

# 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

<b>Summary</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>vi</b>
<b>Certificate</b>	<b>vii</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Functional Data Analysis . . . . .	2
1.1.1 Data Display . . . . .	2
1.1.2 Curve Registration . . . . .	5
1.1.3 Analysis . . . . .	5
1.2 Areas of Research . . . . .	7
1.3 Plan of Thesis . . . . .	8
<b>2 Technical Background</b>	<b>11</b>
2.1 Nonparametric Smoothing Methods . . . . .	11
2.1.1 Basis Functions . . . . .	13
2.1.2 Kernel Estimators . . . . .	18
2.1.3 Local Polynomial Regression . . . . .	21
2.1.4 Penalty Method - Smoothing Splines . . . . .	26
2.1.5 Comparison of Methods . . . . .	29

---

2.2	Controlling Smoothness . . . . .	31
2.2.1	Cross-Validation . . . . .	31
2.3	Methods for Correlated Data . . . . .	34
2.4	Towards Functional Data Analysis . . . . .	35
2.4.1	Classical Longitudinal Data Analysis . . . . .	35
2.4.2	Functional Data Analysis based on Stationarity . . . . .	37
2.5	Functional Data Analysis . . . . .	38
2.5.1	Functional ANOVA and Regression . . . . .	39
2.5.2	Functional Principal Component Analysis . . . . .	42
<b>3</b>	<b>Functional Logistic Regression</b>	<b>46</b>
3.1	Introduction . . . . .	46
3.2	Basis Method . . . . .	46
3.2.1	Estimating Parameters . . . . .	48
3.2.2	Choosing the Basis Dimension . . . . .	50
3.3	Truncated Basis Expansion plus Penalty . . . . .	50
3.3.1	Choosing $m$ and $h$ . . . . .	52
3.4	Model Diagnostics . . . . .	52
3.5	Application to EEG Data . . . . .	54
3.6	Discussion . . . . .	58
<b>4</b>	<b>Functional Data with a Repeated Stimulus</b>	<b>59</b>
4.1	Introduction . . . . .	59
4.2	Continuous Response . . . . .	59
4.2.1	Estimating Parameters . . . . .	62
4.2.2	Model Diagnostics . . . . .	65
4.2.3	Choosing $m$ . . . . .	66
4.3	Binary Response . . . . .	67
4.3.1	Estimating Parameters . . . . .	67
4.3.2	Choosing $m$ . . . . .	70

---

<b>5</b>	<b>The Fetal Heart Rate Data</b>	<b>71</b>
5.1	Introduction . . . . .	71
5.2	Study Description . . . . .	72
5.2.1	Description of Covariates . . . . .	74
5.3	Risk Category . . . . .	75
5.3.1	Controls . . . . .	80
5.4	PDI at 18 Months . . . . .	80
5.5	Discussion . . . . .	85
<b>6</b>	<b>Functional Mean and Covariance Modelling</b>	<b>88</b>
6.1	Previous Approaches . . . . .	88
6.2	The Basis Method . . . . .	91
6.3	Algorithm Analysis . . . . .	96
6.4	Simulation Study . . . . .	99
6.5	Examples . . . . .	102
6.5.1	EEG Data . . . . .	103
6.5.2	Gait Data . . . . .	105
6.6	Discussion . . . . .	106
<b>7</b>	<b>Conclusion</b>	<b>107</b>
7.1	Thesis Contribution . . . . .	107
7.2	Recommendations for Future Research . . . . .	108
<b>A</b>	<b>Infant Developmental Assessment</b>	<b>110</b>
<b>B</b>	<b>Notation and Abbreviations</b>	<b>112</b>
B.1	Abbreviations . . . . .	112
B.2	Notation . . . . .	113
<b>C</b>	<b>Software Documentation</b>	<b>118</b>
<b>D</b>	<b>Publications</b>	<b>129</b>
	<b>Bibliography</b>	<b>130</b>

# 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

1.1	Examples of functional data. . . . .	3
1.2	Curves corresponding to an estimated mean curve and the highest and lowest scores on the first principal component for the hip data. . . . .	4
2.1	Scatter plot of monthly birth rates in the United States. . . . .	13
2.2	Basis estimators for the birth rate data. . . . .	17
2.3	Kernel estimators for the birth rate data. . . . .	20
2.4	Local linear polynomial estimators for the birth rate data. . . . .	22
2.5	The Epanechnikov kernel and its equivalent kernel for local polynomial curve estimation. . . . .	23
2.6	Cubic smoothing spline estimators for the birth rate data. . . . .	27
2.7	Fourier coefficients for the birth rate data, the low pass filter and the resulting Fourier coefficients for the smoothing spline estimator. . . . .	28
2.8	Cross-validation and generalised cross-validation plots for a kernel estimator of the birth rate data. . . . .	33
3.1	Example of an ROC Curve. . . . .	54
3.2	EEG recordings for 91 subjects from the Frontal Lobe position of the brain. . . .	55
3.3	Raw and simple average EEG tracings for males and females. . . . .	55
3.4	Cross-validation plots for functional logistic regression of the EEG data. . . . .	56
3.5	Estimated functional parameter using $m = 15$ Fourier basis functions for the EEG data. . . . .	56
3.6	Predicted probabilities of being female, split by known sex, in the EEG data. . .	57

---

5.1	Cross-validation plot for basis size selection for Risk Category. . . . .	77
5.2	Parameter estimates from functional logistic regression for Risk Category. . . . .	78
5.3	Histogram of probabilities of a high risk birth split by observed response. . . . .	79
5.4	ROC Curves for the functional and simple logistic regression models for Risk Category. . . . .	79
5.5	Normal probability plot for the PDI scores at 18 months. . . . .	81
5.6	Cross-Validation plot for basis size selection for PDI, 18 months. . . . .	83
5.7	Functional time parameter for PDI, 18 months, and its effect on sample heart rates.	84
5.8	Time parameter estimates for PDI at 18 months. . . . .	84
5.9	Functional regression model diagnostics for PDI at 18 months. . . . .	85
6.1	Simulated data, its theoretical spectrum, and the known mean function. . . . .	100
6.2	Cross-validation plot for the simulated data. . . . .	101
6.3	Estimated mean functions, eigenvalues and first two eigenfunctions for the simu- lated data. . . . .	102
6.4	Cross-validation plot for the EEG recordings. . . . .	103
6.5	Estimated mean function, covariance function and first two eigenfunctions for the EEG data. . . . .	104
6.6	Estimated mean function, covariance function and first two eigenfunctions for the gait data. . . . .	105

# List of Tables

2.1	Some common kernel functions. . . . .	19
3.1	EEG Data: Summary of classifications for sex using the basis method for functional logistic regression . . . . .	58
5.1	Logistic regression summary for Risk Category. . . . .	76
5.2	Summary of classifications for Risk Category using logistic regression. . . . .	76
5.3	Summary of classifications for Risk Category using functional logistic regression. . . . .	77
5.4	Summary of classifications of Risk Category for the controls using the functional logistic model. . . . .	80
5.5	Regression summary for PDI at 18 months, using only scalar covariates. . . . .	82
5.6	Functional regression summary for PDI at 18 months. . . . .	82