) \end{bmatrix}
\end{aligned}$$

Substituting the basis expansion gives

$$\begin{aligned}
\phi^T(t) &= z^T(t) F \qquad (6.8) \\
F &= [f_1 \ \ldots \ f_r] \\
\Rightarrow \ \Phi &= ZF
\end{aligned}$$

so that finally (6.7) becomes

$$\Rightarrow \ \hat{N}_i = ZFF^T Z^T (Y_i - X\beta_0) \Delta \qquad (6.9)$$

So given β_0, we get b_i (and hence S_b) from (6.3), then calculate f_k (and hence $\phi_k(t)$) from (6.5), (6.6), and finally estimate the random effects $\hat{N}_i$ from (6.9).

Turning to the F-step, given F, and hence $\phi_k(t)$ and $\hat{\nu}_i(t)$, we get β_1 from the least squares

problem $\beta_1 = \arg\min J(\beta)$

$$\begin{aligned} J(\beta) &= n^{-1} \sum \|y_i - \hat{\nu}_i - x^T\beta\|^2 \\ \Rightarrow\ \beta_1 &= \left\langle x, x^T \right\rangle^{-1} \left\langle x^T, \bar{Y} - \hat{\bar{\nu}} \right\rangle \\ &= \left(X^T X\right)^{-1} X^T \left(\bar{Y} - \hat{\bar{N}}\right) \\ &= \left(X^T X\right)^{-1} X^T \left(\bar{Y} - \Delta Z F F^T Z^T \left(\bar{Y} - X\beta_0\right)\right) \end{aligned} \tag{6.10}$$

where $\bar{Y} = n^{-1}\sum_1^n Y_i$ and $\hat{\bar{\nu}} = n^{-1}\sum_1^n \hat{\nu}_i$.

Thus the algorithm iterates between the R-step (6.3), (6.5), (6.6), (6.9) and the F-step (6.10) (which uses only r of the eigenvectors found in (6.5), (6.6)).

Actually, some computational simplifications can be made to speed up the calculations. For example, S_b, which is used in (6.5), can be found at each iteration by a rank one modification to $\bar{S}$:

$$\begin{aligned} \bar{S}_b = Z^T Z S_b Z^T Z &= \bar{S} + Z^T \left(\bar{Y} - X\beta_0\right)\left(\bar{Y} - X\beta_0\right)^T Z \\ \bar{S} &= \frac{1}{n}\sum_1^n \left(Z^T Y_i - Z^T\bar{Y}\right)\left(Z^T Y_i - Z^T\bar{Y}\right)^T \end{aligned} \tag{6.11}$$

So (6.5) can be rewritten as

$$\bar{S}_b f_k = \frac{\lambda_k}{\Delta^2} W f_k$$

Also, using $\hat{\beta}_0 = \left(X^T X\right)^{-1} X^T \bar{Y}$ and $R_{xz} = \left(X^T X\right)^{-1} X^T Z$, we can rewrite (6.10) as

$$\beta_1 = \hat{\beta}_0 + \Delta R_{xz} F F^T Z^T \left(\bar{Y} - X\beta_0\right)$$

where $\hat{\beta}_0$, R_{xz}, $Z^T\bar{Y}$, and $Z^T X$ can be precomputed. This could also be written as

$$\begin{aligned} \beta_1 &= \left(X^T X\right)^{-1} X^T U \left(\bar{Y} - X\beta_0\right) + \beta_0 \\ U &= I_{(p\times p)} - \Delta Z F F^T Z^T \end{aligned} \tag{6.12}$$

Thus, we see that if the algorithm converges, the converged values of F and β satisfy

$$\begin{aligned} \bar{S}_b f_k &= \frac{\lambda_k}{\Delta^2} W f_k \\ \bar{S}_b &= \bar{S} + \left(Z^T\bar{Y} - Z^T X\beta\right)\left(Z^T\bar{Y} - Z^T X\beta\right)^T \end{aligned} \tag{6.13}$$

and

$$X^T U\left(\bar{Y} - X\beta\right) = 0$$
$$\Rightarrow\ X^T U X\beta = X^T U\bar{Y} \tag{6.14}$$

6.3 Algorithm Analysis

We now examine the algorithm in more detail.[2] While (6.13) and (6.14) give the converged values, the algorithm does not result in a unique solution for β. This was established by a perturbation expansion of the log-likelihood $\mathcal{L}$ for the model (further details can be found in Solo and Ratcliffe, 2000).

$$\begin{aligned} Y_l^i &= \mu(t_l) + \sum_1^k \phi_u(t_l) g_u^i + \varepsilon_l^i \\ &= \mu(t_l) + \phi^T(t_l) g_i + \varepsilon_l^i \\ \Rightarrow\ \mathcal{L} &= -\frac{1}{2}\operatorname{tr}\left(S_e\Sigma^{-1}\right) - \frac{1}{2}\ln|\Sigma| \\ &= -\frac{1}{2\sigma^2}\operatorname{tr}(S_e) + \frac{1}{2\sigma^2}\operatorname{tr}\left(\Phi^T S_e \Phi\Delta\right) - \frac{1}{2}(p-k)\ln\sigma^2 \\ &\qquad - \frac{1}{2}\ln|\frac{\Lambda}{\Delta}| - \frac{1}{2}\operatorname{tr}\left(S_e\Phi\Lambda^{-1}\Phi^T\right) \end{aligned}$$

where Y_l^i, $l = 1 \ldots p$ is the time series observed on the i-th individual; g_i are the random effects of variance Λ; ε_l^i is a white noise sequence of variance σ^2/Δ; $e_i = Y_i - \mu$ is the residual; and $S_e = n^{-1}\sum_1^n e_i e_i^T$.

Investigation of the likelihood as $\sigma^2 \to 0$ revealed that the algorithm provides "small noise" maximum likelihood equations. That is, under the small noise condition

$$\sigma^2/\lambda_u \ll 1, \qquad u = 1 \ldots k$$

the maximum likelihhod estimates of F and β are the PCA (6.13), (6.6), and the least squares computation (6.14), as well as (6.11). Further, it was shown that the likelihood equations could be solved to give a unique solution for F but that the value for β is not unique. Thus,

[2] This work is joint with Victor Solo.

a suitable initial value must be chosen for β. To throw some light on this we explore the algorithm in more detail.

Let $\tilde{X} = X\left(X^TX\right)^{-1/2}$ and $\theta = \left(X^TX\right)^{1/2}\beta$, then we can write the iteration (6.12) as (where k is the iteration index)

$$\begin{aligned}
\theta_k &= \tilde{X}^TU\bar{Y} - \tilde{X}^TU\tilde{X}\theta_{k-1} + \theta_{k-1} \\
&= \tilde{X}^TU\bar{Y} + \left(I - \tilde{X}^TU\tilde{X}\right)\theta_{k-1} \\
\Rightarrow\ \theta_k &= \omega + BB^T\theta_{k-1} \\
\omega &= \tilde{X}^TU\bar{Y} \\
B_{(m\times r)} &= \tilde{X}^TZF\sqrt{\Delta}
\end{aligned}$$

Iterating this gives

$$\theta_k = (BB^T)^k\theta_0 + \sum_1^{k-1}(BB^T)^i\omega + \omega \tag{6.15}$$

We now proceed by examining the decomposition of BB^T. We claim all the eigenvalues of this matrix will be ≤ 1. Since the eigenvalues of BB^T are the same as for B^TB, we prove this by showing that B^TB is positive semidefinite with eigenvalues ≤ 1.

Proof: Let $Q_{p\times(p-m)}$ be a matrix of rank $p-m$ such that $Q^TX = 0$, then the projection theorem gives that, with $P_X = X\left(X^TX\right)^{-1}X^T$, $P_Q = Q\left(Q^TQ\right)^{-1}Q^T$,

$$\begin{aligned}
I_p &= P_X + P_Q \\
\Rightarrow\ Z^TZ &= Z^TP_XZ + Z^TP_QZ \\
\Rightarrow\ \Delta F^TZ^TZF &= I_r \\
&= \Delta F^TZ^TP_XZF + \Delta F^TZ^TP_QZF \\
&= B^TB + \Delta F^TZ^TP_QZF
\end{aligned}$$

Thus, B^TB is positive semidefinite with eigenvalues ≤ 1. □

We now partition the eigenvectors of BB^T into three groups:

q_u with eigenvalues $0 < \lambda_u < 1$;

$\hat{q}_u$ with unit eigenvalues;

$\bar{q}_u$ with 0 eigenvalue.

Since these eigenvectors are orthogonal, we can write

$$\begin{aligned} I_m &= \sum q_u q_u^T + \sum \hat{q}_u \hat{q}_u^T + \sum \bar{q}_u \bar{q}_u^T \\ BB^T &= \sum \lambda_u q_u q_u^T + \sum \hat{q}_u \hat{q}_u^T \\ \Rightarrow\ (BB^T)^i &= \sum \lambda_u^i q_u q_u^T + \sum \hat{q}_u \hat{q}_u^T \\ \Rightarrow\ (BB^T)^n \theta_0 &\to \sum \hat{q}_u \hat{q}_u^T \theta_0, \qquad \text{as } n \to \infty \end{aligned} \tag{6.16}$$

Also

$$\begin{aligned} \sum_1^n (BB^T)^i \omega &= \sum q_u q_u^T \omega \frac{\lambda_u - \lambda_u^n}{1 - \lambda_u} \\ &\to \sum q_u q_u^T \omega \frac{\lambda_u}{1 - \lambda_u}, \qquad \text{as } n \to \infty \end{aligned}$$

Using $\hat{q}_u^T \omega = \hat{q}_u^T (I - BB^T)\beta = 0$, we find that $\omega = \sum q_u q_u^T \omega + \sum \bar{q}_u \bar{q}_u^T \omega$. Hence, returning to (6.15), we find that

$$\theta_k \to \theta_\infty = \sum \hat{q}_u \hat{q}_u^T \theta_0 + \sum \frac{q_u q_u^T \omega}{1 - \lambda_u} + \sum \bar{q}_u \bar{q}_u^T \omega$$

Now, note that the minimum norm least squares solution of the iteration (6.15) is given by $\theta_+ = R_+ \omega$ where R_+ is the (unique) Moore-Penrose generalised inverse of $I - BB^T$.

$$\begin{aligned} I - BB^T &= \sum (1 - \lambda_u) q_u q_u^T + \sum \bar{q}_u \bar{q}_u^T, \qquad \text{from (6.16)} \\ \Rightarrow\ R_+ &= \sum \frac{q_u q_u^T}{1 - \lambda_u} + \sum \bar{q}_u \bar{q}_u^T \\ \Rightarrow\ \theta_+ &= \sum \frac{q_u q_u^T \omega}{1 - \lambda_u} + \sum \bar{q}_u \bar{q}_u^T \omega \end{aligned}$$

So we see that if $\theta_0 = 0$ or $\theta_0 = \omega$, then $\theta_\infty = \theta_+$. That is, the iteration converges to the Moore-Penrose solution if $\theta_0 = 0$ or ω, otherwise the uniqueness of the solution is not guaranteed.

Of further interest is the special case of $X = Z$. Then

$$\begin{aligned} B &= \tilde{Z}^T Z F \sqrt{\Delta} \\ &= \left(Z^T Z \Delta\right)^{1/2} F = \tilde{F} \end{aligned}$$

and $B^T B = \tilde{F}^T \tilde{F} = I$. The projection lemma can then be used to show that all the eigenvalues of $B^T B$ are equal to 1.

Proof: Let $\tilde{G}_{q\times(q-r)}$ be a rank $(q-r)$ orthogonal matrix orthogonal to $\tilde{F}$ then by the projection lemma

$$\begin{aligned} I_q &= \tilde{F}\tilde{F}^T + \tilde{G}\tilde{G}^T \\ &= BB^T + \tilde{G}\tilde{G}^T \end{aligned}$$

Hence all the eigenvalues of $B^T B$ are 1. □

So $q_u = 0$ while the eigenvectors $\hat{q}_u = \tilde{f}_u = u$-th column of $\tilde{F}$ since

$$BB^T \tilde{f}_u = \tilde{F}\tilde{F}^T \tilde{f}_u = \tilde{F} e_u = \tilde{f}_u$$

Similarly, $\bar{q}_u = \tilde{g}_u = u$-th column of $\tilde{G}$. Thus

$$\begin{aligned} \theta_+ &= \sum \bar{q}_u \bar{q}_u^T \omega \\ &= \sum \bar{q}_u \bar{q}_u^T \omega_0 \\ &= \tilde{G}\tilde{G}^T \omega_0 \\ \omega_0 &= \tilde{X}^T \bar{Y} \\ \theta_\infty &= \tilde{F}\tilde{F}^T \theta_0 + \tilde{G}\tilde{G}^T \omega \end{aligned}$$

So if $\theta_0 = \omega_0$ we find that $\theta_\infty = \omega_0$. Hence, in this case the iteration converges in one step.

6.4 Simulation Study

We examined the results of our fitting algorithm using a simulation study. Twenty curves were simulated with measurements taken at 100 equispaced time points on an arbitary time scale. The data was generated using a known mean or trend function plus AR(2) noise time

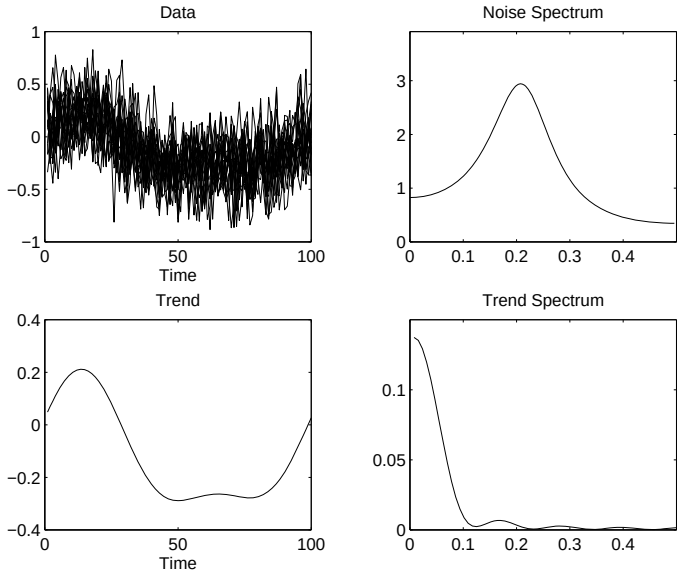


Figure 6.1: *Top left: simulated data using a known mean or trend function and AR(2) noise time series. Top right: theoretical spectrum for the simulated data. Bottom left: true mean function generated using 5 Fourier basis functions. Bottom right: line spectrum for the trend.*

series. The mean or trend function was a linear combination of five Fourier basis functions with randomly generated coefficients. The AR(2) noise time series used for the random effects and error was generated using parameter values of 0.3 and -0.4. The variance of the error was $\sigma^2 = 0.05$. The true eigenvalues of the covariance function can be estimated from the spectrum of the time series (Brillinger (1980)). Figure 6.1 shows the true mean function, and its line spectrum, the theoretical spectrum of the time series (eigenvalues) and the data generated using these functions.

The basis algorithm given in Section 3 was used to estimate the mean and covariance functions of the simulated data. Fourier basis functions were used for the mean, while orthogonal polynomial basis functions were used to estimate the random effects and eigenfunctions.

Cross-validation (Figure 6.2) was used to determine the number of basis functions for both the mean and covariance. The absolute minimum occurred at $m = 20$ (mean basis size) and $q = 4$ (random effect basis size). However, the cross-validation scores dropped at $m = 3, 5, 12,$ and 17 basis functions. In fact, the cross-validation scores are fairly close using anywhere

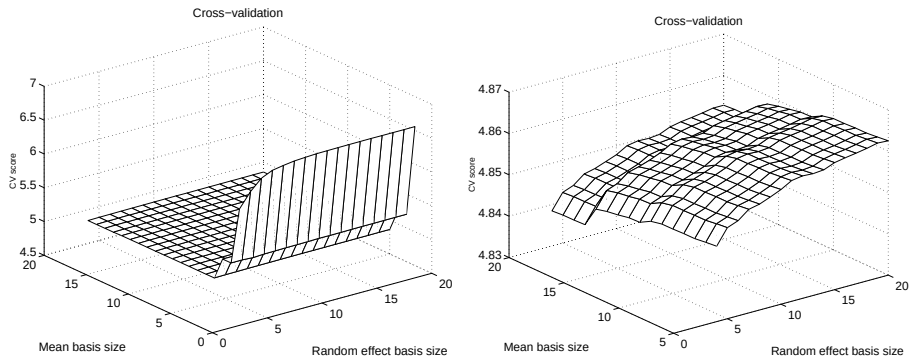


Figure 6.2: *Cross-validation plot for the simulated data.*

from 5 to 20 mean basis functions, and the estimated mean function is relatively the same using any of these values. Using $m = 3$ basis functions, the small change around time 65 is missed. Thus, $m = 5$ and $q = 4$ will be used to model the simulated data.

In Figure 6.3 we show the estimated mean, eigenvalue ratios ($\theta_u = \lambda_u / \sum \lambda_k$), and first two eigenfunctions using $m = 5$, $q = 4$, and summing over $r = 3$ eigenfunctions in the algorithm. The estimated mean function is approximately equal to the true function and the estimated eigenratio resemble the spectrum values. The first eigenfunction accounts for approximately 35% of the variation in the data, while the second and third eigenfunctions account for approximately 24% and 23% respectively. Using four eigenfunctions, the theoretical eigenratios for the first three eigenfunctions would be 32.5%, 29.3% and 22.3%, respectively. The estimated eigenratios are approximately equal to the true values. Eigenfunction 1 approximately represents a random intercept, in that it has the effect of adding a constant to the mean. However, this constant changes from an increase to a decrease around the 75th time point. Eigenfunction 2 represents an increase/decrease in the amplitude of the curves, with a slight time shift in the occurance of the peaks/troughs.

The above results were calculated using $\beta_0 = \omega$, resulting in the unique Moore-Penrose solution. We studied the effect of this start value by examining the results found using different β_0 values. These values included $\beta_0 = 0$, a constant equal to the overall mean of the data, a constant equal to the absolute maximum/minimum value of all the data, and vectors

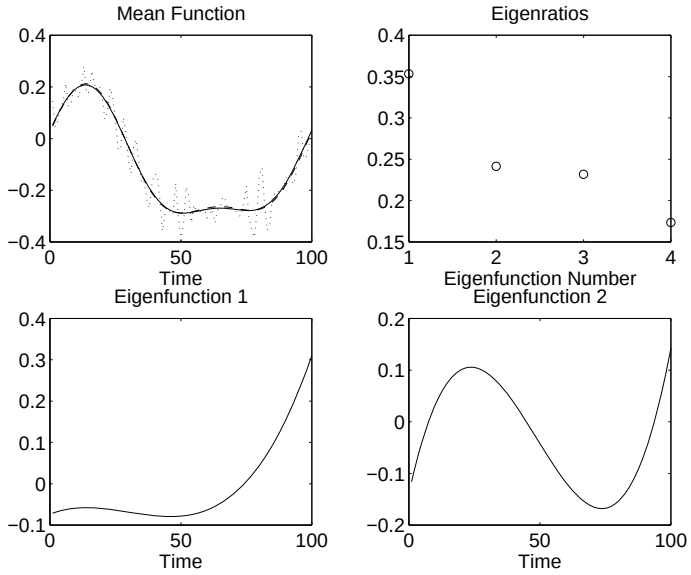


Figure 6.3: *Top left: estimated and true mean functions for the simulated data; (–) mean function estimated using 5 Fourier basis functions, (_) true mean function, (..) estimated mean function using the average of the data at each time point. Top right: estimated eigenratios. Bottom left: estimate of the first eigenfunction. Bottom right: estimate of the second eigenfunction.*

with randomly generated elements. While the number of iterations to convergence varied, we found that the converged values were actually the same for all the different β_0 values tested. The fastest convergence was achieved with $\beta_0 = \omega$ (in both this simulation and other data sets tested). Thus, we favour $\beta_0 = \omega$ in practice as it results in the Moore-Penrose solution with the fastest convergence time.

6.5 Examples

In this section we examine the results from two examples; the electroencephalographic (EEG) recordings described in Section 2.5, and the human gait data from Rice and Silverman (1991). In the following examples we start the iteration with an initial value of $\beta_0 = \omega$.

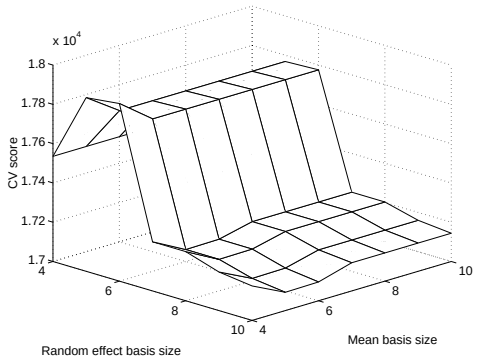


Figure 6.4: *Cross-validation plot for the EEG recordings.*

6.5.1 EEG Data

We return to the EEG data from Section 2.5. We are now interested in modelling the mean and finding the main sources of variation across all of the subjects. To do this, we modelled both the mean function and the random effects using Fourier basis functions. Using cross-validation, shown in Figure 6.4, the optimal mean and random effects basis sizes were $m = 5$, and $q = 10$.

In Figure 6.5 we show the estimated mean function, covariance function and first two eigenfunctions using $m = 5$, $q = 10$. The estimated mean function shows that generally the noise stimulus results in a rise, followed by a fall, in the EEG values before approximately returning to the baseline value. The first eigenfunction accounts for 46.6% of the variability in the data. It represents an approximately linear increase/decrease in the EEG values, and could be captured by a straight line random effect. Eigenfunction 2 accounts for approximately 13.2% of the variability. It represents an increase (decrease) in amplitude of the EEG curves, especially around the 300 milliseconds trough, plus a small time shift. The third important eigenfunction accounted for 12.0% of the variability in the data.

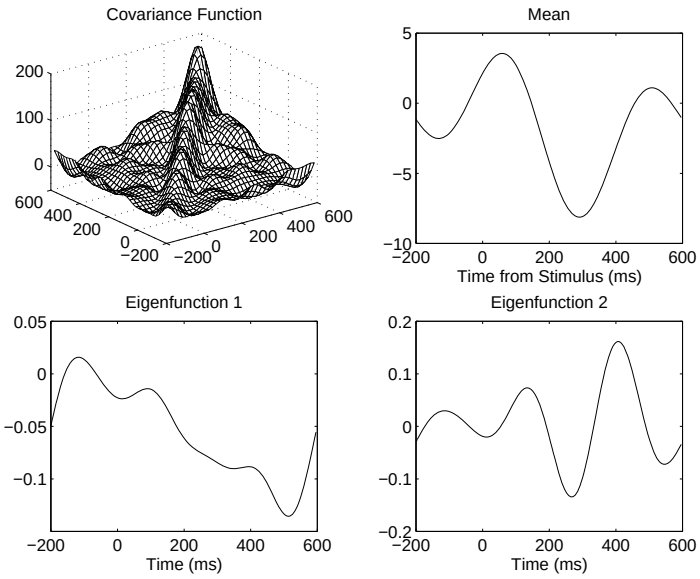


Figure 6.5: *Top left: estimated covariance function for the EEG data using* $q = 10$ *Fourier basis functions. Top right: estimated mean function using* $m = 5$ *Fourier basis functions. Bottom left: estimate of the first eigenfunction. Bottom right: estimate of the second eigenfunction.*

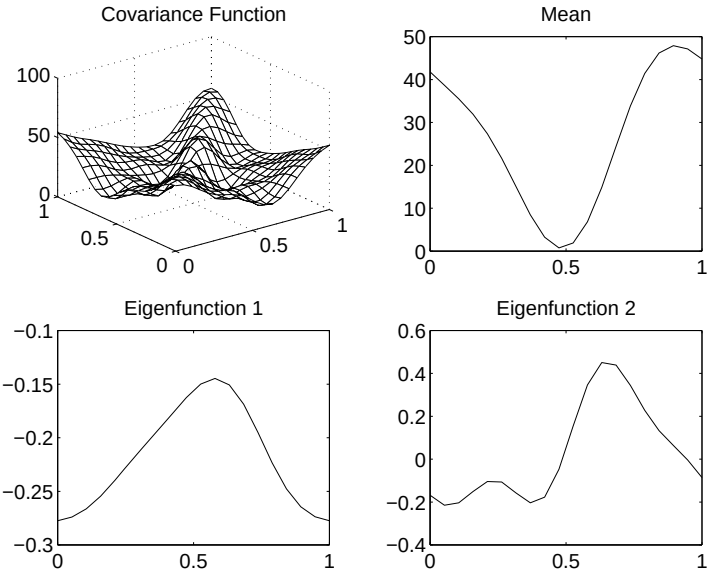


Figure 6.6: *Top left: estimated covariance function for the gait data using $q = 7$ Fourier basis functions. Top right: estimated mean function using $m = 4$ Fourier basis functions. Bottom left: estimate of the first eigenfunction. Bottom right: estimate of the second eigenfunction.*

6.5.2 Gait Data

We now turn our attention to the human gait data analysed in Rice and Silverman (1991), and again in Rice and Wu (1999). Measurements were taken on the angles formed by the hip over a single gait cycle for 39 children. The gait cycles started and ended with the heel on the ground. For each child, 16 to 22 measurements were taken, and then interpolated to give 20 equispaced points. In Rice and Silverman (1991), the interpolated data was used to illustrate eigenfunction analysis while in Rice and Wu (1999) the uninterpolated data was used to illustrate mixed effects modelling. We apply our basis mixed effects model to the interpolated data, which was given in Figure 1.1(c).

Since the data is periodic, it is reasonable to model both the mean and random effects using Fourier basis functions. Using cross-validation, the optimal number of mean and random effects basis functions were found to be $m = 4$ and $q = 7$. There were three important eigenfunctions accounting for 68.2%, 12.1% and 8.0% of the variation. The estimated mean function, covariance function and first two eigenfunctions are displayed in Figure 6.6. These estimates are similar to those obtained by Rice and Wu (1999).

6.6 Discussion

In this chapter, we presented a basis method for the joint modelling of the mean and covariance functions of functional data. The data was modelled as the sum of a smooth mean function and a small number of random effects. The mean was represented as a linear combination of known basis functions while the random effects were modelled using the eigenfunctions of the mean adjusted data, where the eigenfunctions were approximated by a basis expansion.

The iterative algorithm was found to be simple to implement and fast to run. While a unique β is not guaranteed, we've found that the algorithm converges to the Moore-Penrose solution for a wide variety of starting values (β_0).

Chapter 7

Conclusion

7.1 Thesis Contribution

A number of problems in the field of functional data analysis have been considered in this thesis. In particular, we have developed

- a new technique for logistic regression with a functional predictor. A basis function approach was used to generate maximum likelihood parameter estimates via an iterative algorithm. Estimates were found without and with a penalty term, which more explicitly controlled the smoothness of the resulting parameters (Chapter 3);
- new methods for functional regression and functional logistic regression when the functional regressor has a special structure, viz a repeated stimulus. Both the time within a stimulus (functional) and the position of the stimulus were incorporated into the model. A basis function approach was again used (Chapter 4), and;
- a new algorithm for the joint modelling of the mean and covariance functions of functional data, in what is a functional mixed-effects type model. Basis functions were used to model the mean, random effects and eigenfunctions. While a unique solution is not guaranteed, the algorithm converges to the (unique) Moore-Penrose solution for some starting values (Chapter 6).

The new techniques for FDA developed in this thesis were applied to biostatistical data. It was found that

- there is a slight difference in the way males and females process a stimulus, as found by examining the stimulated EEG tracings from the frontal lobe (Fz) position of the brain (Chapter 3);

- the stimulated fetal heart rates can be used to predict the infant's risk category at birth and the infant's development at 18 (and 36) months of age. The infant's sex was also important in predicting the child's psychomotor development at 18 months; with females outperforming males. The importance of the stimulus for the risk category at birth was established through the use of unstimulated control subjects. (Chapter 5), and;

- the average EEG tracing shows a rise and then a fall resulting from the stimulus. The first three eigenfunctions account for 71.8% of the variability in the EEG data and the first two could be represented by a straight line random effect and an increase in the amplitude of the tracings (Chapter 6).

7.2 Recommendations for Future Research

- Functional data analysis techniques do not rely on a stationarity assumption. If the functional data were stationary, then time series methods could be used as they are more efficient. Development of suitable methods for testing for stationarity of functional data is still needed.

- The methods presented in this thesis used the basis function approach to overcome the singularity problems associated with more measurements than subjects. However, other nonparametric smoothing techniques could be used. Whilst there has been some use of smoothing splines and kernels, further investigation of these and other methods, such as local polynomials, is required.

- The estimates generated by FDA techniques depends on the smoothing parameters. In this thesis, we used cross-validation to select the smoothing parameters. However, cross-validation can be time consuming. Further research is needed into this and other

data based selection procedures in order to develop efficient, reliable procedures for smoothing parameter selection.

- In this thesis, we examined the case of equispaced time points. Whilst the techniques can be adapted to random time points, research into its effect on the estimates, model diagnostics, smoothing parameter selection, etc is required.

- Research into other issues and techniques associated with FDA is also needed, for example data display, curve registration, and a true functional discriminant analysis.

Appendix A

Infant Developmental Assessment

Seventy-three infants were tested at 18 months of age using the Bayley Scales of Infant Development (BSID) (Bayley, 1993). All the tests were done by Dr Robyn Dolby[1], a psychologist who has 15 years of experience in infant developmental assessment. She had no knowledge of the clinical details of any of the infants.

The BSID were chosen for this assessment because they are standardised and are the scales most widely used in published research in child development. In addition, the second edition has been written to improve the utility of the scales with clinical populations, including children born at risk due to adverse antenatal factors or delivery complications. The scales consist of:

1. The Mental Development Index (MDI) and Psychomotor Developmental Index (PDI) assess the infant's current level of cognitive, language, personal-social and fine- and gross-motor development.

2. The Behaviour Rating Scale assesses qualitative aspects of the infant's test-taking behaviour. There are large differences between children in how they work, for example: in how much structure they need to be able to concentrate; in how smoothly they move from one task to the next; in how persistent they are; in how quickly they become frustrated; and in whether they are distracted from the task at hand by restless activity or hypersensitivity. The Behaviour Rating Scale has been specifically developed to assess such subtle behavioural difference.

[1]School of Behavioural Sciences, Macquarie University, Sydney, Australia

Outcome assessment also takes into account the contribution of parenting influences upon child development. The impact of parenting on development is evaluated in two ways:

1. globally, by assessing socio-economic indices, including highest level of education and occupation for both mother and father;

2. by examining the level of emotional support and parental involvement available to the child at home. The Home Observation for Measurement of the Environment (HOME) (Caldwell *et al.*, 1967) assesses such support, through interview and observation.

By deliberately measuring parenting influences in such detail, a large proportion of the variability in developmental outcome due to parenting influences can be captured. These parenting influences can then be corrected for to assess how much development can be attributed to antenatal events.

Appendix B

Notation and Abbreviations

B.1 Abbreviations

ANOVA	Analysis of variance
BLUP	Best linear unbiased estimator
BSID	Bayley scales of infant development
CV	Cross-validation
EEG	Electroencephalographic
FANOVA	Functional ANOVA
FDA	Functional data analysis
fPCA	Functional principal component analysis
Fz	Frontal lobe
GCV	Generalised cross-validation
glm	Generalised linear model
HOME	Home observation for measurement of the environment
MISE	Mean integrated squared error
MSE	Mean squared error
NID	Normally and independently distributed
PCA	Principal component analysis
PDI	Psychomotor development index
ROC	Receiver operating characteristic
RSS	Residual sum of squares

B.2 Notation

$\|a\|^2$	$= \langle a, a \rangle$
$\langle a, b \rangle$	$= \int a(t)b(t)dt$

Chapter 2 - Technical background

Let a_n and b_n be sequences of real numbers. Then, as $n \to \infty$

$a_n = o(b_n)$ if and only if $\lim_{n\to\infty} |a_n/b_n| = 0$

$a_n = O(b_n)$ if and only if $\limsup_{n\to\infty} |a_n/b_n| < \infty$

$A(h)$	Hat matrix associated with a smoothing parameter of h, $\hat{g}(t) = A(h)y$
c_k	Basis coefficient
$f(y)$	Kernel density function
$g(t)$	Nonlinear function relating t and y
h	Smoothing parameter / window width
$K(u)$	Kernel function
K_v^*	Equivalent kernel function
m	Number of basis functions
n	Number of time points (Sections 2.1 - 2.3)
	Number of independent subjects (Section 2.4 and elsewhere in the thesis)
p	Number of time points / measurements per subject
t_i	Time associated with y_i
y_i	Observed response
y	Vector of observed responses
$w(t)$	Weight function
ε_i	Error associated with subject i
$\psi_k(t)$	Basis function

Chapter 3 - Functional logistic regression

b	$(m \times 1)$ vector of basis coefficients for $\beta(t)$
c_i	$(m \times 1)$ vector of basis coefficients for subject i
C	$(n \times m)$ matrix of basis coefficients for X
$D(y; \hat{\pi})$	Deviance function
h	Smoothing parameter
$l(\pi; y)$	Log-likelihood function
m	Number of basis functions used to model functional data and parameter
n	Number of subjects
p	Number of time points / measurements per subject
$P(\beta)$	Penalty function; integrated squared second derivative of $\beta(t)$
r	Number of scalar covariates
t	Time
w^*	Diagonal weights matrix
W	$\int \psi(s)\psi^T(s)ds$
$x_i(t)$	Observed functional data at time t
$x(t)$	$(n \times 1)$ vector of functional data at time t
X	$(n \times p)$ matrix of functional data at all t
y_i	Observed response
z_i	$(r + 1 \times 1)$ design vector for subject i scalar covariates
Z	$(n \times r + 1)$ design matrix for scalar covariates
α	$(r + 1 \times 1)$ vector of mean and scalar parameters
$\beta(t)$	Functional parameter
ε_i	Error associated with subject i
η_i	Linear predictor
ν_i	Adjusted response variable
ν	$(n \times 1)$ vector of adjusted response variables
π_i	Probability of success
π	$(n \times 1)$ vector of probabilities

$\psi_k(t)$	Basis function
$\psi(t)$	$(m \times 1)$ vector of basis functions at time t
Ψ	$(m \times p)$ matrix of basis functions at all t

Chapter 4 - Functional data with a repeated stimulus

C_i	$(q \times m)$ matrix of basis coefficients for subject i
D	Functional design matrix for estimating b
E	Functional design matrix for estimating γ
p	Number of time points / measurements per subject within a stimulus
q	Number of stimuli
$x_{i,s}(t)$	Observed functional data at time t within the s^{th} stimulus for subject i
X_i	$(q \times p)$ matrix of functional data for subject i at all (s,t) combinations
$\beta(t)$	Functional parameter within a stimulus
γ	$(q \times 1)$ vector of stimuli parameters

Chapter 6 - Functional mean and covariance modelling

$\langle a, b \rangle$	$\approx \sum_{l=1}^{p} a(l\Delta)b(l\Delta)\Delta$
b_k^i	Basis coefficient for subject i noise
b_i	$(q \times 1)$ vector of noise basis coefficients
e	$(n \times 1)$ vector of residuals
f_k	$(q \times 1)$ vector of eigenfunction basis cofficients
F	$(q \times r)$ matrix of basis coefficients for the dominant eigenfunctions
$\mathcal{L}$	Log-likelihood function of the model
m	Number of mean basis functions
n	Number of subjects
N_i	$(p \times 1)$ vector of subject i noise at all t
p	Number of time points / measurements per subject
q	Number of noise basis functions
r	Number of dominant eigenfunctions
S_b	Sample covariance matrix for the b_i
S_e	Sample covariance matrix for the residuals
t	Time
W	$Z^T Z\Delta$
X	$(p \times m)$ matrix of mean basis functions for all t
$y_i(t)$	Observed functional data for subject i at time t
Y_i	$(p \times 1)$ vector of observed functional data for subject i
Z	$(p \times q)$ matrix of noise basis functions for all t
β_k	Mean basis coefficient
β	$(m \times 1)$ vector of mean basis coefficients
$\hat{\beta}_0$	Initial estimate of β
$\Gamma(s, t)$	Covariance function between times s and t
$\Gamma_n(s, t)$	Sample covariance matrix for Y
Δ	Distance between successive time points

$\zeta(t)$	Noise basis function
θ_k	Eigenvalue ratio ($= \lambda_k / \sum \lambda_u$)
λ_k	Eigenvalue
$\mu(t)$	Mean function
$\nu_i(t)$	Noise (random effect plus residual)
$\phi_k(t)$	Eigenfunction
$\phi(t)$	$(r \times 1)$ vector of dominant eigenfunctions at time t
Φ	$(p \times r)$ matrix of dominant eigenfunctions at all t
$\chi_k(t)$	Mean basis function

Appendix C

Software Documentation

In this appendix we document the software developed in the course of this thesis. All analyses were performed in Matlab using code developed by the author. The software has been organised into a Matlab toolbox called fda.

The fda toolbox is divided into four directories:

fglm	Functional Generalised Linear Modelling
fregrep	Regression Analysis with a Repeatedly Stimulated Functional Regressor
fmucov	Functional Mean and Covariance Modelling
util	Utility Functions

containing the following functions:

fglm	Functional Generalised Linear Modelling
`flog`	Calculate a functional generalised linear model for binary data
`cvflog`	Parameter selection for `flog` via cross-validation

fregrep	Regression Analysis with a Repeatedly Stimulated Functional Regressor
`loadfet`	Load and format repeatedly stimulated functional regressor for all subjects
`fetmod`	Generates and formats basis coefficients for repeatedly stimulated regressor
`flmfet`	Performs functional linear modelling with a repeatedly stimulated regressor
`cvflmfet`	Parameter selection for `flmfet` via cross-validation
`fetglm`	Performs functional generalised linear modelling with a repeatedly stimulated regressor
`cvfetglm`	Parameter selection for `fetglm` via cross-validation

fmucov	Functional Mean and Covariance Modelling
`fmucov`	Performs functional mean and covariance modelling
`cvfmucov`	Parameter selection for `fmucov` via cross-validation

util	Utility Functions
`basisfns`	Generate basis functions
`bsplineb`	Generate B-spline basis functions
`fcvplot`	Plot results from any function performing cross-validation
`fourierb`	Generate Fourier basis functions
`splitgp`	Separates grouped data listed in one column into multiple columns

The software documentation is organised as follows:

- name of function
- a statement of purpose
- a synopsis of the function's syntax
- a description of what the function does

flog

Purpose

Calculate a functional generalised linear model for binary data

Syntax

```
[alpha,beta,prob] = flog(y,x,t,pin)
[alpha,beta,prob] = flog(y,x,t,pin,pred)
[alpha,beta,prob] = flog(y,x,t,pin,pred,link)
[alpha,beta,prob] = flog(y,x,t,pin,pred,link,MAXIT,tol)
```

Description

`[alpha,beta,prob] = flog(y,x,t,pin)` for a binary response vector given in file `y`, $(n \times p)$ functional regressor matrix in file `x`, with measurements taken at times given in file `t`, returns parameter estimates for the constant `alpha`, functional parameter `beta`, and the predicted probabilities of success `prob` from a functional logistic regression.

The number of basis functions `m`, the type of basis functions `btype` and their period or degree `deg`, and the smoothing parameter value `h`, is controlled by `pin = [m,btype,deg,h]`. The possible types of basis functions are: `1` = B-spline, `2` = Fourier.

The parameter estimates `alpha,beta,prob` are found using an adaptation of the standard glm algorithm (McCullagh and Nelder, 1989), as outlined in Chapter 3.

`[alpha,beta,prob] = flog(y,x,t,pin,pred)` includes other scalar covariates (regressors) in the model. Their parameter estimates will be included in `alpha`.

`[alpha,beta,prob] = flog(y,x,t,pin,pred,link)` specifies the link function to be used in the algorithm. If it is not specified, functional logistic regression is performed.

`[alpha,beta,prob] = flog(y,x,t,pin,pred,link,MAXIT,tol)` specifies the maximum number of iterations to perform and the stopping tolerance of the algorithm. The default `MAXIT` is `50` and `tol` is `1e-8`.

cvflog

Purpose

Parameter selection for `flog` via cross-validaton

Syntax

```
cv = cvflog(y,x,t,pin)
cv = cvflog(y,x,t,pin,mvec)
cv = cvflog(y,x,t,pin,mvec,hvec)
cv = cvflog(y,x,t,pin,mvec,hvec,pred)
```

Description

`cv = cvflog(y,x,t,pin)` for a binary response vector given in file `y`, functional regressor matrix in file `x`, with measurements taken at times given in file `t`, returns the CV results for all possible number of basis functions.

The CV scores (`cv.scores`) are calculated using the function given in (3.10). The optimal number of basis functions (`cv.m`) is the value of `m` which maximises the log-likelihood.

`cv = cvflog(y,x,t,pin,mvec)` specifies the number of basis functions (m) to be tested, with `mvec = [mmin mmax stepm]`. The default values are a minimum value of `mmin = 1`, maximum value of `mmax` = the number of subjects, with the values increasing in steps of `stepm = 1`.

`cv = cvflog(y,x,t,pin,mvec,hvec)` specifies the smoothing parameter h values to be tested, with `hvec = [hmin hmax nh]`. The default values are a minimum h value of `hmin = 0`, maximum value of `hmax = 1`, with `nh = 11` values of h tested. The CV scores under this option are calculated using the function given in (3.13).

`cv = cvflog(y,x,t,pin,mvec,hvec,pred)` includes other scalar covariates (regressors) in the model.

loadfet

Purpose

Load and format repeatedly stimulated functional regressor for all subjects

Syntax

```
loadfet(datafile)
```

Description

Given a file `datafile` listing the files containing the repeatedly stimulated functional regressor for each subject, `loadfet(datafile)` loads and formats each file in `datafile` into a single variable for use in later analyses.

---

fetmod

Purpose

Generates and formats basis coefficients for repeatedly stimulated regressor

Syntax

```
fetmod(m)
```

Description

Takes single variable generated by `loadfet` and calculates Fourier basis coefficients for the data, within a stimulus, for each subject. The number of basis functions used is controlled by `m`.

flmfet

Purpose

Performs functional linear modelling with a repeatedly stimulated regressor

Syntax

```
[alpha,beta,gamma,fit,resid] = flmfet(y,m,pred)
[alpha,beta,gamma,fit,resid] = flmfet(y,m,pred,maxit,gamstart,ind)
```

Description

`[alpha,beta,gamma,fit,resid] = flmfet(y,m,pred)` for a continuous response vector given in file `y`, `m` basis functions, and `pred` listing the repeatedly stimulated regressor variable (followed by any optional scalar regressors), returns estimates for the scalar regressor parameters `alpha`, functional parameter `beta`, and stimuli parameters `gamma` from a functional regression with a repeatedly stimulated regressor, as well as the fitted values `fit` and the residuals `resid` of the model.

The parameter estimates `alpha,beta,gamma` are found using the algorithm developed, and described, in Section 4.2.1.

Further, `maxit` specifies the maximum number of iterations, and `gamstart` the starting values for `gamma` (default is a vector of ones). `ind` specifies the algorithm's output mode: 0 (standard output), 1 (quiet) and 2 (extended output).

cvflmfet

Purpose

Parameter selection for `flmfet` via cross-validation

Syntax

```
cv = cvflmfet(y,mvec,pred)
cv = cvflmfet(y,mvec,pred,maxit,gamstart)
```

Description

`cv = cvflmfet(y,mvec,pred)` for a continuous response vector given in file `y` and `pred` listing the repeatedly stimulated regressor variable (followed by any optional scalar regressors), returns the CV results for the number of basis functions (m) as determined by `mvec`.

`mvec = [minm,maxm,sm]` specifies the minimum m value to test (`minm`), maximum value (`maxm`) and the step size (`sm`) to take between `minm` and `maxm`.

The CV scores (`cv.scores`) are calculated using the function given in (4.10). The optimal number of basis functions (`cv.m`) is the value of `m` which minimises the cross-validated residual sum of squares.

Further, `maxit` specifies the maximum number of iterations, and `gamstart` the starting values for `gamma` (default is a vector of ones).

fetglm

Purpose

Performs functional generalised linear modelling with a repeatedly stimulated regressor

Syntax

```
[alpha,beta,gamma,fit,prob] = fetglm(y,m,pred)
[alpha,beta,gamma,fit,prob] = fetglm(y,m,pred,pval,maxit)
```

Description

`[alpha,beta,gamma,fit,prob] = fetglm(y,m,pred)` for a continuous response vector given in file `y`, `m` basis functions, and `pred` listing the repeatedly stimulated regressor variable (followed by any optional scalar regressors), returns estimates for the scalar regressor parameters `alpha`, functional parameter `beta`, and stimuli parameters `gamma` from a functional generalised linear model with a repeatedly stimulated regressor, as well as the fitted values `fit` and the residuals `resid` of the model.

The parameter estimates `alpha,beta,gamma` are found using the algorithm developed, and described, in Section 4.3.1.

Further, `pval` specifies the cut-off probability for classification of the predicted probabilities. The default is `0.5`. `maxit` specifies the maximum number of iterations.

cvfetglm

Purpose

Parameter selection for `fetglm` via cross-validation

Syntax

```
cv = cvfetglm(y,mvec,pred)
cv = cvfetglm(y,mvec,pred,maxit)
```

Description

`cv = cvfetglm(y,mvec,pred)` for a continuous response vector given in file `y` and `pred` listing the repeatedly stimulated regressor variable (followed by any optional scalar regressors), returns the CV results for the number of basis functions (m) as determined by `mvec`.

`mvec = [minm,maxm,sm]` specifies the minimum m value to test (`minm`), maximum value (`maxm`) and the step size (`sm`) to take between `minm` and `maxm`.

The CV scores (`cv.scores`) are calculated using the function given in (3.10). The optimal number of basis functions (`cv.m`) is the value of `m` which maximises the cross-validated log-likelihood.

Further, `maxit` specifies the maximum number of iterations to be performed.

fmucov

Purpose

Performs functional mean and covariance modelling

Syntax

```
[mu,re,phi,lam] = fmucov(y,t,pin,m,q,r)
[mu,re,phi,lam] = fmucov(y,t,pin,m,q,r,maxit)
```

Description

`[mu,re,phi,lam] = fmucov(y,t,pin,m,q,r)` for functional data given in file `y`, with measurements recorded at times given in file `t`, returns the mean `mu`, random effects `re`, eigenfunctions `phi` and eigenvalues `lam` from a functional mean and covariance modelling.

`pin = [mubtype,mudeg,rebtype,redeg]` controls the type of basis functions used for the mean, and their degree/periodicity, plus the type of basis functions used for the random effects and eigenfunctions, and their degree/periodicity. The possible types of basis functions are: `1` = B-spline, `2` = Fourier, `3` = polynomial, `4` = Legendre polynomials. The number of mean basis functions is controlled by `m`, the number of random effects basis functions by `q`, and the number of important eigenfunctions by `r`.

The estimated `[mu,re,phi,lam]` are calculated using the algorithm described in Section 6.2.

Further, `maxit` specifies the maximum number of iterations.

cvfmucov

Purpose

Parameter selection for `fmucov` via cross-validation

Syntax

```
cv = cvfmucov(y,t,pin,mvec,qvec,r)
cv = cvfmucov(y,t,pin,mvec,qvec,r,maxit)
```

Description

`cv = cvfmucov(y,t,pin,mvec,qvec,r)` for functional data given in file `y`, with measurements recorded at times given in file `t`, returns the CV results for choosing the optimal number of mean (`cv.m`) and random effect basis functions (`cv.q`).

`pin = [mubtype,mudeg,rebtype,redeg]` controls the type of basis functions used for the mean, and their degree/periodicity, plus the type of basis functions used for the random effects and eigenfunctions, and their degree/periodicity. The number of important eigenfunctions is specified by `r`.

`mvec = [minm,maxm,stepm]` specifies the number of mean basis functions to be tested: `minm` = minimum m value to test, `maxm` = maximum value, and `stepm` = the step size to take between `minm` and `maxm`. Similarly, `qvec = [minq,maxq,stepq]` specifies the number of random effects basis functions to be tested.

Appendix D

Publications

Papers

The journal editor has indicated likely acceptance subject to revision of the following papers.

- S.J.Ratcliffe, L.R.Leader and G.Z.Heller, "Functional Data Analysis with Application to Periodically Stimulated Fetal Heart Rate Data: I. Functional Regression", Submitted to *Statistics in Medicine*.

- S.J.Ratcliffe, G.Z.Heller and L.R.Leader, "Functional Data Analysis with Application to Periodically Stimulated Fetal Heart Rate Data: II. Functional Logistic Regression", Submitted to *Statistics in Medicine*.

Conference Presentations

- S.J.Ratcliffe and V.Solo, "Functional Mean and Covariance Modelling", In *ASA Proceedings: Joint Statistical Meetings, Section on Statistical Computing*, pages 206-209, August 1999.

- S.J.Ratcliffe and V.Solo, "Some Issues in Functional Principal Component Analysis", In *ASA Proceedings: Joint Statistical Meetings, Section on Statistical Computing*, pages 119-123, August 1998.

Bibliography

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, **22**, 203–217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281, Budapest. Akademia Kiado.

Akaike, H. (1978). A bayesian analysis of the mimimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, **30**, 9–14.

Allen, D. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–127.

Altman, N. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association*, **85**, 749–759.

Anderson, S. and Jones, R. (1995). Smoothing splines for longitudinal data. *Statistics in Medicine*, **14**, 1235–1248.

Antoniadis, A., Gregoire, G., and McKeague, I. (1994). Wavelet methods for curve estimation. *Journal of the American Statistical Association*, **89**, 1340–1353.

Banks, D., Maxion, R., and Olszewski, R. (1995). Comparing methods for multivariate nonparametric regression. In *Proceedings of the Statistical Computing Section of the American Statistical Association*, pages 136–141.

Bayley, N. (1993). *The Bayley Scales of Infant Development*. Harcourt Brace and Co., San Antonio, 2nd edition.

Besse, P. and Ramsay, J. (1986). Principal components analysis of sampled functions. *Psychometrika*, **51**, 285–311.

Billingsley, P. (1961). *Statistical Inference for Markov Processes*. University of Chicago Press, Chicago.

Bornstein, M. and Sigman, M. (1986). Continuity in mental development from infancy. *Child Development*, **57**, 251–274.

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–60.

Box, G. and Jenkins, G. (1976). *Time Series Analysis, Forecasting, and Control*. Holden-Day, San Francisco, 2nd edition.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Brackbill, Y., Kane, J., Manniello, R., and Adamson, D. (1974). Obstetric premedication and infant outcome. *American Journal of Obstetrics and Gynecology*, **118**, 337–384.

Brillinger, D. (1969). The canonical analysis of stationary time series. In P. Krishnaiah, editor, *Multivariate Analysis II*, pages 331–350. Academic Press, New York.

Brillinger, D. (1973). The analysis of time series collected in an experimental design. In P. Krishnaiah, editor, *Multivariate Analysis III*, pages 241–256. Academic Press, New York.

Brillinger, D. (1980). Analysis of variance and problems under time series models. *Handbook of Statistics*, **1**, 237–278.

Brillinger, D. (1981a). *Time Series: Data Analysis and Theory*. Holden-Day Inc., San Francisco, expanded edition.

Brillinger, D. (1981b). *Time Series: Data Analysis and Theory*. Holden-Day, San Francisco, expanded edition.

Bronstein, A., Itina, N., and Kamenetsaia, A. (1968). The orienting reaction in newborn children. In L. Varonin, A. Leontiev, A. Luris, E. Sokolov, and O. Vinogradova, editors,

Orienting Reflex and Exploratory Behaviour. Moscow Academy of Pedagogical Sciences of RSFSR.

Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, **93**, 961–976.

Buckwald, J. and Humphrey, G. (1973). An analysis of habituation in specific sensory systems. In E. Steller and J. Sprague, editors, *Progress in Physiological Psychology*, volume 5, pages 1–75. Academic Press, New York.

Caldwell, B., Bradley, R., and Staff (1967). *Home Observation for Measurement of the Environment*. Center for Child Development and Education, University of Arkansas at Little Rock, Little Rock, Arkansas.

Castro, P., Lawton, W., and Sylvestre, E. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics*, **28**, 329–337.

Chiu, S.-T. (1989). Bandwidth selection for kernel estimate with correlated noise. *Statistics and Probability Letters*, **8**, 347–354.

Chiu, S.-T. (1990). Why do bandwidth selectors tend to choose smaller bandwidths and a remedy. *Biometrika*, **77**, 222–226.

Chui, C. (1992). *An Introduction to Wavelets*. Academic Press, San Diego.

Clark, R. (1975). A calibration curve for radiocarbon dates. *Antiquity*, **49**, 251–266.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Cleveland, W. and Devlin, S. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610.

Cook, R. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. Chapman and Hall, New York.

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

Cox, M. (1972). The numerical evaluation of B-spline. *Journal of the Institute of Mathematics and Its Applications*, **10**, 134–149.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377–403.

de Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory*, **6**, 50–62.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer, New York.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the E-M algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Diggle, P. and Hutchinson, M. (1989). On spline smoothing with autocorrelated errors. *Australian Journal of Statistics*, **31**(1), 166–182.

Diggle, P., Liang, K., and Zeger, S. (1994). *Analysis of Longitudinal Data*. Oxford University Press.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**, 425–455.

Eisenberg, R., Coursin, D., and Rupp, N. (1966). Habituation to an acoustic pattern as an index of differences among human neonates. *Journal of Auditory Research*, **6**, 239–248.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, **21**, 196–216.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

Fan, J. and Lin, S. (1998). Test of significance when data are curves. Technical report, Dept. Statistics, UNC-Chapel Hill.

Fan, J. and Zhang, J. (1998). Functional linear models for longitudinal data. Technical report, Dept. Statistics, UNC-Chapel Hill.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1995). On nonparametric estimation via local polynomial regression. Discussion Paper 9511, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

Faraway, J. (1997). Regression analysis for a functional response. *Technometrics*, **39**, 254–261.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications*. Wiley, New York, 3rd edition.

Ferguson, I., Lenman, J., and Johnston, B. (1978). Habituation of the orbicularis oculi reflex in dementia and dyskinetic states. *Journal of Neurology, Neurosurgery and Psychiatry*, **41**, 824–828.

Fisher, R. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, **22**, 700–725.

Friedman, J. and Silverman, B. (1989). Flexible parsimonious smoothing and additive modeling. *Journal of the American Statistical Association*, **90**, 1179–1188.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In T. Gasser and M. Rosenblatt, editors, *Smoothing Techniques for Curve Estimation*, pages 23–68. Springer-Verlag, Heidelberg.

Gasser, T., Müller, H.-G., and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B*, **11**, 171–185.

Goldstein, H. (1979). *The Design and Analysis of Longitudinal Studies: Their Role in the Measurement of Change*. Academic Press, London.

Goutis, C. (1998). Second-derivative functional regression with applications to near infra-red spectroscopy. *Journal of the Royal Statistical Society, Series B*, **60**, 103–114.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall, London.

Hamming, R. (1989). *Digital Filters.* Prentice Hall, New Jersey, 3rd edition.

Hanley, J. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. *Critical Reviews in Diagnostic Imaging*, **29**, 307–335.

Hannan, E. (1961). The general theory of canonical correlation and its relation to functional analysis. *Journal of the Australian Mathematical Society*, **2**, 229–242.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge.

Härdle, W. (1991). *Smoothing Techniques with Implementation in S.* Springer, New York.

Härdle, W., Hall, P., and Marron, S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *Journal of the American Statistical Association*, **83**, 86–99.

Hart, J. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Series B*, **53**(1), 173–187.

Hart, J. and Wehrly, T. (1986). Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association*, **81**, 1527–1546.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, New York.

Holloway, F. and Parsons, O. (1971). Habituation of the orienting response in brain damaged patients. *Psychophysiology*, **8**, 623–634.

Hutt, S. and Hutt, C. (1964). Hyperactivity in a group of epileptic (and some non-epileptic) brain damaged children. *Epilepsia*, **5**, 334–351.

Hutt, S., Hutt, C., Lee, D., and Dunstead, C. (1965). A behavioural and electroencephalographic study of autistic children. *Journal of Psychiatric Research*, **3**, 181–197.

James, G. and Hastie, T. (1999). Principal component models for sparse functional data. presented at Joint Statistical Meeting 1999, Baltimore.

Jeffrey, W. and Cohen, L. (1971). Habituation in the human infant. In H. Reese, editor, *Advances in Child Development and Behaviour*, volume 6, pages 63–97. Academic Press, New York.

Jenkins, G. and Watts, D. (1968). *Spectral Analysis and its Applications*. Holden-Day, San Francisco.

Jones, M. and Rice, J. (1992). Displaying the important features of large collections of similar curves. *American Statistician*, **16**, 140–145.

Jones, R. (1993). *Longitudinal Data with Serial Correlation: A State-Space Approach*. Chapman and Hall, New York.

Kanwal, R. (1971). *Linear Integral Equations, Theory and Technique*. Academic Press, New York.

Kendall, M. and Ord, J. (1990). *Time Series*. Oxford University Press, New York, 3rd edition.

Kenward, M. (1987). A method for comparing profiles of repeated measurements. *Applied Statistics*, **36**, 296–308.

Kneip, A. (1994). Nonparametric estimation of common regressors for similar curve data. *Annals of Statistics*, **22**, 1386–1427.

Kneip, A. and Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *Annals of Statistics*, **20**, 1266–1305.

Kohn, R., Ansley, C., and Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the American Statistical Association*, **86**, 1042–1050.

Kohn, R., Ansley, C., and Wong, C.-M. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika*, **79**, 335–346.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Leader, L., Baillie, P., Martin, B., and Vermeulen, E. (1982). The assessment and significance of habituation to a repeated stimulus by the human fetus. *Early Human Development*, **7**, 211–219.

Lewis, M. (1971). Individual differences in the measurement of early cognitive growth. exceptional infant 2. In J. Hellmuth, editor, *Studies in Abnormalities*, pages 172–210. Brunner Mazel, New York.

Li, K.-C. (1984). Consistency for cross-validated nearest neighbor estimates in nonparametric regression. *Annals of Statistics*, **12**(1), 230–240.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics*, **15**, 958–975.

Liang, K.-Y., Zeger, S., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**, 3–40.

Lindsey, J. (1993). *Models for Repeated Measurements*. Oxford University Press, Oxford.

Macauley, F. (1931). *The Smoothing of Time Series*. National Bureau of Economic Research, New York.

Madison, L., Adubato, S., Madison, J., Nelson, R., Anderson, J., Erickson, J., Kuss, L., and Goodlin, R. (1986). Fetal response decrement: True habituation? *Journal of Developmental and Behavioral Pediatrics*, **7**(1), 14–20.

Marron, J. (1996). A personal view of smoothing and statistics. In W. Härdle and S. Schimek, editors, *Statistical Theory and Computational Aspects of Smoothing*, pages 1–9.

McCall, R. and Carriger, M. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development*, **64**, 57–79.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2nd edition.

Muenz, L. and Rubinstein, L. (1985). Markov models for covariate dependence of binary sequences. *Biometrics*, **41**, 91–101.

Müller, H. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association*, **82**, 231–238.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its Applications*, **84**, 66–72.

Olshen, R., Biden, E., Wyatt, M., and Sutherland, D. (1989). Gait analysis and the bootstrap. *Annals of Statistics*, **17**, 1419–1440.

Parzen, E. (1962). On estimation of a probability density and mode. *Annals of Mathematical Statistics*, **35**, 1065–1076.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Pezzulli, S. and Silverman, B. (1993). Some properties of smoothed principal components analysis for fuctional data. *Computational Statistics*, **8**, 1–16.

Piegl, L. and Tiller, W. (1997). *The NURBS Book*. Springer, Berlin, 2nd edition.

Priestly, M. and Chao, M. (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society, Series B*, **34**, 385–392.

Ramsay, J. (1982). When the data are functions. *Psychometrika*, **47**, 379–396.

Ramsay, J. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society, Series B*, **60**, 365–375.

Ramsay, J. (2000). Functional components of variation in handwriting. *Journal of the American Statistical Association*, **95**, 9–15.

Ramsay, J. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B*, **53**, 539–572.

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. Springer, New York.

Ramsay, J., Altman, N., and Bock, R. (1994). Variation in height acceleration in the Fels growth data. *The Canadian Journal of Statistics*, **22**, 89–102.

Ramsay, J., Wang, X., and Flanagan, R. (1995). A functional data analysis of the pinch force of human fingers. *Applied Statistics*, **44**, 17–30.

Ramsay, J., Munhall, K., Gracco, V., and Ostry, D. (1996). Functional data analysis of lip motion. *Journal of the Acoustical Society of America*, **99**, 3718–3727.

Ratcliffe, S. and Solo, V. (1999). Functional mean and covariance modelling. In *ASA Proceedings: Joint Statistical Meetings, Section on Statistical Computing*, pages 206–209, Baltimore, Marylands.

Ratcliffe, S., Leader, L., and Heller, G. (2000a). Functional data analysis with application to periodically stimulated fetal heart rate data: I. Functional regression. Submitted to *Statistics in Medicine*.

Ratcliffe, S., Heller, G., and Leader, L. (2000b). Functional data analysis with application to periodically stimulated fetal heart rate data: II. Functional logistic regression. Submitted to *Statistics in Medicine*.

Reinsch, C. (1967). Smoothing by spline functions. *Numerische Mathematik*, **10**, 177–183.

Rice, J. and Rosenblatt, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. *Annals of Statistics*, **11**(1), 141–156.

Rice, J. and Silverman, B. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B*, **53**, 233–243.

Rice, J. and Wu, C. (1999). Nonparametric mixed effects models for unequally sampled noisy curves.

Robinson, G. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, **6**, 15–32.

Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. *Numerische Mathematik*, **10**, 177–183.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65–78.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process*, **26**, 43–49.

Schoenberg, I. (1964). Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences of the United States of America*, **52**, 947–950.

Schumaker, L. (1981). *Spline Functions.* Wiley, New York.

Schwarz, G. (1978). Estimating dimension of a model. *Annals of Statistics*, **6**, 461–464.

Seber, G. (1977). *Linear Regression Analysis.* Wiley, New York.

Shumway, R. (1970). Applied regression and analysis of variance for stationary time series. *Journal of the American Statistical Association*, **65**, 1527–1546.

Shumway, R. (1988). *Applied Statistical Time Series Analysis.* Prentice Hall, New Jersey.

Shumway, R., Tai, R., Tai, L., and Pawitan, Y. (1983). Statistical analysis of daily London mortality and associated weather and pollution effects. Technical Report 53, Dept. Statistics, UC-Davis.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Silverman, B. (1995). Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society, Series B*, **57**, 673–689.

Silverman, B. (1996). Smoothed functional principal components analysis by choice of norm. *Annals of Statistics*, **24**, 1–24.

Simonoff, J. (1996). *Smoothing Methods in Statistics.* Springer, New York.

Solo, V. (1997). Continuous-discrete functional principal components analysis. submitted to Journal of the Royal Statistical Society, Series B.

Solo, V. (2000). A simple derivation of the smoothing spline. *The American Statistician*, **54**(1), 40–45.

Solo, V. and Ratcliffe, S. (2000). Functional mean and covariance modelling with a basis-kernel method. To be submitted.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50**.

Staniswalis, J. and Lee, J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, **93**, 1403–1418.

Stone, C. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**, 595–620.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, **36**, 111–147.

Thompson, R. and Glansman, D. (1966). Neural and behavioural mechanisms of habituation and sensitisation. In T. Tighe and R. Leaton, editors, *Habituation*, pages 49–93. Lawrence Earlbaum Associates, Hillsdale, NJ.

Tizard, B. (1968). Habituation of EEG and skin potential changes in normal and severely sub-normal children. *American Journal of Mental Deficiency*, **73**, 16–43.

Utreras, F. (1988). Boundary effects on convergence rates for Tikhonov regularization. *Journal of Approximation Theory*, **54**, 235–249.

Vonesh, E. and Chinchilli, V. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Marcel Dekker, New York.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wahba, G. and Wold, S. (1975). A completely automatic French curve: Fitting spline functions by cross-validation. *Communications in Statistics*, **4**, 1–17.

Wand, M. and Jones, M. (1995). *Kernel Smoothing.* Chapman and Hall, London.

Wang, K. and Gasser, T. (1997). Alignment of curves by dynamic time warping. *Annals of Statistics*, **25**, 1251–1276.

Wang, K. and Gasser, T. (1999). Synchronizing sample curves nonparametrically. *Annals of Statistics*, **27**, 439–460.

Wang, Y. (1998). Mixed-effects smoothing spline ANOVA. *Journal of the Royal Statistical Society, Series B*, **60**, 159–174.

Watson, G. (1964). Smooth regression analysis. *Sankhya, Series A*, **26**, 359–372.

Whittaker, E. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, **41**, 63–75.

Whittle, P. (1958). On the smoothing of probability density functions. *Journal of the Royal Statistical Society, Series B*, **20**, 334–343.

Zeger, S. and Karim, M. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

Zeger, S., Liang, K.-Y., and Albert, P. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.