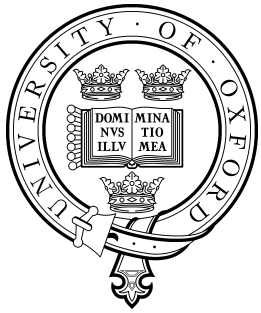
Mathematical Modelling, Forecasting and Telemonitoring of Mood in Bipolar Disorder



P.J. Moore

Somerville College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2014

This thesis is dedicated to
my wife Irene

Acknowledgements

The author wishes to acknowledge the valuable support and direction of his DPhil supervisors at the Oxford Centre for Industrial and Applied Mathematics (OCIAM), Max Little, Patrick McSharry and Peter Howell. Thanks also to John Geddes who has supported the project and provided access to mood data, and to Guy Goodwin, both of the Department of Psychiatry. Thanks to Will Stevens and Josh Wallace, who managed the data. Thanks to Karin Erdmann, my advisor at Somerville College. And thanks to my assessors during the project: Irene Moroz, Paul Bressloff and Gari Clifford whose comments in the intermediate examinations strengthened the work.

Particular thanks are due to Athanasios Tsanas, who has been a source of encouragement, ideas and discussion. Also to Siddharth Arora and Dave Hewitt for their valuable comments and advice. Thanks to all at Oxford who advised on the project: whenever I asked to meet, the answer was invariably positive. And thanks to OCIAM staff and students for providing a great academic environment. Finally, thank you to my wife Irene, who has been a constant source of support and encouragement, and to my parents, Bernard and Mary Moore.

Abstract

This study applies statistical models to mood in patients with bipolar disorder. Three analyses of telemonitored mood data are reported, each corresponding to a journal paper by the author. The first analysis reveals that patients whose sleep varies in quality tend to return mood ratings more sporadically than those with less variable sleep quality. The second analysis finds that forecasting depression with weekly data is not feasible using weekly mood ratings. A third analysis shows that depression time series cannot be distinguished from their linear surrogates, and that nonlinear forecasting methods are no more accurate than linear methods in forecasting mood. An additional contribution is the development of a new k-nearest neighbour forecasting algorithm which is evaluated on the mood data and other time series. Further work is proposed on more frequently sampled data and on system identification. Finally, it is suggested that observational data should be combined with models of brain function, and that more work is needed on theoretical explanations for mental illnesses.

Contents

1 Introduction **1**

1.1 The project . . . 1

1.1.1 Declaration . . . 2

1.1.2 Original contributions . . . 2

1.1.3 Thesis structure . . . 3

1.2 Psychiatry . . . 4

1.2.1 Psychiatric diagnosis . . . 5

1.2.2 Classification of psychiatric conditions . . . 6

1.3 Bipolar disorder . . . 7

1.3.1 Subtypes . . . 8

1.3.2 Rating scales . . . 9

1.3.3 Aetiology and treatment . . . 10

1.3.4 Lithium pharmacology . . . 11

1.4 Models . . . 12

1.4.1 Nonlinear oscillator models . . . 12

1.4.2 Computational psychiatry . . . 15

1.4.3 Data analyses . . . 19

1.4.4 Time series analyses . . . 20

2 Statistical theory **23**

2.1 Statistical models . . . 23

2.1.1 Statistics . . . 24

2.1.2 Probability . . . 25

2.1.3 Inference . . . 27

2.2 Supervised learning . . . 29

2.2.1 Regression . . . 29

2.2.2 Classification . . . 32

2.2.3 Model evaluation and inference . . . 33
2.3 Time series . . . 37
2.3.1 Properties of time series . . . 38
2.3.2 Stochastic processes . . . 38
2.3.3 Time series forecasting . . . 40
2.4 Gaussian processes . . . 44
2.4.1 Gaussian process regression . . . 45
2.4.2 Optimisation of hyperparameters . . . 47
2.4.3 Algorithm for forecasting . . . 47

3 Correlates of mood **49**
3.1 Mood data sets . . . 49
3.1.1 The Oxford data set . . . 50
3.2 Non-uniformity of response . . . 54
3.2.1 Measuring non-uniformity . . . 55
3.2.2 Applying non-uniformity measures . . . 62
3.2.3 Correlates of non-uniformity . . . 64
3.3 Correlates of depression . . . 67
3.3.1 Measuring correlation . . . 67
3.3.2 Applying autocorrelation . . . 68
3.3.3 Applying correlation . . . 70
3.4 Conclusions . . . 75

4 Forecasting mood **77**
4.1 Analysis by Bonsall *et al.* . . . 78
4.1.1 Critique . . . 79
4.2 Time series analysis . . . 80
4.2.1 Data selection . . . 80
4.2.2 Stationarity . . . 80
4.2.3 Detrended fluctuation analysis . . . 81
4.3 Forecasting . . . 83
4.3.1 In-sample forecasting . . . 83
4.3.2 Out-of-sample forecasting . . . 85
4.3.3 Non-uniformity, gender and diagnosis . . . 88
4.4 Conclusions . . . 91

5 Mood dynamics **95**
5.1 Analysis by Gottschalk *et al* . . . 96
5.1.1 Critique . . . 96
5.2 Surrogate data analysis . . . 97
5.2.1 Theory . . . 97
5.2.2 Data selection . . . 98
5.2.3 Autocorrelation . . . 99
5.2.4 Nonlinearity . . . 100
5.3 Forecasting . . . 105
5.3.1 Data selection . . . 106
5.3.2 Gaussian process regression . . . 106
5.3.3 Forecasting . . . 112
5.4 Conclusions . . . 113

6 Nearest neighbour forecasting **115**
6.1 K-nearest neighbour forecasting . . . 115
6.1.1 Method of analogues . . . 116
6.1.2 Non-parametric regression . . . 119
6.1.3 Kernel regression . . . 120
6.2 Current approaches . . . 121
6.2.1 Parameter selection . . . 121
6.2.2 Instance vector selection . . . 122
6.2.3 PPMD Forecasting . . . 123
6.3 Evaluation . . . 125
6.3.1 Lorenz time series . . . 126
6.3.2 ECG data . . . 128
6.3.3 Mood data . . . 131
6.4 Conclusions . . . 133

7 General conclusions **135**
7.1 Conclusions . . . 135
7.1.1 Time series properties . . . 135
7.1.2 Mood forecasting . . . 136
7.1.3 Mood dynamics . . . 137
7.2 Future work . . . 138
7.2.1 Mood data . . . 138
7.2.2 Techniques . . . 138

7.2.3 System identification 139
7.3 Summary 142

A Appendix A **143**
A.1 Statistics for time series and patient data 143
A.2 Statistics split by gender 144
A.3 Statistics split by diagnostic subtype 144
A.4 Interval analysis 150

Bibliography **152**

List of Figures

1.1 Van der Pol oscillator model for a treated bipolar patient 13
1.2 Lienard oscillator model for a treated bipolar patient 14
1.3 Markov model of thought sequences in depression 17

2.1 Bivariate Gaussian distributions 26
2.2 Examples of time series 37

3.1 Sample time series from two patients 50
3.2 Flow chart for data selection - main sets 51
3.3 Distribution of age and time series length 52
3.4 Scatter plot of time series length 53
3.5 Response interval medians and means 53
3.6 The effect of missing data on Gaussian process regression 54
3.7 Illustration of resampling 56
3.8 Effect of resampling on high and low compliance time series 57
3.9 Time series with compliance of 0.5 59
3.10 Time series with continuity of 0.8 60
3.11 Continuity versus compliance for patients 62
3.12 Continuity versus compliance for gender and diagnosis sets 63
3.13 Mean weekly delay in response 64
3.14 Variability of sleep against continuity 65
3.15 Correlograms for depression time series 69
3.16 Time series exhibiting seasonality of depression 69
3.17 Flow chart for data selection - correlation analysis 1 70
3.19 Autocorrelation for symptom time series 72
3.20 Flow chart for data selection - correlation analysis 2 73
3.21 Pairs of time plots which correlate 74
3.22 Pairwise correlations between time series. 74

4.1 Data selection for forecasting . . . 80
4.2 Change in median depression over the observation period . . . 81
4.3 Illustration of nonstationarity . . . 82
4.4 Scaling exponent of time series . . . 83
4.5 Relative error reduction of smoothing over baseline forecasts . . . 84
4.6 Forecast error against first order correlation . . . 85
4.7 Distribution of out-of-sample errors . . . 87
4.8 Proportion of imputed points . . . 89
4.9 Out-of-sample errors for resampled time series . . . 90
4.10 Relative error against non-uniformity measures . . . 90
4.11 Out-of-sample errors for male and female patients . . . 91
4.12 Out-of-sample errors for BPI and BPII patients . . . 92

5.1 Flow chart for data selection - surrogate analysis . . . 99
5.2 Depression time series . . . 100
5.3 Shuffle surrogates . . . 101
5.4 CAAFT surrogates . . . 102
5.5 Surrogate analysis of nonlinearity - 1 . . . 103
5.6 Surrogate analysis of nonlinearity - 2 . . . 103
5.7 Surrogate analysis of nonlinearity - 3 . . . 104
5.8 Surrogate analysis of nonlinearity - 4 . . . 104
5.10 Flow chart for data selection - forecasting . . . 106
5.11 Mood time series . . . 107
5.12 Sample draws from a Gaussian process . . . 107
5.13 Gaussian process forecasting . . . 111
5.14 Forecast error *vs.* retraining period . . . 111

6.1 A reconstructed state space . . . 118
6.2 K-nearest neighbour forecasting . . . 124
6.3 A reconstructed state space with weighting . . . 125
6.4 Attractor for PPMD evaluation . . . 126
6.5 Lorenz time series set for PPMD evaluation . . . 126
6.6 PPMD forecast method applied to the Lorenz time series . . . 128
6.7 ECG time series set for PPMD evaluation . . . 128
6.8 PPMD forecast method applied to an ECG time series . . . 130
6.9 Depression time series used for PPMD evaluation . . . 131
6.10 PPMD forecast method applied to an depression time series . . . 132

7.1 Cognition as multi-level inference 140

A.1 Distribution of mean mood ratings 145
A.2 Distribution of dispersion of mood ratings 145
A.3 Mean ratings for symptoms of depression 145
A.4 Time series age and length for males and females 146
A.5 Mean mania ratings for males and females 146
A.6 Standard deviation of depression for males and females 147
A.7 Symptoms of depression - females 147
A.8 Symptoms of depression - males 147
A.9 Time series age and length for BPI and BPII patients 148
A.10 Mean mania ratings for BPI and BPII patients 148
A.11 Standard deviation of depression for BPI and BPII patients 149
A.12 Symptoms of depression for BPI patients 149
A.13 Symptoms of depression for BPII patients 149
A.14 Analysis of gaps in time series 150
A.15 Distribution of response intervals 151

List of Tables

1.1 Diagnostic axes from the DSM-IV-TR framework 7
1.2 DSM-IV-TR bipolar disorder subtypes 9
1.3 Rating scales for depression and mania 10
1.4 Analyses of mood in bipolar disorder 19

2.1 Prediction using Gaussian process regression 48

3.1 Diagnostic subtypes among patients 51
3.2 Age, length and mean mood 52
3.3 Correlation between depression symptoms and continuity 65
3.4 Age, length and mean mood for depression symptom analysis . . . 71
3.5 Age, length and mean mood for time series correlation analysis . . . 73

4.1 Age, length and mean mood for selected time series 80
4.2 Out-of-sample forecasting methods 87
4.3 Out-of-sample forecasting results 88

5.1 Statistics for the eight selected time series 99
5.2 Statistics for the six selected time series 106
5.3 Gaussian process forecast methods 109
5.4 Likelihood for GP covariance functions 110
5.5 Forecast error for GP covariance functions 110
5.6 Forecast methods used 112
5.7 Forecast error for different methods 113
5.8 Diebold-Mariano test statistic for out-of-sample forecast results . . . 114

6.1 Validation error for the Lorenz time series 127
6.2 Validation error for an ECG time series 129
6.3 Out-of-sample errors for ECG data 130

6.4 Validation error for kernel variants on ECG data 131
6.5 Next step forecast errors for depression time series 133

List of Abbreviations

5-HT 5-hydroxytryptamine or serotonin, a neurotransmitter

AAFT Amplitude adjusted Fourier transform surrogate data

AR Autoregressive model

AR1 Autoregressive model order 1

AR2 Autoregressive model order 1

AIC Akaike information criterion

ARIMA Auto regressive integrated moving average

ARMA Auto regressive moving average

ASRM Altman self-rating mania scale

BN1 Mean threshold autoregressive model order 1

BN2 Mean threshold autoregressive model order 2

BNN Mean threshold autoregressive model order n

BON Mean threshold autoregressive model

BIC Bayesian information criterion

BP-I Bipolar I disorder

BP-II Bipolar II disorder

BP-NOS Bipolar disorder not otherwise specified

CAAFT Corrected amplitude adjusted Fourier transform surrogate data

DFA Detrended fluctuation analysis

DSM The Diagnostic and Statistical Manual of Mental Disorders

DSM-IV-TR DSM edition IV - text revision

DSM-V DSM edition V

ECG Electocardiogram

EMS European Monetary System

FNN Fractional nearest neighbour model

FFT Fast Fourier transform

GPML Gaussian processes for machine learning software

IAAFT Iterated amplitude adjusted Fourier transform surrogate data

ICD International Statistical Class^n of Diseases and Related Health Problems

IDS Inventory of Depressive Symptomatology

KNN K-nearest neighbour model

LOO Leave-out-one cross-validation technique

MA Moving average model

MAE Mean absolute error

MDP Markov decision process

MS-AR Markov switching autoregressive model

OCD Obsessive-compulsive disorder

PPMD Prediction by partial match version D

PPD PPMD using median estimator

PPK PPMD using distance kernel

PPT PPMD using time kernel

PSR LIFE psychiatric status rating scale

PST Persistence model

QIDS Quick Inventory of Depressive Symptomatology scale

QIDS-SR Quick Inventory of Depressive Symptomatology - self report scale

RMS Root mean square

RMSE Root mean square error

S11 SETAR model with 2 regimes and order 1

S12 SETAR model with 2 regimes and order 2

SETAR Self-exciting threshold autoregressive model

SSRI Selective serotonin re-uptake inhibitor

TAR Threshold autoregressive model

TIS Local linear model

UCM Unconditional mean model

1

Introduction

This chapter provides an introduction to the project and sets the context for the research. The context is given in terms of psychiatry, bipolar disorder and theoretical models. Psychiatric illness, diagnosis and classification are discussed. Bipolar disorder is described along with assessment methods, contributory factors and treatment. Theoretical models of the disorder are described in detail and then literature that is directly relevant to the study is reviewed.

1.1 The project

This project began when I was working part-time in the Department of Psychiatry to support my DPhil which was, to start with, on automatic speaker recognition. I was aware that the department had collected a large database of mood data and wondered about studying its properties. It was a comparatively rare database of mood time series, and there existed only a few relevant papers so after some discussion with my supervisors I embarked on the current study. I analysed the time series and published some valuable results both about the data and the techniques that I developed for the analysis [98][97][96]. However, as the work progressed the limits of having no control data and only weekly sampling of variables became increasingly clear. I tried to obtain other data sets but most were either unreleased academic collections or were commercially sensitive.

I began the project using observational data but I have found this insufficient to draw any deep inferences about the disorder. Part of the reason undoubtedly lies in limitations of the data, its frequency and range. However, I suggest that

even with a richer set of data, there will remain limits on what it can reveal. To make more progress in understanding mental illness the data must be combined with realistic models of brain function, yet we are experiencing a rapid increase in data at a time when psychiatry still has no coherent theoretical basis. A new approach to modelling psychopathology is the idea of a *formal narrative*, which is based on a generative model of cognition. Details of the approach are given in the section on Future work in Chapter 7. However, the focus of this thesis is on mood data and its analysis. The work presented below covers statistical analysis of the data, prediction and the techniques used for these tasks.

1.1.1 Declaration

The content of this thesis is my own. Where information has been derived from other sources, I have indicated this in the text. I often used the first person plural in the text but this is simply a stylistic choice.

1.1.2 Original contributions

- A statistical analysis of mood data was presented and findings made on correlates between symptoms and sampling uniformity. For example, patients whose sleep varies in quality tend to return ratings more sporadically. Measures of non-uniformity for telemonitored data were constructed for the analysis. This work is presented in Chapter 3.
- A feasibility study for mood prediction using weekly self-rated data was conducted. A wide variety of forecasting methods was applied and the results compared with published work. This study is given in Chapter 4.
- A study of mood dynamics in bipolar disorder was conducted and the results were compared with previously published work. I showed that an existing claim of nonlinear dynamics was unsubstantiated. This work is presented in Chapter 5.
- A novel k-nearest neighbour forecasting method was developed and evaluated on mood, synthetic and ECG data. A software kit is published on my website at www.pjmoore.net. This work is presented in Chapter 6.

1.1.3 Thesis structure

This chapter, **Chapter 1**, introduces the thesis and sets the context of the research. **Chapter 2** is a short introduction to statistical theory, time series analysis and forecasting. The body of research for the thesis is in the next four chapters, three of which extend analyses in journal papers.

Chapter 3 is about correlates of mood in a set of time series from patients with bipolar disorder and extends the analysis in the paper, *Correlates of depression in bipolar disorder* [98]. The Oxford mood data is introduced and its statistical qualities are described, including an analysis of sampling non-uniformity. Non-uniformity is handled in two ways. First by selecting appropriate methods for measuring correlation and spectra. Second by developing measures of non-uniformity for mood telemonitoring.

Chapter 4 addresses the question of whether mood in bipolar disorder can be forecast using weekly time series and extends the paper, *Forecasting depression in bipolar disorder* [97]. The Oxford time series are analysed for stationarity and roughness and a range of time series methods are applied. A critique is made of a paper by Bonsall *et al.* [11] suggesting that their models may have a poor fit to the data.

Chapter 5 applies nonlinear analysis and forecasting methods to a particular subset of the Oxford time series and extends the paper, *Mood dynamics in bipolar disorder* which is currently under review for the International Journal of Bipolar Disorders. A critique of Gottschalk *et al.* [55] is made: this paper reports chaotic dynamics for mood in bipolar disorder. Surrogate data methods are applied to assess autocorrelation and nonlinear dynamics. Linear and nonlinear forecasting methods are compared for prediction accuracy.

Chapter 6 presents a k-nearest neighbour forecasting algorithm for time series. Some theoretical background to k-nearest forecasting is given and in this context the new algorithm is described. The algorithm is then evaluated on synthetic time series, ECG data and the Oxford bipolar depression time series.

The final chapter **Chapter 7** covers general conclusions and future work. **Appendix A** gives statistical summaries for the Oxford mood data.

1.2 Psychiatry

Psychiatry faces an ongoing crisis. The debate occasionally rises into public consciousness, but it has a long history: the recent controversy following (and preceding) the publication of DSM-V[1] is the latest chapter in a history that goes back at least as far as the antipsychiatry movement in the 1960s. Criticisms of DSM-V have brought to a focus concerns that have been voiced before: the medicalisation of normal human experience, cultural bias and controversies over inclusion/exclusion of conditions. More fundamental concerns have also been raised about the nature of mental illness and the validity of diagnoses.

Within the specialty itself, some psychiatrists have defined and analysed the problems. Goodwin and Geddes [54] suggest that the reliance on schizophrenia as a model condition had been a mistake. Difficulties with delineating schizophrenia as a diagnosis and questions over its explanation have led to conceptual challenges. They argue that bipolar disorder would have made a more certain 'heartland' or core disorder because it is easier to define within the medical model and provides a clearer role for the specialty's expertise than does schizophrenia. More broadly, Craddock *et al.*[26] in a 'Wake-up call for British psychiatry' criticise the downgrading of medical aspects of care in favour of non-specific psychosocial support. They point out the uneasiness that colleagues feel in defending the medical model of care and the difficulty in continuing to use the term *patient*. This is commonly being replaced with *service user*, despite patients preferring the older description [88]. They note a tendency to characterise a medical psychiatric approach as being narrow, biological and reductionist.

Katschnig [75] observe six challenges, three internal to the profession and three from outside.

1. Decreasing confidence about diagnosis and classification
2. Decreasing confidence about therapies
3. Lack of a coherent theoretical basis
4. Client discontent
5. Competition from other professions
6. Negative image of psychiatry both from inside and outside medicine

Out of the six challenges to psychiatry listed by Katschnig, the lack of a coherent theoretical basis stands out as causal. Katschnig comments that psychiatry is split

[1] DSM is a diagnostic manual which is described in Section 1.2.1.

into many directions and sub-directions of thought. He says, 'Considering that a common knowledge base is a core defining criterion of any profession, this split is a considerable threat to our profession.' Psychiatry possesses no satisfactory explanations for schizophrenia, bipolar disorder, obsessive-compulsive disorder (OCD) nor other psychiatric conditions. And according to Thomas Insel, research and development in therapies have been 'been almost entirely dependent on the serendipitous discoveries of medications' [92].

The tone of debate is becoming increasingly negative: Kingdon [76] asserts that 'Research into putative biological mechanisms of mental disorders has been of no value to clinical psychiatry' while both White [135] and Insel [66] propose to regard mental disorders as brain disorders. And the arguments become polarised, with parties finding themselves cast at one end of a nature-nurture, biological-social or mind-brain spectrum.

1.2.1 Psychiatric diagnosis

Authoritative definitions of mental illness can appear to be imprecise. Many dictionaries or encyclopaedias employ the term *normal* (or *abnormal*) when referring to cognition or behaviour, and the term *mind* is often used. For example, the Oxford English Dictionary refers to *'a condition which causes serious abnormality in a person's thinking or behaviour, especially one requiring special care or treatment'*. This definition raises the question of what is normal thinking or behaviour, and how it relates to the context of that action. Another approach is to make an analogy with physical sickness and introduce the notion of distress: both mental and physical illnesses can cause pain. This still implies some kind of default state or health, presumably of the brain. But normal psychological function is harder to define in objective terms than normal physiological operation. Blood pressure, for example, can be given usual limits in terms of a standard physical measure, but it is more difficult to define limits on human behaviour.

In practical terms, the criteria for mental illness are defined by a manual. One such manual is *The Diagnostic and Statistical Manual of Mental Disorders (DSM)* [2] published by the American Psychiatric Association. It is commonly used in the US, the UK and elsewhere for assessing and categorising mental disorders. Publishing criteria does not, of course, solve the problems with defining mental illness, and there is continuing controversy over what should and should not be

included. It does, however, allow conditions to be labelled[2], and appropriate therapy to be given. And importantly, the use of accepted criteria facilitates research into specific conditions.

1.2.2 Classification of psychiatric conditions

Attempts to classify mental illness date back to the Greeks and before. The earliest taxonomies, for example the *Ayur Veda* [28], a system of medicine current in India around 1400 BC, were based on a supernatural world view. Hippocrates (460-377 BC) was the first to provide naturalistic categories [3]. He identified both *mania* and *melancholia*, concepts which are related to, though broader than the current day equivalents. The modern system of classification (or *nosology*) is based on the work of the German psychiatrist Emil Kraepelin (1856-1926). His approach was to group illnesses by their course[3] and then find the combination of symptoms that they had in common.

The first attempt at an international classification system was made in 1948 when the World Health Organisation added a section on mental disorders to the *Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death (ICD-6)* [139]. This section was not widely adopted and the United States in particular did not use it officially. An alternative was published in the US, the first edition of The Diagnostic and Statistical Manual of Mental Disorders *(DSM-1)*. Development of the ICD section on mental disorder continued under the guidance of the British psychiatrist Erwin Stengel, and this later became the basis for the second revision of the DSM [3]. Both texts continue to be developed, and while the latest revision of the ICD section *(ICD-10)* is more frequently used and more valued in a clinical setting, DSM-IV is more valued for research [91]. Having been through five revisions, the most commonly used version of the DSM was published in 2000, and is referred to as DSM-IV Text Revision *(DSM-IV-TR)*. A more recent version, DSM-V, was published in 2013.

1.2.2.1 DSM-IV-TR axes

The DSM-IV-TR provides a framework for assessment by organising mental disorders along five axes or domains. The use of axes was introduced in DSM-III and

[2]Labelling obviously has both benefits and drawbacks.

[3]The course of an illness concerns the typical lifetime presentation, such as the progression of the illness over time.

has the purpose of separating the presenting symptoms from other conditions which might predispose the individual or contribute to the disorder.

DSM-IV-TR Axis	*Disorder*
Axis I	Clinical Disorders
Axis II	Developmental and Personality Disorders
Axis III	General Medical Condition
Axis IV	Psychosocial and Environmental Factors (Stressors)
Axis V	Global Assessment of Functioning

Table 1.1: The five diagnostic axes from the DSM-IV-TR framework.

The DSM-IV-TR axes are summarised in Table 1.1. Axis I comprises specific clinical disorders, for example bipolar II disorder, that the individual first presents to the clinician. It includes all mental health and other conditions which might be a focus of clinical attention, apart from personality disorder and mental retardation. The remaining four axes provide a background to the presenting disorder. Axis II includes personality and developmental disorders that might have influenced the Axis I problem, such as a personality disorder. Axis III lists medical or neurological conditions that are relevant to the individual's psychiatric problems. Axis IV lists *psychological stressors* or stressful life events that the individual has recently faced: individuals with personality or developmental disorders are likely to be more sensitive to such events. Axis V assesses the individual's level of functioning using the Global Assessment of Functioning Scale *(GAF)*.

1.3 Bipolar disorder

Bipolar disorder is a condition affecting mood and featuring recurrent episodes of *mania* and *depression* which can be severe in intensity. Mania is a condition in which the sufferer might experience racing thoughts, impulsiveness, grandiose ideas and delusions. Under these circumstances, individuals are liable to indulge in activities which can be damaging both to themselves and to those around them. Depression is characterized by low mood, insomnia, problems with eating and weight, poor concentration, feelings of worthlessness, thoughts of death or suicide, a lack of general interest, fatigue and restlessness. Both states are characterized by conspicuous changes in energy and activity levels which are increased in mania and decreased in depression [49].

The frequency and severity of mood swings vary from person to person. Many

people with bipolar disorder have long periods of normal mood when they are unaffected by their illness while others experience rapidly changing moods or persistent low moods that adversely affect on their quality of life [71]. Although manic and depressive mood swings are the most common, sometimes *mixed states* occur in which a person experiences symptoms of mania and depression at the same time. This often happens when the person is moving from a period of mania to one of depression although for some people the mixed state appears to be the usual form of episode. Further, some sufferers of bipolar disorder experience a milder form of mania termed *hypomania* which is characterised by an increase in activity and little need for sleep. Hypomania is generally less harmful than mania and individuals undergoing a hypomanic episode may still be able to function effectively [68].

1.3.1 Subtypes

DSM-IV-TR defines four subtypes of bipolar disorder and these are summarised in Table 1.2. Bipolar I disorder is characterised by at least one manic episode which lasts at least seven days, or by manic symptoms that are so severe that the person needs immediate hospital care. In Bipolar II disorder there is at least one depressive episode and accompanying hypomania. The condition termed *cyclothymia* refers to a group of disorders whose onset is typically early, are chronic and have few intervening euthymic[4] periods. The boundary between cyclothymia and the other categories is not well-defined and some investigators believe that it is simply a mild form of bipolar disorder rather than a qualitatively distinct subtype.

Bipolar NOS is a residual category which includes disorders that do not meet the criteria for any specific bipolar disorder. An example from this category is of the rapid alteration (over days) between manic and depressive symptoms that do not meet the minimal duration criteria for a manic episode or a major depressive episode. If an individual suffers from more than four mood episodes per year, the term *rapid cycling* is also applied to the disorder. This may be a feature of any of the subtypes.

[4] *Euthymia* is mood in the normal range, without manic or depressive symptoms.

Subtype	*Characteristics*
Bipolar I Disorder	At least one manic episode which lasts at least seven days, or manic symptoms that are so severe that the person needs immediate hospital care. Usually, the person also has depressive episodes, typically lasting at least two weeks.
Bipolar II Disorder	Characterised by a pattern of at least one major depressive episode with accompanying hypomania. Mania does not occur with this subtype.
Cyclothymia	Characterised by a history of hypomania and non-major depression over at least two years. People who have cyclothymia have episodes of hypomania that shift back and forth with mild depression for at least two years.
Bipolar NOS	A classification for symptoms of mania and depression which do not fit into the categories above. NOS stands for 'not otherwise specified'

Table 1.2: DSM-IV-TR bipolar disorder subtypes.

1.3.2 Rating scales

Rating scales may be designed either to yield a diagnostic judgement of a mood disorder or to provide a measure of severity. The former *categorical* approach tends to adhere to a current nosology such as documented in DSM-TR-IV and consists of examinations administered by the clinician or *schedules*. Such diagnostic tools are important for determining eligibility for treatment and, for example, help from social services. Measurement of severity or *dimensional instruments* are important for management of a condition, and for research. Dimensional instruments may be administered by the clinician or the patient and are designed or adapted for either use. The two scales used in this study are described next, one measuring depression and the other mania.

A rating scale used for depression is the *Quick Inventory of Depressive Symptomatology - Self Report (QIDS-SR$_{16}$)* [115] which comprises 16 questions. This self-rated instrument has acceptable psychometric qualities including a high validity [115]. Its scale assesses the nine DSM-IV symptom domains for a major depressive episode, as shown in Table 1.3. Each inventory category can contribute up to 3 points and the maximum score for each of the 9 domains is totalled, giving a total possible score of 27 on the scale. Most scales for mania have been designed for rating by the clinician rather than for self-rating because it was thought that the condition would vitiate accurate self-assessment. However some self-rated

QIDS Category (depression)	*ASRM Category (mania)*
Sleep *(4 questions)*	Feeling happier or more cheerful than usual
Feeling sad	Feeling more self-confident than usual
Appetite/weight *(4 questions)*	Needing less sleep than usual
Concentration	Talking more than usual
Self-view	Being more active than usual
Death/suicide	
General interest	
Energy level	
Slowed down/Restless *(2 questions)*	

Table 1.3: Rating scales for depression and mania. The QIDS Scale for depression is shown in the left hand column. There is more than one question for domains 1, 3 and 9 and the score in these cases is calculated by taking the maximum score over all questions in the domain. The QIDS score is the sum of the domain scores and has a maximum of 27. The Altman self-rating mania scale is shown in the right hand column. In this case each question can score from $0-4$ giving a maximum possible score of 20.

scales for mania have been assessed for reliability (self-consistency) and validity (effectiveness at measurement) [1]. The *Altman Self-Rating Mania Scale (ASRM)* is comprised of 5 items, each of which may can contribute up to 4 points, giving a total possible score of 20 on the scale. For both depression and mania ratings, a score of 0 corresponds to a healthy condition and higher scores correspond to worse symptoms. The schema for mania is shown in Table 1.3.

1.3.3 Aetiology and treatment

The aetiology[5] of bipolar disorder is unknown but it is likely to be multi-factorial with biological, genetic, psychological and social elements playing a part [49]. Psychiatric models of the illness suggest a vulnerability, such as a genetic predisposition, combined with a precipitating factor which might be a life event or a biological event such as a viral illness. Treatment includes both psychological therapy and medication to stabilise mood. Drugs commonly used in the UK are *lithium carbonate*, *anti-convulsant medicines* and *anti-psychotics*. Lithium carbonate is commonly used as a first line treatment either on its own (*monotherapy*) or in combination with other drugs, for example the anti-convulsants *valproate* and *lamotrigine*. Anti-psychotics are sometimes prescribed to treat episodes of mania or hypomania and include *olanzapine*, *quetiapine* and *risperidone* [102].

[5] *Aetiology* refers to the cause of a disease.

The mood stabilising effects of lithium[6] were first noted by John Cade, an Australian psychiatrist [17]. Cade was trying find a toxic metabolite in the urine of patients who suffered from mania by injecting their urine into guinea pigs. He was using lithium only because it provides soluble compounds of uric acid, which he was investigating. The animals injected with lithium urate became lethargic and unresponsive to treatment, so he then tried lithium carbonate and found the same effect. Assuming that this was a psychotropic effect[7], Cade first tried the treatment on himself, then on patients. In all the cases of mania that he reported, there was a dramatic improvement in the patients' conditions. Applying the treatment to patients with schizophrenia and depression, he found that the therapeutic effect of lithium was specific to those with bipolar disorder [93].

Cade's results were published in the *Medical Journal of Australia* in 1949 but the adoption of lithium as a mood stabiliser was slow [17] [60]. Although it has been commonly used in the UK it found less acceptance in the US [41], and was not approved by the Food and Drug administration until 1970. Concerns remain about lithium's toxicity: its therapeutic index (the lethal dose divided by the minimum effective dose) is low, there are long-term side effects, and there is the possibility of rebound mania following abrupt discontinuation of treatment [23].

1.3.4 Lithium pharmacology

One view of bipolar disorder is as resulting from a failure of the self-regulating processes (or *homeostatic mechanisms*) which maintain mood stability [87]. Some evidence for the cellular mechanisms is derived from studies on the action of mood stabilisers. Lithium in particular has several actions: it appears to displace sodium ions and reduces the elevated concentration of intracellular sodium in bipolar patients. It also has an effect on neurotransmitter signalling and interacts with several cellular systems [137]. It is not known which, if any, of these actions is responsible for its therapeutic effect.

One hypothesis for the action of lithium in bipolar disorder has generated particular interest. In the 1980s the biochemist Mike Berridge and his colleagues suggested that the depletion of inositol is the therapeutic target [9]. Inositol is a naturally occurring sugar that plays a part in the phosphoinositide cycle which regulates neuronal excitability, secretion and cell division. Lithium inhibits an enzyme which is essential for the maintenance of intracellular inositol levels.

[6] Lithium carbonate is commonly referred to as 'lithium'.

[7] In retrospect, it is possible that the animals were just suffering from lithium poisoning [93].

Furthermore Cheng *et al.* [22] found evidence that the mood stabiliser valproic acid limits mood changes by acting on the same signalling pathway. The inositol depletion hypothesis for lithium is just one possible cellular mechanism for the therapeutic effect of mood stabilizers and remains neither refuted nor confirmed. However, this kind of hypothesis can be relevant to the mathematical modelling of treatment effects in bipolar disorder. Cheng *et al.* [22] use a physical analogy to explain mood control, suggesting that it is like the action of a sound compressor which limits extremes by attenuating high and amplifying low volumes to keep music at an optimal level. In modelling mood following treatment changes, it may be possible to incorporate such a mechanism and thereby improve the validity of the model.

1.4 Models

Attempts at modelling mood in bipolar disorder have been constrained by the scarcity of data in a form suitable for mathematical treatment. Suitability in this context implies a useable format – that is, numerical time series data – and a frequency and volume high enough for analysis. We first review two models that do not use observational data directly. Daughtery *et al.*'s [29] oscillator model uses a dynamical systems approach to describe mood changes in bipolar disorder. Secondly, the field of *computational psychiatry* [95] derives models using a combination of computational and psychiatric approaches. These fundamental modelling approaches can provide insights into the dynamics of bipolar disorder without assimilating data. We then turn to analyses that are based on mood data and summarise the kinds of analysis and the measurements that were applied. Finally we introduce two time series analyses of data [11][55] that are similar to those reported in this study.

1.4.1 Nonlinear oscillator models

Daughtery *et al.* [29] use a theoretical model based on low dimensional limit cycle oscillators to describe mood in bipolar II patients. This framework was intended to provide an insight into the dynamics of bipolar disorder rather than to model real data. However the authors intended to motivate data collection and its incorporation into the model, and their paper has inspired further work [94], [4]. Daughtery *et al.* model the mood of a treated individual with a van der Pol

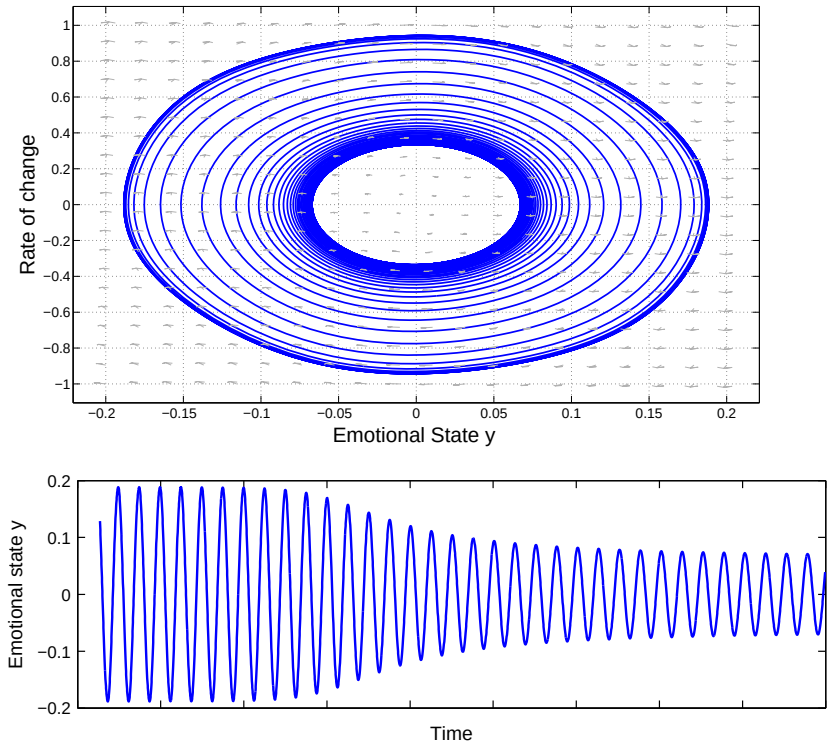


Figure 1.1: Van der Pol oscillator model for a treated bipolar patient with a forcing function of $g(y, \dot{y}) = \gamma y^4 \dot{y}$ modelling treatment. The upper panel shows a phase portrait and the lower panel shows a time plot. There are two limit cycles: the inner limit cycle is stable while the outer is unstable. As time increases the trajectory approaches the smaller, stable limit cycle. The amplitude of the mood oscillations in time thus decreases until it reaches a minimum level corresponding to that of a functional individual. The time plot shows a trajectory starting within the basin of attraction of the smaller limit cycle.

oscillator,

$$\ddot{y} - \alpha\dot{y} + \omega^2 y - \beta y^2 \dot{y} = g(y, \dot{y}) \tag{1.1}$$

where y denotes the patient's mood rating, $\dot{y}$ is the rate of change of mood rating with time, β determines amplitude and α, ω determine damping and frequency respectively. Treatment is modelled as an autonomous[8] forcing function $g(y, \dot{y}) = \gamma y^4 \dot{y}$ which represents all treatment, including mood stabilisers, antidepressants and psychological therapies. Since normal individuals normally experience some degree of mood variation, those individuals who suffer from bipolar disorder are defined as having a limit cycle of a certain minimum amplitude.

In an untreated state $g(y, \dot{y}) = 0$, the model oscillates with a limit cycle whose amplitude is determined by the parameters α and β. The application of treatment is simulated by applying the forcing function $g(y, \dot{y})$.

[8] *Autonomous* means that the forcing function depends only on the state variables.

The existence of limit cycles is analysed with respect to parameter values α, β and γ and the biologically relevant situation of two limit cycles is found when $\beta/\gamma < 0$ and $\beta^2 > 8\alpha\gamma > 0$. Parameter values of α = 0.1, β = -100 and γ = 5000 yield the phase portrait for a treated bipolar II patient shown in Figure 1.1. The smaller of the limit cycles is stable while the larger limit cycle is unstable. This leads to an incorrect prediction that if an individual remains undiagnosed for too long and their mood swings are beyond the basin of attraction of the smaller limit cycle, then they are untreatable.

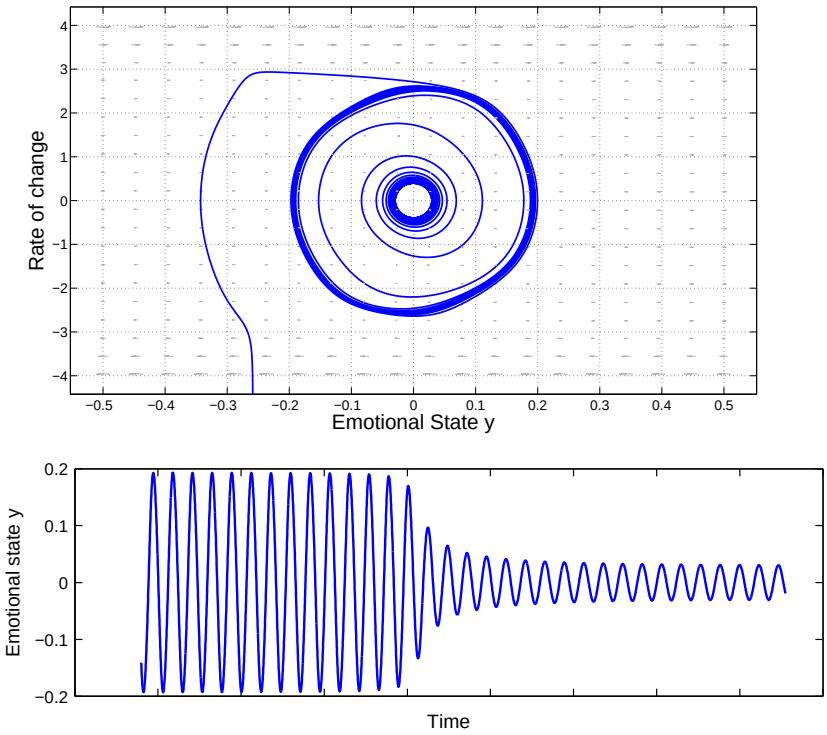


Figure 1.2: Lienard oscillator model for a treated bipolar patient with treatment $g(y, \dot{y})$ modelled by a polynomial in $\dot{y}$. The upper panel shows a phase portrait and the lower panel shows a time plot. There is a large stable limit cycle, a smaller, unstable limit cycle (which almost overlays it) and a small stable limit cycle within it. The smallest limit cycle represents the mood swings which remain under treatment. The largest stable limit cycle prevents a patient who is under treatment from having unbounded mood variations which could occur as a result of some perturbation. The time plot shows a trajectory starting within the basin of attraction of the smaller limit cycle.

A second model is introduced, based on the Lienard oscillator which has the form,

$$\ddot{y} + f(y)\dot{y} + h(y) = g(y, \dot{y}) \tag{1.2}$$

The forcing function $g(y, \dot{y})$ is configured according to whether a patient is treated

or untreated. For a treated patient, the model yields the phase portrait shown in Figure 1.2. In this case, there is a large stable limit cycle, an unstable limit cycle just within it and a smaller stable cycle inside that, representing the mood swings which remain under treatment. The larger limit cycle prevents a patient who is under treatment from having unbounded mood variations which could occur as a result of some perturbation.

Daughtery and his co-authors propose generalisations of their limit cycle modelling of bipolar disorder, including an examination of the bifurcations that occur in their models and an enhancement to model the delay in treatment taking effect. They suggest that employing their modelling framework along with clinical data will lead to a significantly increased understanding of bipolar disorder.

1.4.2 Computational psychiatry

Computational psychiatry is a subdiscipline which attempts to apply computational modelling to phenomena in psychology and neuroscience. For example, reinforcement learning methods are used to simulate trains of thought and to examine the effect of drugs on the model. First the theory for reinforcement learning is given followed by an example application.

1.4.2.1 Reinforcement learning

Reinforcement learning is a form of *unsupervised* machine learning. Supervised learning assumes the existence of examples provided by an external supervisor. Unsupervised learning attempts to find relationships and structure in unlabelled data. With reinforcement learning an *agent* tries a variety of actions and progressively favours those which subsequently give a reward. Modern reinforcement learning dates from the 1980s [128] and has inherited work from both the psychology of animal behaviour and from the problem of optimal control. One approach to the problem developed by Richard Bellman and others uses a functional equation which is solved using a class of methods known as *dynamic programming*. Bellman also introduced the discrete stochastic control process known as the *Markov decision process (MDP)* [8]. An MDP is in state s at time t, and moves randomly at each time step to state s' by taking action a and gaining reward $r(s,a)$. In a Markov decision process [128], a *policy* is a mapping from a state $s \in \mathcal{S}$ and an action $a \in \mathcal{A}(s)$ to the probability $\pi(s,a)$ of taking action a when in state s.

VALUE FUNCTIONS Most reinforcement learning algorithms are based on estimating *value functions*, which are functions of states or state-action pairs that estimate how beneficial it is for the process to be in a given state. The benefit is defined in terms of future reward or *expected return*. Since what the process expects to receive in the future depends on the policy, value functions are defined with respect to specific policies. The *value* $V^\pi(s)$ of a state s under a policy π is the expected return when starting in state s and following π thereafter. From [31] and [128, p134],

$$V^\pi(s) = \mathcal{E}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + .. \,|s_t = s\right] \tag{1.3}$$

$$= \mathcal{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,|s_t = s\right] \tag{1.4}$$

where r_t is the reward at time t, and $0 \leq \gamma \leq 1$ is a discount factor which determines the present rate of future rewards: a reward received k time steps in the future will be worth only γ^{k-1} times what it would be if it were received in the current time step. From (1.4) we see that,

$$V^\pi(s) = \mathcal{E}\left[r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,|s_t = s\right] \tag{1.5}$$

$$= \mathcal{E}\left[r_{t+1} + \gamma V^\pi(s_{t+1}) \,|s_t = s\right] \tag{1.6}$$

The method of *temporal difference prediction* allows the estimation of the change in value function without waiting for all future values of r_t. We define the temporal difference error δ_t as follows

$$\delta_t = r_{t+1} + \gamma \hat{V}^\pi(s_{t+1}) - \hat{V}^\pi(s) \tag{1.7}$$

where $\hat{V}^\pi(s)$ is an *estimated* value of state s under policy π. The algorithm for estimating state values then consists of incrementing the state values by $\alpha\delta_t$, where α is a learning rate parameter, as each new state is visited.

1.4.2.2 MODELLING DEPRESSION

The uncertainty over the action of lithium and other mood stabilisers was described in Section 1.3.4. In particular Cheng *et al.* [22] conjecture that valproic acid moderates mood by a bidirectional action on the phosphoinositide signalling pathway. A parallel can be seen with the role of serotonin (5-HT) in depression:

in both cases there is a therapeutic agent which has multiple, opponent effects which are not well understood. Serotonin is a neuromodulator[9] which plays an important role in a number of mental illnesses, including depression, anxiety and obsessive compulsive disorder. The role that serotonin plays in the modulation of normal mood remains unclear: on the one hand, the inhibition of serotonin reuptake is a treatment for depression; on the other, serotonin is strongly linked to the prediction of aversive outcomes. Dayan and Huys [31] have addressed this problem by modelling the effect of inhibition on trains of thought.

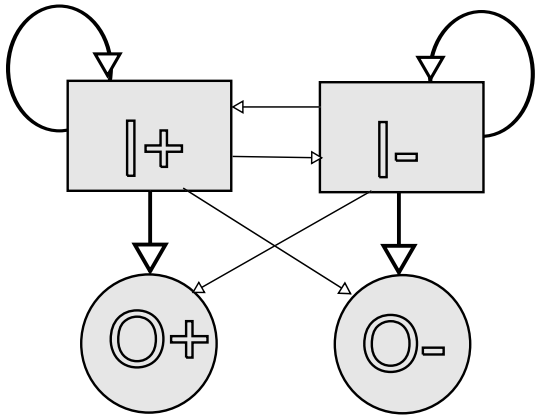


Figure 1.3: Markov model of thought from Dayan and Huys [31]. The abstract state space is divided into observable values of mood $\mathcal{O}$ and internal states $\mathcal{I}$. Transition probabilities are represented by line thickness: when the model is in an internal state, it is most likely to transition either to itself or to its corresponding affect state.

Figure 1.3 shows the state space diagram for the trains of thought. The model is a simple abstraction which uses four states: two are internal belief states $(\mathcal{I}_+, \mathcal{I}_-)$ and two are terminal affect states $(\mathcal{O}_+, \mathcal{O}_-)$ where the subscripts denote positive and negative affect respectively. The state $\mathcal{I}_+$ leads preferentially to the terminal state $\mathcal{O}_+$ and the state $\mathcal{I}_-$ leads preferentially to the terminal state $\mathcal{O}_-$. Transitions between states are interpreted as actions, which in the context of the study are identified with thoughts.

The internal abstract states $(\mathcal{I}_+, \mathcal{I}_-)$ are realised by a set of 400 elements each and the terminal states $(\mathcal{O}_+, \mathcal{O}_-)$ are realised by a set of 100 elements each. Each of the terminal states is associated with a value $r(s)$ where $(r(s) \geq 0, s \in \mathcal{O}_+)$ and $(r(s) < 0, s \in \mathcal{O}_-)$. The values are drawn from a 0-mean, unit variance Gaussian distribution, truncated about 0 according to which set $(\mathcal{O}_+, \mathcal{O}_-)$ it is assigned. In

[9] A neuromodulator simultaneously affects multiple neurons throughout the nervous system. A neurotransmitter acts across a synapse.

the model, the policy π^0 applies as follows: each element of $\mathcal{I}_+$ has connections to three randomly chosen elements also in $\mathcal{I}_+$, three to randomly chosen elements in $\mathcal{O}_+$ and one each to randomly chosen elements in $\mathcal{I}_-$ and $\mathcal{O}_-$. Similarly, each element of $\mathcal{I}_-$ has connections to three randomly chosen elements also in $\mathcal{I}_-$, three to randomly chosen elements in $\mathcal{O}_-$ and one each to randomly chosen elements in $\mathcal{I}_+$ and $\mathcal{O}_+$.

1.4.2.3 Modelling inhibition

The neuromodulator 5-HT is involved in the inhibition of actions which lead to aversive states, and this effect is represented by a parameter α_{5HT} which modifies the transition probabilities in the Markov model. The transition probability is given by

$$p_{5HT}(s) = \min(1, \exp(\alpha_{5HT} V(s))) \tag{1.8}$$

where $V(s)$ is the value of state s. High values of α_{5HT} will cause those trains of thought which lead to negative values of $V(s)$ being terminated as a result of the low transition probability. On the other hand, those trains of thought which have a high expected return (a positive value of $V(s)$) will continue. Thoughts that are inhibited are restarted in a randomly chosen state $\mathcal{I}$. When $\alpha_{5HT} = 0$, the estimated values match their true values within the limits determined by the learning error and the random choice of action. With α_{5HT} set to 20, the low valued states are less well visited and explored, leading to an over optimistic estimation for aversive states. In this case aversive states are less likely to be visited, leading to an increase in the average reward.

The experiment involves training the Markov decision process using a fixed level of α_{5HT} and manipulating this level once the state values are acquired. A model is trained with a policy $\pi_{\alpha_{5HT}}$, $\alpha_{5HT} = 20$ and the steady state transition probabilities are found for $\alpha_{5HT} = 0$ by calculating the probability of inhibition for each state. Two effects are observed. Firstly, the average value of trains of thought is reduced, because negative states are less inhibited. Secondly, the surprise at reaching an actual outcome is measured by using the prediction error

$$\Delta = r(s, a) - \hat{V}^{\alpha_{5HT}}(s) \tag{1.9}$$

for the final transition from an internal state $s \in \{\mathcal{I}_+, \mathcal{I}_-\}$ to a terminal affect state $s \in \{\mathcal{O}_+, \mathcal{O}_-\}$. It is found that the average prediction error for transitions into the negative affect states $\mathcal{O}_-$ becomes much larger when inhibition is reduced. These

results suggest that 5-HT reduction leads to unexpected punishments, large negative prediction errors and a drop in average reward. They accord with selective serotonin re-uptake inhibitors (SSRIs) being a first line treatment for depression and resolve the apparent contradiction with evidence that 5-HT is linked with aversive rather than appetitive outcomes.

1.4.2.4 Applicability

This application of reinforcement learning provides a psychological model for depression in contrast to data-driven models or methods based on putative underlying dynamics of mood. The power of the model is in suggesting possible mechanisms for mood dysfunction and in allowing experiments which could not easily be accomplished *in vivo*. The model could potentially be extended to bipolar disorder by extending the Markov model to include states for mania as well as depression. This would then allow experiments with mood stabilisers to be performed which would otherwise be impractical or unethical. However, for this study a new database of time series is available so we take a data driven approach to modelling.

1.4.3 Data analyses

Until recently most analyses of mood in bipolar disorder have been qualitative. Detailed quantitative data has been difficult to collect: the individuals under study are likely to be outpatients, their general functioning may be variable and heterogeneous across the cohort. The challenges involved in collecting mood data from patients with bipolar disorder has influenced the kinds of study that have been published. A survey of data analyses is given in Table 1.4

Authors	*Subjects*	*Analysis*	*Scale*	*Mood metrics*
Wehr(1979) *et al.* [134]	BP1/2 (n=5)	LG	Bunney-Hamburg	None
Gottschalk(1995) *et al.* [55]	BP (n=7)	TS	100 point analogue	Linear, nonlinear
Judd(2002) [71]	BP1 (n=146)	LG	PSR scales	Weeks at level
Judd(2003) *et al.* [70]	BP2 (n=86)	LG	PSR scales	Weeks at level
Glenn(2006) *et al.* [52]	BP1 (n=45)	TS	100 point analogue	Approx. entropy
Bonsall(2012) *et al.* [11]	BP1/2 (n=23)	TS	QIDS-SR	Linear, nonlinear
Moore(2012) *et al.* [97]	BP1/2 (n=100)	TS	QIDS-SR	Linear, nonlinear
Moore(2013) *et al.* [98]	BP1/2 (n=100)	TS	QIDS-SR	Linear, nonlinear

Table 1.4: Analyses of mood in bipolar disorder. *LG* denotes a longitudinal analysis and *TS* a time series analysis.

Detailed data has been taken from a small number of patients [55][134] or more general data from a larger number [70][71]. The article by Wehr and Goodwin [134] uses twice daily mood ratings for five patients. Judd [71] and Judd *et al.*[70] measure patients' mood using the proportion of weeks in the year when symptoms are present. This kind of measurement lacks the frequency and the resolution for time series analysis.

The paucity of suitable data has also constrained the kinds of measure used for analysis of mood. Until recently the primary measures used have been the mean and standard deviation of the ratings from questionnaires [110], although other measures have been used. Pincus [109] has introduced *approximate entropy* which is a technique used to quantify the amount of regularity and the predictability of fluctuations in time-series data. It is useful for relatively small datasets and has since been applied to both mood data generally [142] and to mood in bipolar disorder [52]; in the latter case, 60 days of mood data from 45 patients was used for the analysis. Gottschalk *et al.* [55] analysed daily mood records from 7 rapid cycling patients with bipolar disorder and 28 normal controls. The participants in this study kept mood records on a daily basis over a period of 1 to 2.5 years. The mood charts were evaluated for periodicity and correlation dimension and they inferred the presence of nonlinear dynamics, a claim that was later challenged by [79] and defended in [56].

1.4.4 Time series analyses

Two papers are directly relevant to this study because they address the dynamics of depression in bipolar disorder using time series analysis techniques. The first and more recent study was by Bonsall *et al.* [11] who applied time series methods to depression time series from patients with bipolar disorder. They used a data set similar to that in this project: time series from 23 patients monitored over a period of up to 220 weeks were obtained from the Department of Psychiatry in Oxford. The patients were divided into two groups of stable and unstable mood. The authors fitted time series models to the two groups and found that the two groups were described by different models. They concluded that there were underlying deterministic patterns in the mood dynamics and suggested that the models could characterise mood variability in patients.

Identifying mood dynamics is very challenging whereas empirical mood forecasting can be tested more easily. The effectiveness, or otherwise, of forecasting

using weekly mood ratings is an important question for management of the disorder. We address this question using out-of-sample forecasts to estimate the expected prediction error for depression forecasts and comparing the results with baseline forecasts. The results are given in Chapter 4, which includes a full review and discussion of the Bonsall *et al.* [11] paper.

The paper by Gottschalk *et al.* [55] was published in 1995 and dealt with 7 patients having a rapid-cycling course. Data was sampled on a daily basis in contrast to this study and to Bonsall *et al.* [11] where weekly data is used. Gottschalk *et al.* [55] used a surrogate data approach with nonlinear time series techniques to study the dynamics of depression. They also examined mood power spectra for patients and controls. They found a difference between the power spectral decay with frequency for patients and controls. They also found a difference in the correlation dimension for these two groups. From these findings, they inferred the presence of chaotic dynamics in the time series from bipolar patients. A full review and discussion of their conclusions, including the criticism by Krystal *et al.* [79], is given in Chapter 5.

2

Statistical theory

Introduction

This chapter provides a short introduction to statistical models, learning methods and time series analysis. The objective is to give some theoretical background to techniques that are applied in the thesis. The structure of this chapter is as follows. Section 1 covers statistical models and probability, including Bayes Theorem. Section 2 reviews the field of supervised learning including regression, classification and model inference, drawing especially on Hastie *et al.* [59]. Section 3 covers time series analysis and stochastic processes. Finally Section 4 covers Gaussian process regression.

2.1 Statistical models

A model is a representation which exhibits some congruence with what it represents. An important quality of a model is its *usefulness*, by contrast to its correctness. For example a tailor's dummy used for designing clothes is not anatomically correct except where certain sizes and proportions have to be true. Even these proportions are an abstraction from a diverse range of sizes in the population. Salient qualities and relationships are reflected in the model and detail is hidden. A *mathematical* model is expressed in mathematical language, for example in terms of variables and equations. A tailor's dummy is more convenient than a human in most cases, and in turn, a mathematical model is more convenient than a physical model. For this reason mathematical, or computational,

models are increasingly taking over from physical models in product design. Just as language allows debate about external referents, mathematical models facilitate the discussion of specific entities or phenomena. They can help in describing and explaining a system and they are used for predicting its behaviour. And importantly, mathematical models are *communicable* and so facilitate their criticism and in turn, their improvement.

All models encapsulate *assumptions* or invariant properties which are assumed to be true. Rigid assumptions might lead to poor representation, whereas relaxed assumptions can make a model less ambitious in its description. We can characterise both extremes of this range as *fundamental* and *formal* models. Fundamental models are based on well founded prior knowledge, such as the relation between current and voltage in an electrical circuit. In contrast, formal models are constructed from empirical data with more general, less ambitious assumptions. For example exponential smoothing is used in the prediction of time ordered data. It assumes exponentially decreasing influence of past values, but it does not encapsulate specific knowledge of a domain.

2.1.1 Statistics

Statistics is the principled management of data for the purposes of description and explanation. A statistical model is a formalism of the relationship between sets of data. Observations can be presented in two ways. The first, more modest approach is to document and describe them as they are, for example using points on a graph. The data may be scattered without any meaningful pattern, and with no obvious cause for their generation. However if they tend to lie on a straight line, it is reasonable to infer a linear relation between the two variables in the population from which the samples were drawn. The first approach of data exposition is classed as *descriptive statistics* and the second, *inferential statistics*.

Inference allows for prediction and simulation. If we observe two clusters of light in the sky each with a different centre and spread, we might infer that the sources are distinct in some way. We could then predict the likely source of a new observation by observing its location either side of a line between the clusters. Alternatively, if we go further and represent two stars directly, we can simulate observations. This distinction corresponds to the difference between a *discriminative model* and a *generative model* in statistics.

2.1.2 Probability

An important aspect of real data is uncertainty. A measurement of even a fixed quantity will fluctuate because there is error inherent in observation, and if the quantity varies the finite sample of observations leads to uncertainty. Probability theory is the calculus of uncertainty and it provides a structure for its management. We first state the rules of probability as follows.

The Rules of Probability.

$$\textbf{sum rule} \qquad p(X) = \sum_{Y} p(X, Y) \tag{2.1}$$

$$\textbf{product rule} \qquad p(X, Y) = p(Y|X)\, p(X) \tag{2.2}$$

□

We define the conditional probability as the probability of one event given another. It is especially important in statistical learning where we would like to find the source of an event given an observation. By combining the product rule for the two possible conditional probabilities, $p(Y|X)$ and $p(X|Y)$ we obtain *Bayes theorem*, an essential element of statistical learning,

Bayes Theorem.

$$p(Y|X) = \frac{p(X|Y)\, p(Y)}{p(X)} \tag{2.3}$$

□

2.1.2.1 Probability distributions

We can use histograms to visualise a distribution of values, and in the limit of an infinite number, we use a *probability density* for the variable. The density is expressed as a function of the value of the random variables, and is called a probability density function or pdf. An useful property of functions in this context is the average of its values weighted by their probability. This is called the *expectation* of a function and for a discrete distribution it is defined [10],

$$\mathcal{E}[f] = \Sigma_x p(x) f(x) \tag{2.4}$$

The *variance* of a function is given by,

$$var[f] = \mathcal{E}[(f(x)^2] - \mathcal{E}[f(x)]^2 \tag{2.5}$$

or for a random variable X,

$$var[X] = \mathcal{E}[X^2] - \mathcal{E}[X]^2 \tag{2.6}$$

For two random variables, the *covariance* is given by,

$$cov[X] = \mathcal{E}_{X,Y}[\ (X - \mathcal{E}[X])\,(Y - \mathcal{E}[Y])\ ] \tag{2.7}$$

where $\mathcal{E}_{X,Y}$ denotes averaging over both variables. The standard deviation σ_X is equal to the square root of the variance.

GAUSSIAN DISTRIBUTIONS An important distribution is the Gaussian distribution which for D variables, $x_1 .. x_D$, the Gaussian pdf has the form,

$$\mathcal{N}(\mathbf{y}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \tag{2.8}$$

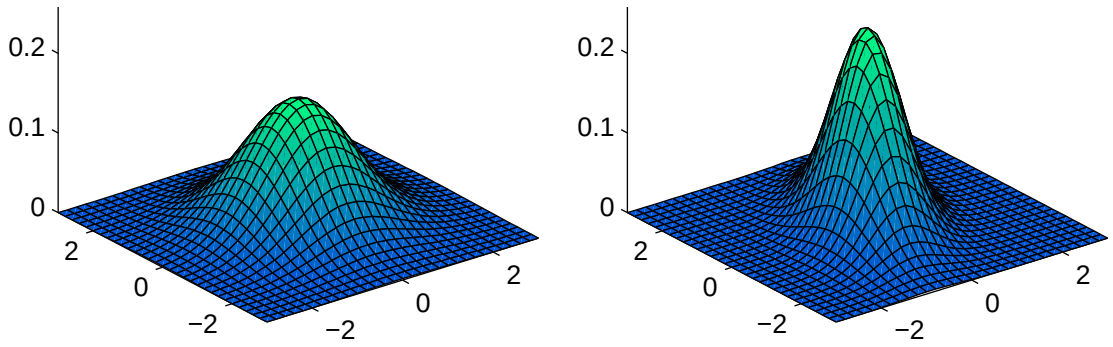


Figure 2.1: Joint distributions of two Gaussian random variables. (a) is a distribution with a unit covariance matrix so that there is no correlation between the two variables. (b) has off-diagonal terms in the covariance matrix giving rise to an skewed, elliptical form.

where Σ is the covariance between variables expressed as a D x D matrix and $\boldsymbol{\mu}$ is a D–dimensional mean vector. Two bivariate Gaussian distributions with different covariance matrices are illustrated in Figure 2.1. The multivariate Gaussian is used in Gaussian process regression which can be used for time series forecasting.

2.1.3 Inference

It was from a bivariate Gaussian distribution that Sir Francis Galton began to develop the idea of correlation between random variables. In 1885, he plotted the frequencies of pairs of childrens' and parents' height as a scatterplot and found that points with the same values formed a series of concentric ellipses [82]. Three years later he noted that the coefficient r measured the 'closeness of the co-relation'. In 1895, Karl Pearson developed the *product-moment correlation coefficient* [107], which is in use today,

> *Pearson's product-moment correlation coefficient.*
>
> $$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{[\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2]^{\frac{1}{2}}} \tag{2.9}$$
>
> □

where $\bar{X}$ denotes the average of X. This definition is based on a sample. For a population, the character ρ is used for the coefficient,

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{2.10}$$

So the correlation can be seen as rescaled covariance. The standardisation limits the range of ρ to the interval between -1 and +1. Correlation, like covariance, is a measure of linear association between variables, but its standardisation makes for easier interpretation and comparison. The definition of correlation is extended to time series in section 2.3 and its application to non-uniform time series is explained in Chapter 3.

2.1.3.1 Statistical testing

The correlation coefficient gives a standardised linear measure of association between two variables. However an association can arise by chance, so there is a need to quantify the uncertainty of the correlation coefficient. A *null hypothesis* is postulated, for example that the two random variables are uncorrelated. Assuming that the null hypothesis is true, the probability of seeing data at least as extreme as that observed, the *p-value*, is calculated. This value is then used to reason about the data: for example a value close to 1 shows little evidence against the null hypothesis.

The p-value itself is subject to some misinterpretation and misuse, for example Gigerenzer [51] asserts that hypothesis tests have become a substitute for thinking about statistics. Lambdin [81] makes a similar point and claims that psychology's obsession with null hypothesis statistical testing has resulted in 'nothing less than the sad state of our entire body of literature'. In this study, p-values are used, but we usually state them rather than relating them to a prescribed 5% level to imply a conclusion.

2.1.3.2 Kolmogorov-Smirnov test

For comparing distributions in forecasting we also use the *Kolmogorov-Smirnov test* [78]. The null hypothesis for this test is that the samples are drawn from the same distribution, and the test statistic is defined,

$$D_{m,n} = \sup_x |F_m^*(x) - G_n^*(x)| \tag{2.11}$$

where F_m^* and G_n^* are the empirical cumulative distributions of two sample sets, m and n are the sample sizes, and sup is the least upper bound of a set. The p-value is the probability of seeing data that is at least as extreme as that observed, assuming that the distributions are the same. For the Kolmogorov-Smirnov test, the test statistic $D_{m,n,p}$ is tabulated against sample sizes and p-values, so that the data is significant at level p for $D_{m,n} \geq D_{m,n,p}$.

2.1.3.3 Diebold-Mariano test

The Diebold-Mariano test [34] compares the predictive accuracy of two forecasting methods by examining the forecast errors from each model. The null hypothesis of the test is that the expected values of the loss functions are the same,

$$H_0 : \mathcal{E}[L(\epsilon_1)] = \mathcal{E}[L(\epsilon_2)] \tag{2.12}$$

where ϵ_1 and ϵ_2 are the forecast errors for each method. The Diebold-Mariano test statistic for one step ahead predictions is,

$$S_{DM} = \frac{\bar{d}}{\sqrt{\frac{var(d)}{T}}} \sim \mathcal{N}(0,1) \tag{2.13}$$

where d is $L(\epsilon_1) - L(\epsilon_2)$ and T is the number of forecasts. Since the statistic is distributed normally, the null hypothesis that the methods have equal predictive accuracy is rejected at the 5% level for absolute values above 1.96.

2.2 Supervised learning

This section introduces the field of statistical learning and draws on Hastie *et al.* [59] for structure and content. Statistical learning is about finding relationships between data. An important area involves the relationship between independent and dependent variables or *input* and *output* data. For example in spam classification, the input is the message and the output is classified as either spam or genuine email. In automatic speech recognition the input is a sound waveform and the output is text. The data can be *categorical* such as the colours red, green and blue, *ordered categorical*, for example, small, medium and large, or *quantitative* as with the real numbers.

The process of learning generally starts with *training data* which is used to create a model. When the training data is comprised of outputs Y associated with inputs X, then the process is known as *supervised learning* because the model can learn by comparing its outputs $f(X)$ with the true outputs Y. It can be seen either in terms of an algorithm which learns by example or as a function fitting problem. Models are often subdivided into two kinds: *regression*, when the output variables are quantitative and *classification* when the output variables are categorical.

2.2.1 Regression

One criterion for comparing $f(X)$ with outputs Y is the *residual sum of squares*,

$$RSS(f) = \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{2.14}$$

This is a popular criterion for regression problems, but minimising $RSS(f)$ does not uniquely define f. Hastie *et al.* [59, p33] define three approaches to resolving the ambiguity,

1. Use linear basis functions of the form $\sum \theta h(x)$, as in linear regression.
2. Fit f locally rather than globally, as for example in k–nearest neighbour regression.
3. Add a functional $J(f)$ that penalises undesirable functions. Regularisation methods such as Lasso, and Bayesian approaches fall into this category.

The discussion in Section 2.1 distinguished fundamental from formal models depending on the modelling assumptions. The contrast can be seen by comparing

two examples from approaches 1) and 2): linear fitting and k-nearest neighbour regression (kNN). A linear model is fit globally to the data using the *RSS* criterion to set its parameters. By contrast kNN does not assume linearity, so the model can mould itself to the data[1]. There is trade-off between fitting to the training data and generalising to new data. The dilemma can be interpreted in Bayesian terms (2.3) where we assume a prior form for the function, and update the prior with the training data. This kind of approach falls into category 3), and an example is that of Gaussian process regression, described in section 2.4.1.

2.2.1.1 Linear regression

Linear systems are characterised by the principle of superposition. That is, the response to a linear sum of inputs is equal to the linear sum of responses to the individual inputs. They have a number of advantages compared with nonlinear models in that there is a large body of knowledge to help with model choice and parameter estimation. They are conceptually simpler than their nonlinear counterparts and can have a lower risk of overfitting the data that they are trained on, compared with nonlinear models. An intrinsic disadvantage, though, is that real systems are often nonlinear - for example speech production has been shown to be nonlinear [84]. However, in practice linear models are often used as a convenient approximation to the real system.

A linear regression model assumes that the regression function $f(X)$ is linear. It explains an output variable Y as a linear combination of known input variables X with parameters $\boldsymbol{\beta}$ plus an error term ϵ. Following Hastie *et al.* [59, p44],

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon \tag{2.15}$$

If we assume that the additive error ϵ is Gaussian with $\mathcal{E}[\epsilon] = 0$ and $var(\epsilon) = \sigma^2$, then by minimising $RSS(f)$ we find,

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2) \tag{2.16}$$

The distribution of parameters $\boldsymbol{\beta}$ is multivariate normal, as illustrated in Figure 2.1. The convenience of a linear model with these assumptions becomes clear: the coefficients can be tested for statistical significance using a standardised form, the *Z-score*.

[1]When the neighbourhood in a local regression model covers the input space, the model becomes global.

The least squares estimates of the parameters β have the smallest variance among all unbiased estimates, but they might not lead to the smallest prediction error. Accuracy can be improved by shrinking or removing some parameters, and this leads to further advantages: for example a subset of inputs may be more interpretable and reveal the most important effects. One approach is to derive a subset of coefficients using a selection algorithm. A simple approach is simply to find $RSS(f)$ for all possible subsets, although other more sophisticated algorithms can be used. Subset selection discards variables, leading to a higher variance of the predicted function. An alternative is to use *shrinkage methods* which impose a penalty on the size of the coefficients. Two popular shrinkage methods are *ridge regression*, which adds a scaled $L2$ penalty to $RSS(f)$ and *Lasso*, which adds an $L1$ penalty.

2.2.1.2 Nonlinear regression

Linear models can be easier to interpret because the relation between inputs and outputs is simple to describe. A linear model might also be necessary if the sample size is small and when a more flexible model might overfit the data. This is liable to occur when the dimension of the input space is high and the model is complex. This section describes two general approaches to moving beyond linearity when it is appropriate. The first approach, linear basis expansion, breaks down the function $f(X)$ into multiple functions. The second approach, kernel smoothing, fits models to neighbourhoods of each training point x_i.

Basis expansions Where a more complex model is needed, the linear basis functions of the form $\sum \theta h(x)$ introduced in section 2.2.1 provide a way to move beyond linearity. In linear regression the basis functions are identity functions $h(x) = x$, but more generally they can be transformations of X, such as polynomials or logarithms. However, with the additional flexibility comes a need to control the complexity. Hastie *et al.* [59, p140] identify three approaches: restrict the number of basis functions, select the basis functions or regularise the coefficients.

In the context of least squares regression, subset selection is an example of a selection method, and Lasso a regularisation method. Piecewise polynomial regression is an example of the first method, where a fixed number of polynomials are fitted in each interval of the function domain X. When the intervals and nodes between them called *knots* are fixed, the piecewise polynomials are called

regression splines. Their parameters are the polynomial order, the number of knots and their location. An alternative method is to use X to define the knot positions. We consider the following penalised residual sum of squares [59, p151],

$$PRSS(f,\lambda) = \sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int (f''(t))^2 dt \tag{2.17}$$

where λ is a *smoothing parameter* which governs the degree of smoothing. The first term is a data fit term and the second a complexity term which penalises curvature in the function. The solution which minimises (2.17) is a cubic spline with knots at X. Complexity can vary from $\lambda = \infty$, which is a linear least squares fit, to $\lambda = 0$ which is unconstrained. The smoothing parameter λ can be chosen using *cross-validation*, a method which estimates the prediction error for different values and is described in section 2.2.3.

KERNEL SMOOTHING Kernel smoothing applies a weighting function $K_\lambda(x_0, x_i)$ or *kernel* to each training point x_0 and neighbouring point x_i where λ determines the width of the neighbourhood. One example is k–nearest neighbour regression, which in its simplest form uses a rectangular weighting function to select from X, and takes the average output in this region. This kernel leads to a discontinuous function because the averages change abruptly between kernels. A smoother function can be obtained by letting the weight decay smoothly with distance, rather than using a cut-off value. The parameter λ then controls smoothness, as in the case of basis expansions. However, locally weighted averages can show bias at the edge of the domain so a natural extension is *local linear regression*, which fits a linear model to each neighbourhood by minimising over the entire domain X, [59, p195],

$$\min_{\alpha(x_0),\beta(x_0)} \sum_{i=1}^{N} K_\lambda(x_0, x_i)[y_i - \alpha(x_0) - \beta(x_0)x_i]^2 \tag{2.18}$$

The function is then estimated as $\hat{f}(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0)x_0$. Both k–nearest neighbour regression and local linear regression are used for forecasting mood in Chapter 5.

2.2.2 CLASSIFICATION

The squared error loss function seems inappropriate for classification, particularly when the output categories are not ordered in any way. A more suitable loss function is the *zero-one* loss function, where the supervisor gives zero for a correct

classification and one for an incorrect classification. Under this rule, the optimal classifier is,

$$\hat{G}(x) = \underset{k \in 1,2,..K}{\operatorname{argmax}}\ p(G = g_k | X = x) \tag{2.19}$$

where K is the number of categories in the output variable G. This classifier is known as the *Bayes classifier* and is equivalent to choosing the most probable class from the conditional distribution $p(G|X)$. For example G might be a Boolean class representing spam or no spam, and X the features derived from an email message. In simulated problems $p(G|X)$ may be known exactly, but in real world applications there is a need for a model derived from output-input pairs. The joint probability $p(G, X)$ can be decomposed in two different ways using the product rules (2.2),

Classification approaches.

generative approach $$p(G, X) = \underline{p(G)\, p(X|G)} \tag{2.20}$$

discriminative approach $$p(G, X) = p(X)\, \underline{p(G|X)} \tag{2.21}$$

□

In both cases, the objective is to find the conditional probability $p(G|X)$. In the *generative* approach to classification, the full joint density is estimated and $p(G|X)$ is found using Bayes theorem (2.3). The *discriminative* approach models the conditional distribution $p(G|X)$ directly from the data. Dawid [30] calls the generative approach the *sampling paradigm* and the discriminative approach the *diagnostic paradigm*.

2.2.3 Model evaluation and inference

The data that is used for optimising the parameters is called the training set because it is from this data that the values are learned. Good performance on the training data does however not guarantee good performance on new data. For example, a polynomial of high degree will fit closely to the training data and follow the noise values but without representing the underlying function. Using least squares evaluation, the error will decrease with the polynomial order, but a model with a zero training error is overfitted to the training data, and will not generalise well.

One approach to avoid overfitting is to evaluate the model on a *validation set* that is separate from the data that the model has been trained on. To provide independence from the validation set, a third *test set* is sometimes used for the final evaluation. Where there is only one model or the data set is relatively small this process may be unrealistic, and *cross-validation* can be used. This technique involves partitioning the available data into S groups, using $S-1$ of the groups as training data and making an evaluation using the remaining set. The procedure is repeated for all S choices of the held-out group. When data is very scarce, it is possible to use $S = N$, where N is the total number of data points giving a *leave-one-out* technique. A similar approach is that of *bootstrapping* in which, rather than using subsets, new data sets are derived from the original set by sampling with replacement.

2.2.3.1 Model errors

We wish to estimate the error associated with a model in order to choose the best one and to estimate its error on new data. We define the following kinds of error for this purpose using a loss function L [59, ch7],

1) The *training error* is the average loss over the training set,

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i)) \tag{2.22}$$

2) The *expected prediction error* is an estimate of the prediction error on a new sample set and is defined,

$$Err = \mathcal{E}[L(Y, \hat{f}(X))] \tag{2.23}$$

Cross-validation estimates the expected prediction error directly, but the process can prove to be computationally expensive. Since the training error decreases with model complexity there is a need for a performance measure that compensates for over-fitting. Various information criteria have been proposed which penalise model complexity in order to estimate the expected prediction error. One example is the *Akaike information criterion (AIC)* [59, p230],

$$AIC = -\frac{2}{N}\, loglik + 2\frac{d}{N} \tag{2.24}$$

where *loglik* is the maximum log-likelihood of the data given the parameters, d is the number of parameters and N is the number of points in the training set. A related criterion is the *Bayesian information criterion* [59, p233] which has the form,

$$BIC = -2\,loglik + log(N)d \tag{2.25}$$

Both AIC and BIC are used when the model is fitted using a maximum likelihood estimator[2]. The AIC is derived by considering the bias of the training error $\overline{err}$ in estimating an ideal in-sample error. The BIC arises from the probability of each model conditional on the data: the lowest BIC corresponds to the model with the highest posterior probability. A more general Bayesian approach to model selection is to find the posterior distribution over the models. This is described in section 2.2.3.3.

2.2.3.2 Maximum likelihood

The model errors in the previous section were defined using a loss function which for regression is usually $RSS(f)$. A more general approach is to consider a *likelihood function* $L(\mathbf{w}; Y)$ which is the probability of the data Y given parameters $\mathbf{w}$. It can be seen as mapping the non-random parameter variables $\mathbf{w}$ to the random variable Y. If we assume that the data that we are modelling has a Gaussian distribution with mean $f(t, \mathbf{w})$ at time t, and a constant variance σ^2 then,

$$L(\mathbf{w}; Y) \sim \mathcal{N}(Y|f(t, \mathbf{w}), \sigma^2) \tag{2.26}$$

Assuming that the data $\mathbf{y}$ are drawn independently from this distribution, then the likelihood function is given by,

$$L(\mathbf{w}; Y) = \prod_{n=1}^{N} \mathcal{N}(y_n|f(t_n, \mathbf{w}), \sigma^2) \tag{2.27}$$

The negative log likelihood function is given by,

$$-\log L(\mathbf{w}; Y) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} (f(t_n, \mathbf{w}) - y_n)^2 + N \log(\sigma) + \frac{N}{2} \log(2\pi) \tag{2.28}$$

Maximising the likelihood function is equivalent to minimising the negative log likelihood because the negative logarithm is a monotonically decreasing function. The maximum likelihood solution for the model parameters $\mathbf{w}_{ML}$ is then given by minimising the first term on the right hand side of (2.28). Maximising likelihood, under the assumption of a Gaussian noise distribution, is therefore equivalent to minimising a sum of squares error function.

[2]Maximum likelihood estimation is explained in the next section.

2.2.3.3 Bayesian model comparison

The AIC and BIC information criteria take account of the *number* of model parameters and can be termed *penalized likelihood* approaches to model complexity. A more fully Bayesian approach takes account of the uncertainty in parameter *values*. Bayes' theorem takes the form,

$$p(\mathbf{w}|Y) = \frac{p(Y|\mathbf{w})\, p(\mathbf{w})}{p(Y)} \tag{2.29}$$

where $p(Y|\mathbf{w})$ is the likelihood function as before and $p(\mathbf{w})$ is the prior distribution of parameter values. In the classical or *frequentist* paradigm, the parameters $\mathbf{w}$ in the likelihood function are considered to be fixed, and estimates of the error in parameters are obtained by considering the distribution over different data sets of Y. Hence the maximum likelihood provides an *estimate* $\mathbf{w}_{ML}$ of the correct value of the parameters. From a Bayesian viewpoint, there is only one data set Y and the uncertainty in parameters is expressed through a probability distribution over the random variable $\mathbf{w}$.

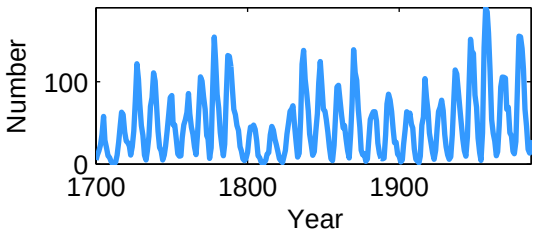
A Bayesian treatment involves integrating over all parameter values, so that the *model evidence* can be written in the form,

$$p(Y|\mathcal{M}_i) = \int p(Y|\mathbf{w}, \mathcal{M}_i)\, p(\mathbf{w}|\mathcal{M}_i)\, d\mathbf{w} \tag{2.30}$$

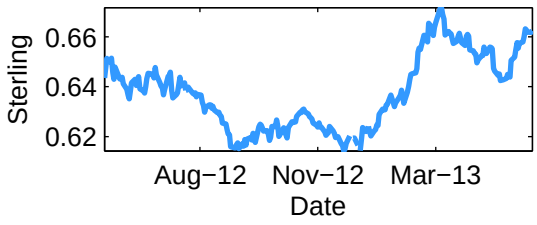
where Y is the observed data in the training set and $\mathcal{M}_i$ is the i-th model under consideration. The model evidence is also called the *marginal likelihood* because it can be seen as a likelihood function in the space of models where the model weights have been marginalised out. It allows the consideration of the likelihood of the *model*, rather than the *weights* of a model, to have generated the data. Following from this, a simple approach is to use the model with the highest marginal likelihood to make predictions. This technique is applied to the choice of hyperparameters in Gaussian process regression described in section 2.4.2.

2.3 Time series

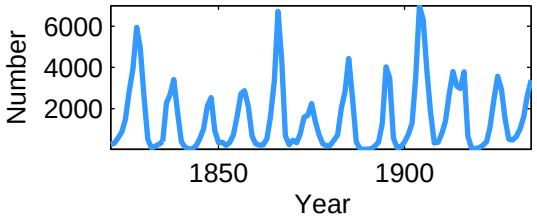
Time series analysis involves the description, explanation and prediction of observations taken sequentially in time [20]. Description implies the use of numerical and graphical descriptive statistics such as time plots. Time series analysis has become especially important in recent decades as more measurement data has become available. Analysis of this data can allow effective modelling and forecasting of significant events which can be of much value. Time series may be derived from a variety of applications, including biological signals, engineering applications, human related activities and natural processes.



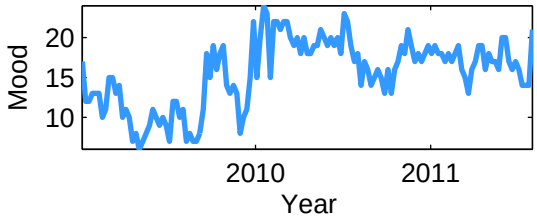
(a) Sunspot data



(b) Sterling - dollar exchange rate



(c) Lynx trapped in the Mackenzie River



(d) Mood variation with time

Figure 2.2: Four examples of time series from astronomical, economic, ecological and psychiatric domains. (a) shows the change of sunspot numbers with time, a graph that is known to cycle approximately on an eleven year period with a variation in the height of the peaks. (b) is a time plot of the dollar-sterling exchange rate from mid-2012 to mid-2013. It exhibits a variation over a period of days and a steady rise during the first quarter of 2013. (c) is a plot of the number of lynx trapped in the Mackenzie River district of Canada between 1821 and 1934. The time plot shows a seasonal peak, and troughs which drop close to zero. It was used in a paper by P.A.P Moran [99] and has since been re-analysed by a number of investigators. (d) is a graph of self-monitored mood for a patient with bipolar disorder. There is little obvious periodicity or other regular change with time in this case.

2.3.1 Properties of time series

Some example time plots are shown in Figure 2.2, illustrating a variety of qualities that can characterise a time series. Observations from natural sources inevitably include *noise* both because of naturally generated random behaviour and measurement error. A second quality of a series is the *level*, which can be defined as the expected value or mean of the series. The *trend* is the variation in level over a period; one way of measuring trend is to fit a linear model and investigate the slope: the exchange rate data in Figure 2.2(b) exhibits an upward trend during the first quarter of 2013. The *seasonality* of a time series is the tendency to repeat a pattern of a certain periodicity. Seasonality is apparent in both the sunspot data in Figure 2.2(a) and the lynx trapping data in Figure 2.2(c).

2.3.2 Stochastic processes

A time series can be understood as a realisation of a *stochastic process*, which is defined as an ordered collection of random variables defined at a set of time points [20]. Formally, it may be represented $\{Y(t) : t \in T\}$ where $Y(t)$ is defined on a probability space $(\Omega, \mathcal{F}, P)$. The random variables are indexed by time t and map the sample space Ω to a set S. The *index set* T is usually either a positive real value $T = [0, \infty)$ or a discrete value $T = \{0, 1, 2..\}$. Set S is called the *state space* and similarly is continuous ($S = \mathbb{R}$) or discrete ($S = \mathbb{Z}$).

The tools of statistical modelling can be used, with the constraint that there is usually only a single observation at time t. In time series modelling we do not usually consider the joint distribution of the responses but instead use the *moments* of the process[3].

Functions of a stochastic process.

$$\textbf{Mean} \qquad \mu(t) = \mathcal{E}[Y(t)] \tag{2.31}$$

$$\textbf{Variance} \qquad \sigma^2(t) = var[Y(t)] \tag{2.32}$$

$$\textbf{Autocovariance} \qquad \gamma(t_1, t_2) = \mathcal{E}[\ [Y(t_1) - \mu(t_1)]\,[Y(t_2) - \mu(t_2)]\ ] \tag{2.33}$$

□

[3]Gaussian process modelling is an exception. Since the multivariate Gaussian distribution is completely described by its mean and covariance, it follows that second-order stationarity implies strict stationarity for Gaussian processes.

A time series is referred to as *stationary* if the statistical properties do not change over time. It is described as *strongly stationary* if the joint distribution of $Y(t_1), ..., Y(t_k)$ is the same as $Y(t_1 + \tau), ..., Y(t_k + \tau)$ for all $\tau > 0$. A weaker form, *second order stationarity*, is defined as a process with a constant mean μ and whose autocorrelation function depends only on τ, which is called the *lag*. We can then define the autocovariance as a function of τ,

Functions of a second-order stationary process.

Autocovariance $$\gamma(\tau) = cov(Y(t), Y(t + \tau)) \tag{2.34}$$

Autocorrelation $$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} \tag{2.35}$$

□

The *sample autocorrelation* function can be derived from the definition of the correlation coefficient in (2.9) with some simplifying assumptions. It is defined in Chapter 3 where the application to non-uniform time series is developed.

2.3.2.1 Noise processes

Noise is a factor in most time series and is generated both by the underlying process, for example thermal noise, and by measurement error. If the noise contains equal power for all ranges of frequency it is referred to as *white noise* by analogy with white light which contains all visual frequencies. White noise has zero mean and is uncorrelated in time, and as such it is a stationary process. An example of naturally generated white noise is thermal noise that can be generated by a resistive electronic component,

While white noise has a flat power spectrum, *pink noise* has an equal amount of noise power in each octave. That is, the power spectral density of pink noise is proportional to the reciprocal of the frequency, $P(f) \propto f^{-1}$. Natural pink noise sources include electronic devices, such as carbon resistors, and semiconductors. In this case, it is referred to as *flicker noise*. There is even evidence that long sequences of music exhibit pink noise [133]. A third important kind of noise is *brownian noise* which has $P(f) \propto f^{-2}$. The different forms of noise can be distinguished using *detrended fluctuation analysis*, which is applied to the mood time series in Chapter 4.

2.3.3 Time series forecasting

Stochastic time series forecasting can be traced back at least to Yule [144] who postulated that an individual time series can be regarded as the realization of a stochastic process. The autoregressive (AR) and moving average (MA) models are based on this concept. In 1970, Box and Jenkins published the influential text *Time Series Analysis: Forecasting and Control*, in which they outlined a three stage, iterative process for model identification, estimation and verification [13]. ARIMA models and their variants have been widely applied because the behaviour of a diverse range of time series can be described using relatively few parameters.

Exponential smoothing methods[4] originated in the post-war decades with the work of Brown [16], Holt [64] and Winters [138]. They were, until the 1980s, often considered to be a collection of ad hoc techniques for univariate extrapolation but have since their inception been progressively underpinned by a statistical foundation. Muth [101] demonstrated that simple exponential smoothing (SES) gave the optimal forecast for a random walk with additive noise. Box and Jenkins [13] showed that simple exponential smoothing was equivalent to an ARIMA(0,1,1) model. Work on exponential smoothing has continued, including taxonomies of methods, and state space interpretations [65].

Regime switching models, such as self-exciting threshold AR (SETAR) have been introduced by Tong [131] and, in the context of the current application, were applied to bipolar depression data by Bonsall *et al.* [11]. A review of linear and nonlinear forecasting methods is given by De Gooijer and Hyndman [32], and in De Gooijer and Kumar [33]. We now briefly describe the structure of each model used in the experiments in this chapter.

2.3.3.1 Autoregressive models

A regression model expresses the response variable y_t as the sum of a function f called the *regression function* and an error term,[5]

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-r}) + \epsilon_t \tag{2.36}$$

If the regression function can be written as a linear combination of the regression coefficients, the model is said to be a linear regression model. The autoregressive (AR) model explains the response vector of a time series in terms of its past

[4] Exponential smoothing is explained in Section 2.3.3.2.

[5] The form of model given here is an additive error regression model.

values, rather than separate predictor variables. This method was introduced by Yule [144] in 1927 in a paper in which he applies the model to sunspot numbers. An AR model can be written,

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + z_t \tag{2.37}$$

where $\{Z_t\}$ is purely random process and p is the order of the model, which is referred to as an *AR(p)* model.

Autoregressive models are commonly used with a moving average *(MA)* model which is defined,

$$y_t = \sum_{i=0}^{q} \beta_i z_{t-i} \tag{2.38}$$

where q is the order of the model, which is referred to as an *MA(q)* model. For an MA process, the response is the weighted sum of past random errors and there is no memory of past value.

The autoregressive and moving average models can be combined by addition to create an autoregressive moving average *(ARMA)* model,

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=1}^{q} \beta_i z_{t-i} + z_t \tag{2.39}$$

ARMA models are specified by the orders of the AR and MA model respectively and denoted *ARMA(p,q)*. ARMA modelling makes an assumption of weak stationarity whereas some time series exhibit trends[6]. One approach to a time varying population mean is to detrend the data by taking the difference between samples. This process is a differentiation of the original series which gives rise to the name autoregressive integrated moving average *(ARIMA)* model.

Regime shifting models Linear models such as the autoregressive model described above cannot capture some of the sudden changes that may occur in mood time series. To allow greater flexibility, we use *regime shifting* models which allows for separate parameterization in each model state or *regime*. One approach is to incorporate one or more *thresholds* into the model; when a threshold is reached, the model parameters are changed from one set to another set which is trained to the new conditions. This *threshold autoregressive (TAR)* approach was taken by Bonsall *et al.* [11] in which a linear model was fitted above and below the mean

[6]Trends in time series are explained in Section 2.3.

of the time series. This kind of model is in effect locally linear because a linear model is used in each regime.

The *Self-exciting Threshold Autoregressive (SETAR)* model was introduced by Tong [131] to extend the concept of the AR model to multiple regimes with the regime controlled by a threshold variable. A SETAR model is defined,

$$y_t = \sum_{i=1}^{p} \alpha_i^j y_{t-i} + z_t^j \tag{2.40}$$

where α_i^j is the i-th parameter under the j-th regime. The regime j corresponds to an interval on the real line, chosen by the value of a threshold variable u_t such that $r_{j-1} < u_t < r_j$ and $-\infty = r_0 < r_1 < \ldots < r_k = +\infty$. The threshold variable u_t is set equal to a past value of the time series which is used to detect the change of regime. Such a model of autoregressive order p and with k regimes is written *SETAR(k,p)*.

2.3.3.2 Exponential smoothing

Exponential smoothing refers to a class of methods which are used for forecasting with time series. The idea appears to have originated in 1944 with a US Navy operations research analyst, Robert G. Brown [48]. In the 1950s he extended his methods to include terms for trend and seasonality and he applied the approach to forecasting the demand for military hardware. The idea is that a forecast value is a weighted combination of past observations, with more recent observations being given more weight. The name derives from the fact that the weights decrease exponentially with the lag of the observation.

Forecasting with exponential smoothing The simplest form of exponential smoothing takes a previous forecast and adjusts it with a forecast error. If we denote the time series value as y_t, the existing forecast of that value as $\hat{y}_t$, then the forecast for the next period is,

$$\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t) \tag{2.41}$$

where α is a constant between 0 and 1. Hence,

$$\begin{aligned} \alpha = 1 &\Rightarrow \hat{y}_{t+1} = y_t \\ \alpha = 0 &\Rightarrow \hat{y}_{t+1} = \hat{y}_t \end{aligned} \tag{2.42}$$

When α is close to 1, a greater weight is placed on recent observations and the smoothing or dampening effect is fast. When α is close to 0, smoothing is slower, and past values have more influence. Equation (2.41) may be rewritten as,

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t \tag{2.43}$$

By repeated substitution of the previous forecast value, it becomes,

$$\begin{aligned} \hat{y}_{t+1} = \; & \alpha y_t + \alpha(1-\alpha)y_{t-1} + \\ & \alpha(1-\alpha)^2 y_{t-2} + ... + \\ & \alpha(1-\alpha)^{t-1} y_1 + (1-\alpha)^t \hat{y}_1 \end{aligned} \tag{2.44}$$

This form of the forecasting equation shows that $\hat{y}_{t+1}$ is the sum of a weighted average of past observations and an initial estimate $\hat{y_1}$ scaled by $(1-\alpha)^t$. The weights are in proportion to the geometric progression $\{1, (1-\alpha), (1-\alpha)^2..\}$ which is a discretized exponential decay.

Choice of initial condition The final term $(1-\alpha)^t \hat{y}_1$ in (2.44) is determined by the choice of the initial estimate $\hat{y}_1$, the length of the series t and the value of α. When t is large and α is not close to zero, this product tends towards zero, but for small values of t and α it becomes significant. Several heuristic approaches have been proposed for the initial estimate, for example using either the first data point or the mean of the training set as estimates. More sophisticated approaches are described by Hyndman *et al.* [65]. In the experiments described later, the unconditional mean is used as the initial condition.

2.3.3.3 K-nearest neighbour models

The models described so far use linear combinations of past values as estimators for the mean of the forecast distribution. The threshold models are exceptions to some degree, although even these models are piecewise linear. Linear models assume a linear relationship between the predictors and the regression function $f(x)$. Nearest neighbour methods attempt to approximate the function directly from the training data by using a *neighbourhood*[7]. Nearest neighbour forecasting for time series has commonly been applied in econometrics where it has been used for forecasting gross national product [5] and exchange rates [42]. A full explanation of k-nearest neighbour models is given in Chapter 6, where the method is explored in more detail.

[7] K-nearest neighbour models are also piecewise linear, though with many pieces.

2.4 Gaussian processes

In most forecasting methods the response values are assumed to be indexed at equally spaced intervals, and missing or irregular data is dealt with by a preprocessing step. Preprocessing methods include imputation for missing data and interpolation for uneven spacing, both of which rely on assumptions about the latent function. Gaussian process modeling does not require equally spaced intervals, and makes use of the time indices in learning the model.

The theory for Gaussian process models was established in the 1940s and in recent decades it has become popular in regression and machine learning. It was already well established in geostatistics, a field which involves the spatial distribution of ores and other materials, where the term used is *kriging*. O'Hagan [103] applies the model to several one-dimensional regression tasks. Williams and Rasmussen describe model optimisation [112], and Rasmussen relates Gaussian processes to other models that are commonly used in machine learning [111]. Applications in biomedicine have included heart rate inference [126], neuroimaging inference [118] and neonatal seizure detection [40].

The method uses a Bayesian nonparametric model and assumes a Gaussian prior distribution over the regression function. A clear introduction to Gaussian process regression is to be found in [113] and a fuller description is given in [112, p13] from which the following explanation and algorithm are derived.

> *Gaussian process.* A convenient way to model a time series $\{y_1, y_2, ..., y_N\}$ is as a set of random variables indexed by a set of real values $\{t_1, t_2, ..., t_N\}$ which in this application are time indices. We define a *Gaussian Process* as a collection of random variables, any finite number of which have a joint Gaussian distribution. □

The multivariate Gaussian distribution given by eqn. (2.8) applies to a random vector of dimension D. A Gaussian process is a generalisation of the Gaussian probability distribution to infinite dimensionality. It is specified by a *mean function* $m(\mathbf{t})$ and a *covariance function* $k(\mathbf{t}, \mathbf{t}')$,

$$m(\mathbf{t}) = \mathcal{E}[f(\mathbf{t})] \tag{2.45}$$

$$k(\mathbf{t}, \mathbf{t}') = \mathcal{E}[(f(\mathbf{t}) - m(\mathbf{t}))(f(\mathbf{t}') - m(\mathbf{t}'))] \tag{2.46}$$

This is written,

$$f(\mathbf{t}) \sim \mathcal{GP}(m(\mathbf{t}), k(\mathbf{t}, \mathbf{t}')) \tag{2.47}$$

If the index set is infinite, a Gaussian process can be interpreted as a distribution over functions. Each function generated by the process is the result of drawing from this distribution.

It is convenient to set the mean function (2.45) to be zero to simplify the analysis. In this case, the Gaussian process is defined by the covariance function (2.46) which specifies the covariance between pairs of output random variables. As such it determines the properties of the possible functions generated by the process. A commonly used covariance function is the squared exponential,

$$k(t_1, t_2) = \sigma^2 \exp\left(-\frac{(t_1 - t_2)^2}{2\,l^2}\right) \tag{2.48}$$

where the arguments t_1, t_2 are a pair of index values which are to be tested for covariance: when the input variables are close, the covariance is near to unity. The function has two parameters – called *hyperparameters* and collectively denoted $\boldsymbol{\theta}$ – which control the properties of functions generated by the process. The first hyperparameter σ^2 represents the signal variance while the second l is a length scale which represents how far along the axis the values become uncorrelated: a higher value corresponds to a more slowly varying signal.

2.4.1 Gaussian process regression

According to the definition of a Gaussian process, any subset of the function range has a joint Gaussian distribution. For a set of random variables which have a joint Gaussian distribution, the covariance matrix can be partitioned as follows,

$$p\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}\right) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} A & C^T \\ C & B \end{bmatrix}\right) \tag{2.49}$$

where A and B are the covariance matrices for $\mathbf{a}$ and $\mathbf{b}$ respectively, and C is the matrix of covariances between pairs of elements of A and B. An important property for forecasting is the conditional distribution of $\mathbf{b}$ given $\mathbf{a}$,

$$p\left(\mathbf{b}|\mathbf{a}\right) \sim \mathcal{N}(CA^{-1}\mathbf{a}, B - CA^{-1}C^T) \tag{2.50}$$

which is obtained using the product rule (2.2). The marginal distribution is,

$$p\left(\mathbf{a}\right) \sim \mathcal{N}(\mathbf{0}, A) \tag{2.51}$$

We rewrite (2.49) to represent the joint distribution between training $\mathbf{f}$ and test $\mathbf{f}_*$ outputs,

$$p\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix}\right) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{t},\mathbf{t}) & K(\mathbf{t},\mathbf{t}_*) \\ K(\mathbf{t}_*,\mathbf{t}) & K(\mathbf{t}_*,\mathbf{t}_*) \end{bmatrix}\right) \tag{2.52}$$

where $K(\mathbf{t},\mathbf{t})$, $K(\mathbf{t}_*,\mathbf{t}_*)$ are the covariance matrices for the training and test sets respectively and $K(\mathbf{t}_*,\mathbf{t})$ and $K(\mathbf{t},\mathbf{t}_*)$ denote the matrix of covariances between pairs of elements of the training and test sets. For clarity in the notation, we write K, K_{**}, K_* and K_*^T respectively for these matrices. The vector $\mathbf{f}$ represents a perfect observation: noise is assumed to be zero for now.

Equation (2.52) represents a joint prior distribution of functions from a Gaussian process prior. The posterior process is derived by conditioning the prior distribution (over functions) on the observations $\mathbf{f}$ and the time indices $\mathbf{t}$ and $\mathbf{t}_*$. From (2.50) and (2.52), the probability of the test set given the training set is,

$$p\left(\mathbf{f}_*|\mathbf{f},\mathbf{t},\mathbf{t}_*\right) \sim \mathcal{N}(K_*K^{-1}\mathbf{f}, K_{**} - K_*K^{-1}K_*^T) \tag{2.53}$$

Equation (2.53) allows for mean and variance forecasts of the output variables given the time indices of training and test sets and the output values of the training set.

Noise The training and test outputs in (2.52) include no noise terms whereas we have access only to observations $\mathbf{y}$ which include noise. We represent independent, identically distributed Gaussian noise ϵ with variance equal to σ_n^2 such that $\mathbf{y} = f(\mathbf{t}) + \boldsymbol{\epsilon}$. It is represented by adding the variance to diagonal elements of the covariance matrix K of the Gaussian prior distribution. The joint distribution is now between observed values $\mathbf{y}$ and the value $\mathbf{f}_*$ which is to be forecast so (2.52) becomes,

$$p\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix}\right) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K+\sigma_n^2 I & K_*^T \\ K_* & K_{**} \end{bmatrix}\right) \tag{2.54}$$

In this case the probability of the test set given the training set is,

$$p\left(\mathbf{f}_*|\mathbf{y},\mathbf{t},\mathbf{t}_*\right) \sim \mathcal{N}(K_*(K+\sigma_n^2 I)^{-1}\mathbf{y}, K_{**} - K_*(K+\sigma_n^2 I)^{-1}K_*^T) \tag{2.55}$$

The expectation and covariance are therefore,

$$\mathcal{E}\left[\mathbf{f}_*|\mathbf{y},\mathbf{t},\mathbf{t}_*\right] = K_*(K+\sigma_n^2 I)^{-1}\mathbf{y} \tag{2.56}$$

$$\mathcal{C}ov[\,\mathbf{f}_*] = K_{**} - K_*(K+\sigma_n^2 I)^{-1}K_*^T \tag{2.57}$$

Where there is a single test point with time index t_* the covariance between it and the training points is a vector which we denote $\mathbf{k}_*$ and the covariance for the test point is a scalar $k(t_*, t_*)$. In this case, the mean and variance are,

$$\bar{f}_* = \mathbf{k}_*^{\mathbf{T}}(K + \sigma_n^2 I)^{-1}\mathbf{y} \tag{2.58}$$

$$\mathcal{V}[f_*] = k(t_*, t_*) - \mathbf{k}_*{}^T(K + \sigma_n^2 I)^{-1}\mathbf{k}_* \tag{2.59}$$

It can be seen that (2.58) is a linear combination of the response variables $\mathbf{y}$.

2.4.2 Optimisation of hyperparameters

The marginal likelihood in (2.30) is rewritten with an explicit dependence on $\mathbf{t}$ and $\boldsymbol{\theta}$ with the function values $\mathbf{f}$ in place of weights $\mathbf{w}$,

$$p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{t}, \boldsymbol{\theta})\, p(\mathbf{f}|\mathbf{t}, \boldsymbol{\theta})\, d\mathbf{f} \tag{2.60}$$

We can obtain $p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta})$ directly by noting that $\mathbf{y} \sim \mathcal{N}(0, K + \sigma_n^2 I)$ and using (2.8),

$$\mathcal{N}(\mathbf{y}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})\right) \tag{2.61}$$

to obtain

$$\log\, p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2| - \frac{D}{2}\log 2\pi \tag{2.62}$$

where D is the length of the training data. A more complete Bayesian approach would integrate over all hyperparameter values to derive a probability conditioned only on the time indices but in practice this evaluation is difficult. Instead the likelihood $p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta})$ is maximised with respect to the hyperparameters $\boldsymbol{\theta}$ to provide an approximation.

2.4.3 Algorithm for forecasting

The algorithm described here first maximises the likelihood function of (2.62) to determine the values of the hyperparameters and noise variance and then uses these values in (2.58) and (2.59) to solve the mean and variance of the Gaussian process at the test point.

Table 2.1 summarises the steps needed for forecasting by Gaussian process regression. Since the prior process is assumed to have zero mean, the bias is first removed from the input time series. Using the unbiased signal, the hyperparameters $\boldsymbol{\theta}$ and noise variance σ_n^2 are then estimated by maximising the log marginal

likelihood $log\ p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ (2.62) for the given covariance function and the unbiased signal. The gradient of the marginal likelihood with respect to $\boldsymbol{\theta}$ can be found analytically and optimized using a conjugate gradient method. Next, equations (2.58) and (2.59) are used to find the forecast mean of the distribution at the test point t_*. The last step is to add the original signal bias to give the forecast value of the mood rating.

Step	*Estimated variables*
Subtract mean from the input	$\mathbf{y} = \mathbf{y}_{in} - \bar{\mathbf{y}}_{in}$
Optimise hyperparameters and noise	$\boldsymbol{\theta}, \sigma_n^2$ (eqn. 2.62)
Predict unbiased mean and variance	$\bar{f}_*, \mathcal{V}[f_*]$ (eqns. 2.58 and 2.59)
Add mean to estimate	$f_{estimated} = \bar{f}_* + \bar{\mathbf{y}}_{in}$

Table 2.1: Prediction using Gaussian process regression. The mean of the input series is subtracted from the series, the hyperparameters are optimised, the predictive equations applied, and the mean of the original series added to the result.

3

Correlates of mood

Introduction

The last chapter set the context of the research and provided a literature review. This chapter introduces the time series data which was collected by the Department of Psychiatry in Oxford. The data is presented using descriptive statistics, and the extent of non-uniformity in the time series is documented. Some new measures of missing and irregular responses are introduced and their correlation with mood and demographic qualities is investigated. Methods for analysing spectral properties of non-uniform time series are reviewed and used to detect seasonality in the depression time series. Finally the pairwise correlations between depression symptoms and correlations between individual patients are examined.

3.1 Mood data sets

The review of psychiatric literature in Chapter 1 illustrates how quantitative data on bipolar disorder has been hard to obtain until recently. The situation has begun to change in the last decade partly because of the widespread use of mobile phones and the internet as means of communication. There now exist mood monitoring websites, such as *Moodscope* (www.moodscope.com) which requires the user to complete a 20 point standard scale. Moodscope is intended for general use rather than specifically for patients with mood disorders. Other projects are designed for therapeutic purposes such as *Buddy* (www.buddyapp.co.uk), which

supports psychological therapy services. There are also databases of mood collected for academic studies, although their availability is at the discretion of the research investigators.

For this project data were collected as part of the *OXTEXT* (oxtext.psych.ox.ac.uk) programme which investigates the potential benefits of mood self-monitoring for people with bipolar disorder. OXTEXT uses the *True Colours* (truecolours.nhs.uk) self-management system for mood monitoring which was initially developed to monitor outpatients with bipolar disorder. Each week, patients complete a questionnaire and return the results as a digit sequence by text message or email. The Oxford mood monitoring system has generated a large database of mood time series which has been used for studying the longitudinal course of bipolar disorder [12] and for non-linear approaches to characterising mood by Bonsall *et al.* [11]. The data in the current study used the same telemonitoring techniques and the same rating scales as [11], but the studies are otherwise independent.

3.1.1 The Oxford data set

The OXTEXT data used in this study comprises time series from 153 patients with bipolar disorder who were monitored between 2006 and 2011. Some data relating to age, gender and diagnosis is available although it is incomplete. The mood data is returned approximately each week and comprises answers to standard self-rating questionnaires for both depression and mania. Details of the QIDS and ASRM rating scales used are given in Section 1.3.2. Examples of depression and mania time series for two patients are shown in Figure 3.1.

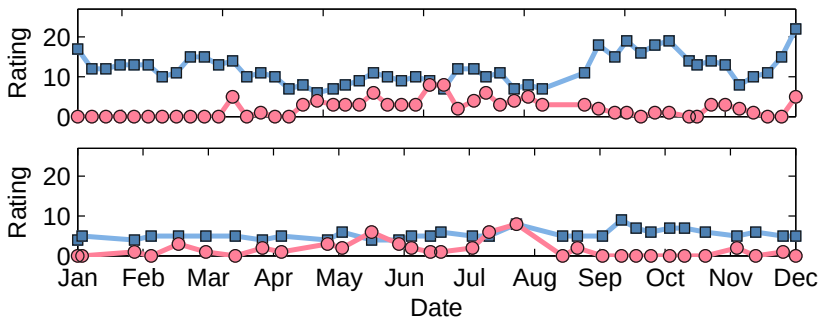


Figure 3.1: Sample time series from two patients for the year 2009. Each plot shows depression ratings (square markers) and mania ratings (circle markers). The top plot from the first patient is approximately uniform with some gaps and the lower plot is a non-uniform series.

3.1.1.1 Data selection

The initial set of 153 patients is first cleaned by removing repeated response values, that is those which share the same time stamp. These repeats arise when a patient resubmits a rating score either by mistake or in order to correct an earlier response. Assuming that earlier values are being corrected, we remove repeated responses by taking the most recent in the sequence. We then create Set A (n=93) with patients whose time series have at least 25 data points, or approximately six months duration. The joint distribution of gender and diagnosis for Set A is shown in Table 3.1.

	Diagnosis			
	Unknown	*BP I*	*BP II*	*BP NOS*
Female	22	21	16	1
Male	6	20	6	1

Table 3.1: Distribution of diagnostic subtypes for patients in Set A

Two further subsets are then created from Set A. The first subset has equal numbers of male and female patients all of whom have a diagnosis of BPI disorder and this labelled as Set G. The second subset has equal numbers of patients with BPI and BPII disorder and this is labelled Set D. Figure 3.2 illustrates the data selection process.

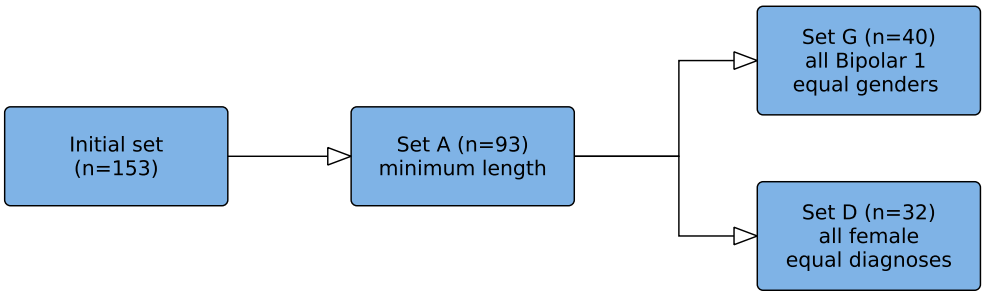
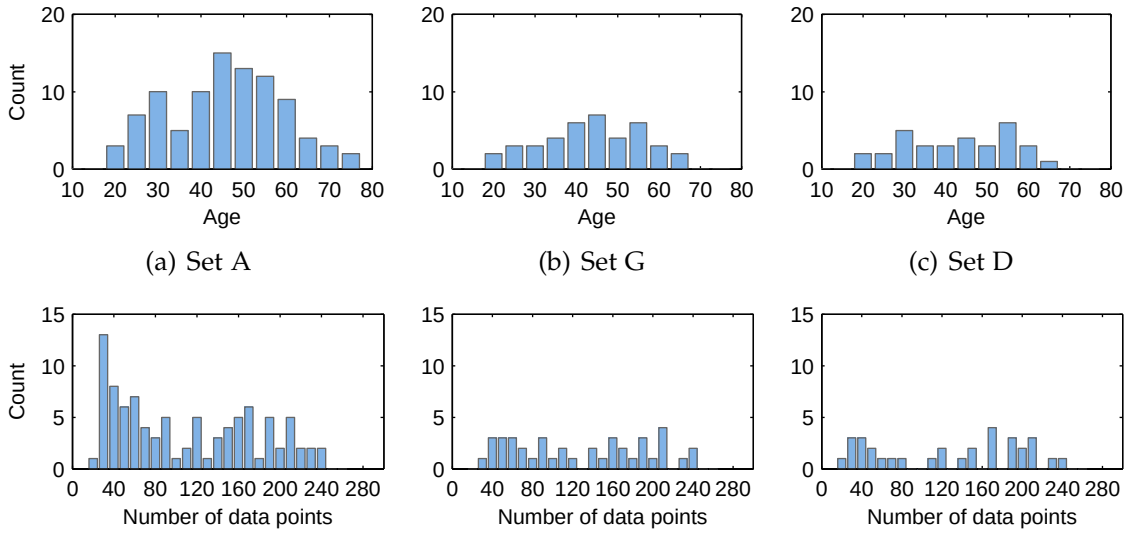


Figure 3.2: The Department of Psychiatry has provided a data set of 153 patients, each having time series of mania and depression. From this initial data set, Set A (n=93) of time series having a minimum length of 25 data points is selected. Two further subsets are then selected from Set A. Set G (n=40) has equal numbers of each gender, all with a diagnosis of Bipolar I disorder. Set D (n=32) has equal numbers of patients having Bipolar I and Bipolar II diagnoses, all of whom are female. The selection algorithm matches patients by time series length. Where no patient of matching length can be found, the search range is progressively widened until one or more matches is found.

3.1.1.2 Data description

The distribution of age and time series length for the Sets A, G and D is shown in Figure 3.3, and summaries of the distributions, including those for mean mood, are given in Table 3.2.



(a) Set A (b) Set G (c) Set D

(d) Set A (e) Set G (f) Set D

Figure 3.3: Distribution of age and number of data points for subsets of patients. Set A (n=93) comprises patients who have returned at least 25 data points, corresponding to a period of approximately six months. Set G (n=40) has equal numbers of male and female patients all with a diagnosis of Bipolar I disorder. Set D (n=32) has numbers of patients with a diagnosis of Bipolar I and Bipolar II all of whom are female.

		n	*Age in years*	*Length points*	*Depression QIDS*	*Mania ASRM*
Set A	*All*	93	45 (20)	95 (121)	6.3 (5.2)	1.9 (2.7)
	Female	60	45 (22)	92 (118)	7.2 (5.6)	2.0 (2.6)
	Male	33	45 (15)	111 (127)	5.0 (4.7)	1.6 (2.9)
Set G	*All*	40	45 (20)	128 (125)	6.5 (3.4)	1.8 (3.0)
	Female	20	40 (23)	105 (117)	7.1 (3.3)	2.2 (2.7)
	Male	20	45 (13)	159 (122)	4.9 (4.5)	1.6 (3.9)
Set D	*All*	32	45 (25)	142 (141)	7.7 (6.5)	2.0 (2.7)
	Bipolar I	16	45 (20)	128 (140)	7.1 (2.7)	2.2 (2.9)
	Bipolar II	16	43 (23)	160 (154)	8.7 (8.2)	1.9 (2.3)

Table 3.2: Median (interquartile range) of age, length and mean mood.

It can be seen that age and length are broadly similar across the sets. In Set G, female patients have a higher median depression compared with male patients. In Set D, patients with Bipolar I disorder have lower depression than those with Bipolar II disorder, and the distribution is less dispersed. Histograms for all the sets are given in Appendix A, Figures A.1 to A.13.

3.1.1.3 Interval statistics

We examine the characteristics of Set A by considering the time series length and the time intervals between responses for each patient. Figure 3.4 shows time series length as measured by the number of responses, plotted against the date that the patient's monitoring started.

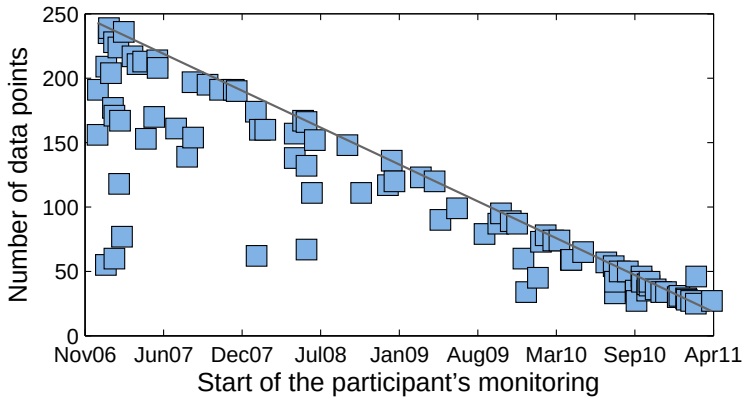


Figure 3.4: A scatter plot of time series length for all patients in Set A where length is the number of responses in the individual time series. The x-axis is the time when the patient enrolled in the monitoring system. A 'compliance' line shows a rate of one response per week.

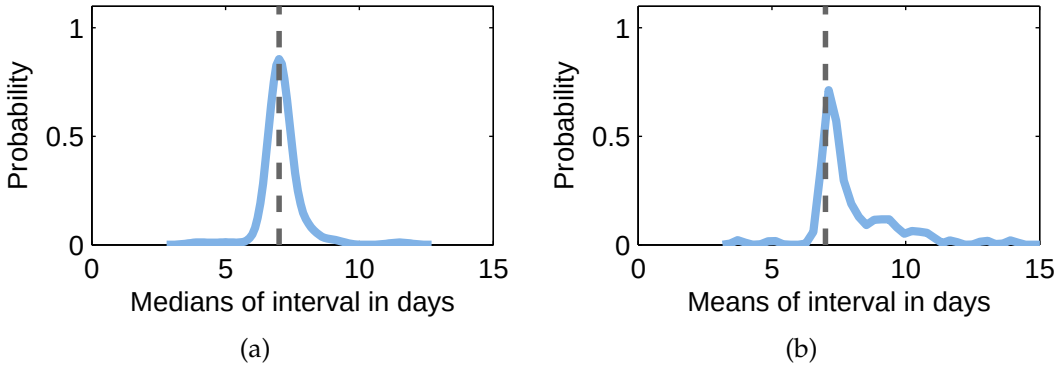


Figure 3.5: Estimated distribution of response interval medians and means for Set A (n=93). A vertical compliance line is drawn at an interval of 7 days.

We next consider the response interval, that is the time between receipt of ratings, for patients in Set A. Figure 3.5(a) shows a probability density estimate for the distribution of the medians of the response interval per individual. Some patients return ratings more frequently than each week – often these ratings are repeated values which have been resent – while others fail to respond for one week or for a longer period. The distribution is symmetrical with a low variance about the mode of seven days, a result that contrasts with Figure 3.4 which is asymmetrical about the compliance line. The reason for this apparent disparity is that some patients have long gaps in their time series, increasing the mean of the response interval but not its median. The distribution of mean intervals is given in Figure 3.5(b) which has a mode at 7.2 days. A more detailed analysis of gaps in the time series is presented in Appendix A.

3.2 Non-uniformity of response

The statistics just presented show that most patients return data on an approximately weekly basis and that some have long gaps in the time series. Time series methods described in standard texts such as Box and Jenkins [13] and Brockwell and Davis [14] assume uniformly spaced data, so any significant departure from uniformity must be addressed.

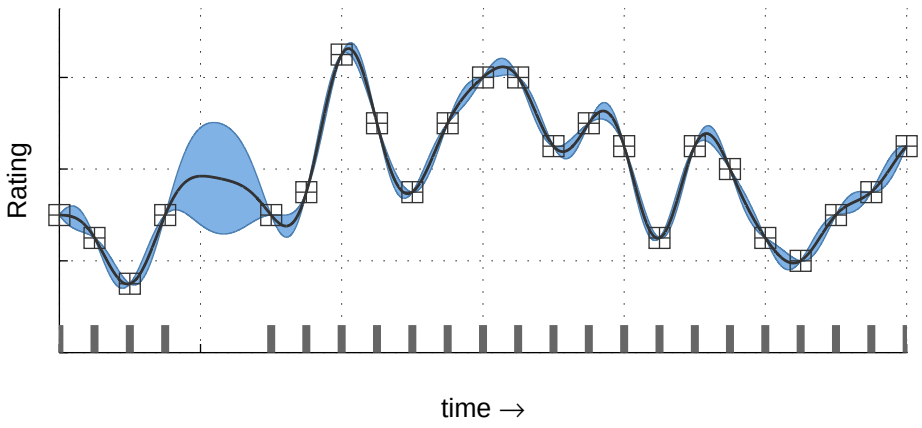


Figure 3.6: The effect of missing data on Gaussian process regression. Error bands for the predictive distribution are one standard deviation each side of the mean which is shown in black. The sampling comb is a regular seven days except for a two week gap. The signal shows a rise in variance in the gap because the responses are separated by much more than the length scale, which is estimated at six days. Here we have used a Gaussian likelihood which, combined with the Gaussian process prior gives a Gaussian process as the posterior. In this case, the posterior can be inferred using linear algebra.

One popular method for handling non-uniformity is to interpolate the data before applying the usual time series methods, but this can fail when there are a large number of gaps in the time series. An alternative is to use methods that take account of non-uniform intervals such as Gaussian process regression [112], which is illustrated in Figure 3.6.

Some work on managing non-uniform intervals assumes specific models, such as AR or ARMA [38] [69]. Eckner [36] provides further references and gives a general framework for handling non-uniformity. In this application we can make only weak assumptions about the time series dynamics partly because the time series are heterogeneous, as illustrated by the descriptive statistics in Section 3.1.1, but also because the dynamics are unknown. It is for this reason that Gaussian process regression is used in Chapters 4 and 5 as a forecasting method.

First we introduce an approach to measuring non-uniformity that is robust to long gaps in the time series. The measures are useful both for indicating data quality and for investigating non-uniformity as a correlate of mood. We then examine the correlation of non-uniformity measures with diagnostic, gender and mood data using the Oxford data set.

3.2.1 Measuring non-uniformity

We introduce two measures for quantifying missing and non-uniform responses in time series. The first which we call *compliance* measures the proportion of real observations in a time series that contains imputed values. The second measure, called *continuity*, quantifies the sampling regularity among those real observations. Both measures are easily derived from a uniformly sampled series with missing data, but here we start from an irregular series and assume that a response is valid for an interval rather than a single point in time. This condition applies when the observations are responses to a questionnaire about the week prior to the response, which is the case for the rating scales used here. We begin with the process of resampling the time series into an equivalent with uniform intervals.

3.2.1.1 Compliance

Figure 3.7 illustrates the resampling process where the actual observations are marked by diamonds and their period of validity is shown by a short horizontal line.

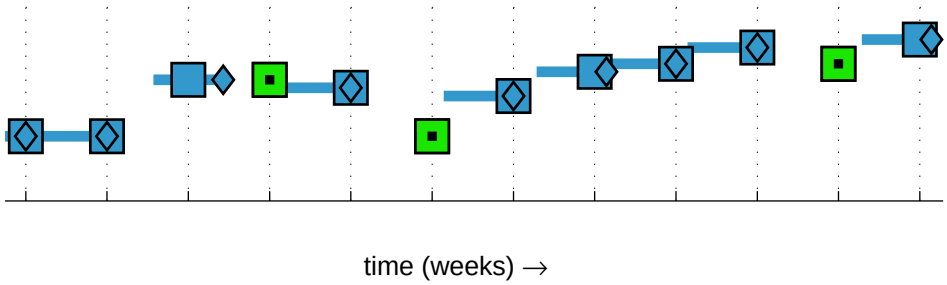


Figure 3.7: Illustration of resampling. Diamond markers represent the original, non-uniform time series and the horizontal lines to the left of each marker show the period over which the response is valid. Square markers represent the resampled series and those with a square central dot are imputed values. The X-axis or 'comb' shows the optimal weekday which when aligned with the original series gives the minimum total distance (deviation) of the sample time from the response time.

The optimal weekday w for the resampled time series is chosen to minimise the total deviation of the original responses from their corresponding resampled position on the X-axis or 'comb' of weekdays. The deviation in this case is the elapsed time to the first response within seven days.

The comb is then populated from the original series as follows. Starting from weekday w at the start, or the last instance before the start, of the time series, we record any response within seven days. We repeat the search from weekday w in the following week and continue until the last response of the time series is reached. If no response is found within seven days, a missing value is imputed. The imputed value itself is chosen for the purposes of illustration here and does not affect the non-uniformity measures.

Figure 3.8 shows the effect of resampling on two example series. Most of the original responses are not shifted while some are moved to an earlier time point and where this cannot be accomplished, an imputation is made.

We define compliance as the proportion of non-imputed values in the resampled time series. Imputations occur when a response is later than the sample period τ which in this application is equal to 7 days. Formally,

$$C_m = \frac{1}{N} \sum_{k=0}^{N-1} \Theta \left[\sum_{i=1}^{N'} \mathbf{1}[k\tau \leq t_i < (k+1)\tau] \right] \tag{3.1}$$

where C_m is compliance which lies between 0 and 1, τ is the uniform sample period and t_i is the i^{th} element of the time vector for the original series which has N' points. N is the number of points in the resampled series and is approximately equal to the number of weeks spanned by the original time series. The function

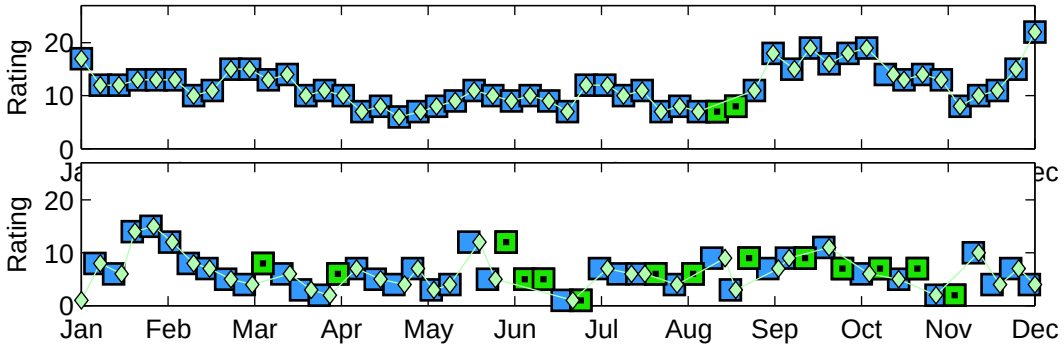


Figure 3.8: Effect of resampling on high and low compliance time series. The original responses are denoted by small diamond markers and the resampled series by a larger square border. Imputed values are shown with a central square dot. The top plot represents an approximately uniform original time series in which resampling preserves the time stamps of the original responses: most diamond markers are centered in the squares. The lower plot illustrates a non-uniform series where many responses are late, shown by diamond markers to the right of center, and some responses are missing, shown by central square dots.

Θ is equal to 0 if its argument is 0 otherwise it is equal to 1, and the indicator operator $\mathbf{1}$ has value 1 for a boolean argument of `true` and 0 for `false`.

As long as the original series covers all the new sample time points, there will be no imputations and the compliance is 100%. For example if responses are returned more often than every week, a uniform series may be derived by discarding some responses while still giving a compliance of 1. A non-uniform series may also exhibit full compliance as long as no response is more than six (more generally, $\tau - 1$) days late. However, longer gaps result in an imputed value being added to the uniform series and compliance being reduced. The measure thus penalises missing data but not additions or late returns.

3.2.1.2 Continuity

A low compliance implies that there is a high proportion of imputed points in the resampled series but it gives no information about their distribution throughout the observed responses. A second measure which we call *continuity* measures the connectedness of non-imputed responses in the resampled time series. To develop the measure, we examine the sequence of points in the resampled series and label them with a state indicator of `P` for imputed and `R` for not imputed. The number of sequential state changes `R` $\rightarrow$ `P` is a count of the discontinuity and we use the ratio of this count to $N_r - 1$, where N_r is the number of `R` states. A simple example

is the sequence RRR PPP R PPP R. Here there are 2 sequential changes of state from R to P out of a total of five R states giving a continuity of 2/4. The sequence RRRRR then has a continuity of 1, and the sequence RPRPR has a continuity of 0. In general we then have

$$C_t = 1 - \frac{1}{N_r - 1}\left(\sum_{k=1}^{N-1} \mathbf{1}[\,(w_k, w_{k+1}) = (\mathtt{R}, \mathtt{P})]\right) \tag{3.2}$$

where C_t is continuity, N is the length of the resampled series and $w_k \in \{\mathtt{R},\mathtt{P}\}$ is the state of the k^{th} data point. The minimum possible continuity occurs when the P states are distributed throughout the time series. In this case,

$$C_{t\,(min)} = 1 - \frac{N_p}{N - N_p - 1} \tag{3.3}$$

$$\approx \begin{cases} \frac{2C_m - 1}{C_m} & \text{if } C_m \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

for $N \gg 1$ where N_p is the number of P states. It can be seen from (3.4) that as the compliance approaches 1, the minimum possible continuity approaches the compliance.

3.2.1.3 Autocorrelation error

The potential effect of imputed values on the autocorrelation calculation is illustrated using an AR(1) process[1]. This model generates a sequence $\{Y_t\}$ from a scaled previous value plus a random term. An AR(1) model can be written,

$$Y_t = \alpha Y_{t-1} + Z_t \tag{3.5}$$

where $\{Z_t\}$ is a sequence of random variables with a mean of zero and variance σ_z^2. The following derivation of the autocorrelation[2] of the process is based on Chatfield [20, p 41]. By repeated substitution in (3.5), we obtain,

$$Y_t = Z_t + \alpha Z_{t-1} + \alpha^2 Z_{t-2} + \ldots \tag{3.6}$$

Then,

$$\mathcal{E}[Y_t] = 0 \tag{3.7}$$

$$var(Y_t) = \sigma_z(1 + \alpha^2 + \alpha^4 + \ldots) \tag{3.8}$$

[1] Autoregressive models are introduced in the context of forecasting in Section 2.3.3.1.
[2] Autocorrelation is defined in Section 2.3.2, equation (2.35).

The series converges for $|\alpha| < 1$ giving,

$$var(Y_t) = \sigma_y^2 = \frac{\sigma_z^2}{1-\alpha^2} \tag{3.9}$$

The autocovariance at lag 1 is given by,

$$\begin{aligned}
\gamma(1) &= \mathcal{E}[Y_t Y_{t+1}] && (3.10)\\
&= \mathcal{E}[(\sum_i \alpha^i Z_{t-i})(\sum_j \alpha^j Z_{t+1-j})] && (3.11)\\
&= \sigma_z^2 \sum_{i=0}^{\infty} \alpha^i \alpha^{i+1} && (3.12)\\
&= \alpha \frac{\sigma_z^2}{1-\alpha^2} \qquad \text{for } |\alpha| < 1 && (3.13)\\
&= \alpha \sigma_y^2 && (3.14)
\end{aligned}$$

The autocorrelation at lag 1 is then given by,

$$\rho = \alpha \tag{3.15}$$

We illustrate the effect of compliance and continuity on the calculation of the autocorrelation coefficient using an example time series.

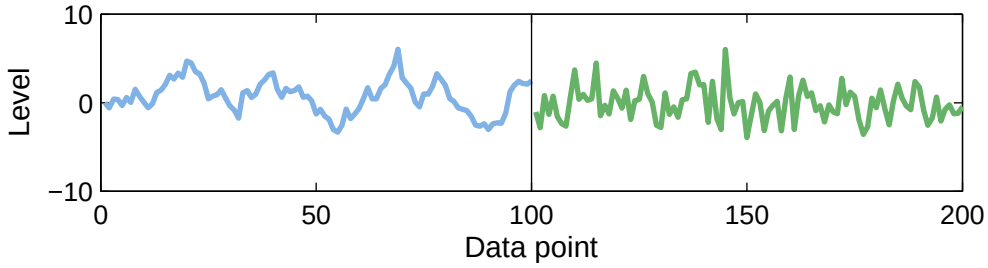


Figure 3.9: The time plot shows an instantiation of an AR(1) process from points 0 to 100 followed by a random time series with the same mean and variance as the AR(1) process at points 101 to 200. The value of α for the AR(1) process is 0.9. Considering the AR(1) process as observations and the random process as imputed points, the overall compliance of the time series is 0.5 and the continuity is close to 1.0.

Figure 3.9 shows an instantiation of an AR(1) process of length 100 points followed by a random process with a Gaussian distribution also of length 100 points giving a composite time series of length $N = 200$. The two segments of the time series have equal variance. If the AR(1) process is considered to be composed of N_r real observations while the $N - N_r$ random data points are imputations, then

the compliance of the overall time series is $N_r/N = 0.5$. The AR(1) time series is uniform in time, and so it has a continuity of close to 1.0.

Chatfield [20, p23] discusses different ways of calculating the sample autocorrelation coefficients for a time series. The general form for arbitrary lag and mean is given in Section 3.3.1 later in this chapter. For a time series of length N and with zero mean $\{y_1, y_2, ..., y_N\}$ the sample autocorrelation coefficient at lag 1 is given by,

$$r = \frac{\sum_{t=1}^{N-1} y_t\, y_{t+1}}{\sum_{t=1}^{N} y_t^2} \tag{3.16}$$

For a mixed time series which is a combination of N_r data points from the AR(1) process followed by $N - N_r$ points from the random process, the autocorrelation is,

$$r_m = \frac{\sum_{t=1}^{N_r} y_t\, y_{t+1} + \sum_{t=N_r+1}^{N-1} y_t\, y_{t+1}}{\sum_{t=1}^{N} y_t^2} \tag{3.17}$$

The term on the right hand side of the numerator is zero for the random time series, so the measured autocorrelation is given by,

$$r_{cpl} = \frac{N_r}{N}\, r_a \tag{3.18}$$

where r_a is autocorrelation of the AR(1) time series. We see that N_r/N is the proportion of real to imputed values, which is the compliance. So if there is a long gap in a time series which is otherwise uniform (a continuity near 1.0), the measured autocorrelation is reduced by the compliance value, assuming random imputations with the same mean as the real observations.

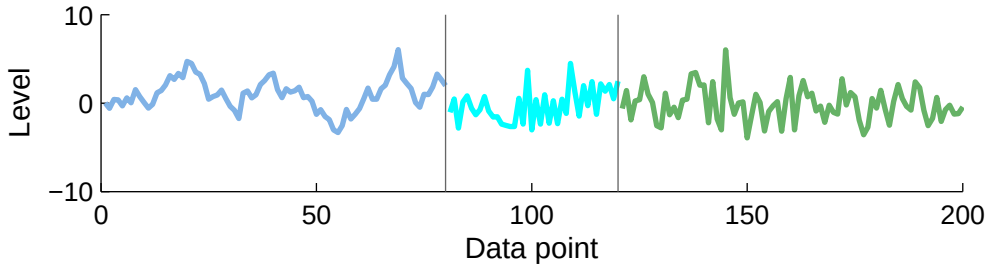


Figure 3.10: The time plot shows an instantiation of an AR(1) process from points 0 to 80 and a random time series at points 121 to 200. The points from 81 to 120 are taken in turn from the random time series then the AR(1) time series to create a non-uniform segment in the middle. Considering the AR(1) process as observations and the random process as imputed points, the overall compliance of the time series is 0.5 and the continuity is 0.8.

Figure 3.10 illustrates an instantiation of an AR(1) process of length $N_r = 100$ points merged with a random process of length $N - N_r = 100$ points. The first

segment of the time series is an AR(1) process which is considered to be the real data. The final segment is from a random process, and is considered to be imputed data. Between them, there is a segment of 40 points which are taken in turn from the random and then the AR(1) time series. The proportion of real to imputed values, or compliance, is therefore 0.5 as before. The proportion of correct intervals among the real points, or continuity, is 0.8.

The merged segment of the time series is composed of one real observation from the AR(1) process followed by one random imputation. By the same reasoning as before, neither the merged segment nor the pure random time series contribute to the autocorrelation. So the measured autocorrelation is,

$$r_s = \frac{\sum_{t=1}^{N_s-1} y_t\, y_{t+1}}{\sum_{t=1}^{N} y_t^2} \tag{3.19}$$

where N_s is the number of observations which have a uniform interval. Comparing this with (3.17) we see that the value of r_{cpl} is reduced by the proportion of uniform points N_s to real points N_r. So the measured autocorrelation becomes,

$$r_{cty} = \frac{N_s}{N_r}\, r_{cpl} \tag{3.20}$$

which is equal to the autocorrelation r_a of the AR(1) time series multiplied both by the continuity and the compliance.

The foregoing analysis was based on a toy model of an AR(1) process merged with a random time series to represent single gaps between real points. In practice, the mood time series have a longer correlation length, and the gaps do not have just one missing point. A time series with a longer correlation length, such as an AR(2) process, would show some correlation if single random points were interspersed in the data. Conversely, gaps are commonly longer then one interval, as the interval analysis in Appendix A shows.

3.2.1.4 Summary

Compliance is the proportion of non-imputed responses and continuity is the proportion of correct intervals among them. Continuity summarises the interval distribution using the probability density located only at the desired interval. The location of the remaining mass, corresponding to the distribution shape, does not influence its value. This property has an advantage over standard dispersion measures (of either the raw or the resampled series) because all intervals longer than the sampling period are classed together.

Long gaps in the time series, when the patient fails to respond for a period, are not reflected in the continuity measure although they do influence the compliance. This property of continuity is relevant to the autocorrelation calculation because time series with high continuity can be treated as uniform for this purpose. Both compliance and continuity can be useful in both selection of near-uniform series for the application of standard methods and for exploring non-uniformity as an informative property in itself.

3.2.2 Applying non-uniformity measures

This section applies the measures of compliance and continuity to the Oxford data set. Using the time series in Set A, we derive the compliance and continuity measures for each patient. A scatter plot of continuity against compliance is shown in Figure 3.11. From (3.4) we see that the minimum continuity tends towards the compliance as the compliance approaches 1. For lower compliance where there is a higher proportion of imputations, the continuity may be lower.

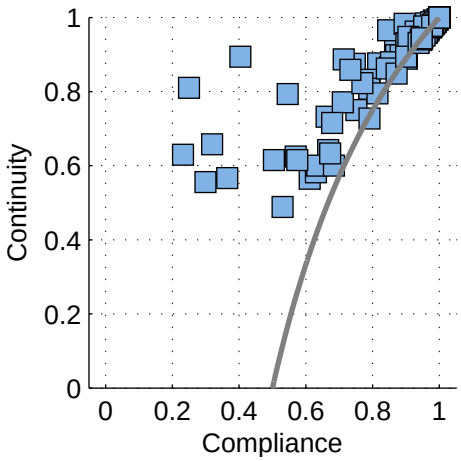


Figure 3.11: Continuity versus compliance for the 93 patients in Set A. The approximate minimum continuity limit $2 - C_m^{-1}$ from (3.4) is shown as a line. As compliance tends towards 1, the minimum possible continuity tends towards compliance. Those series in the upper left of the plot with high continuity and low compliance have large gaps where there is a long sequence of imputed points.

Figure 3.12 shows the compliance against continuity for sets of male and female patients in Set G and for bipolar I and II patients in Set D. No pattern emerges in either case, and a two-sample Kolmogorov-Smirnov test[3] cannot distinguish the pairs of distributions either for compliance or continuity.

[3]The Kolmogorov-Smirnov test is described in Section 2.1.3.1.

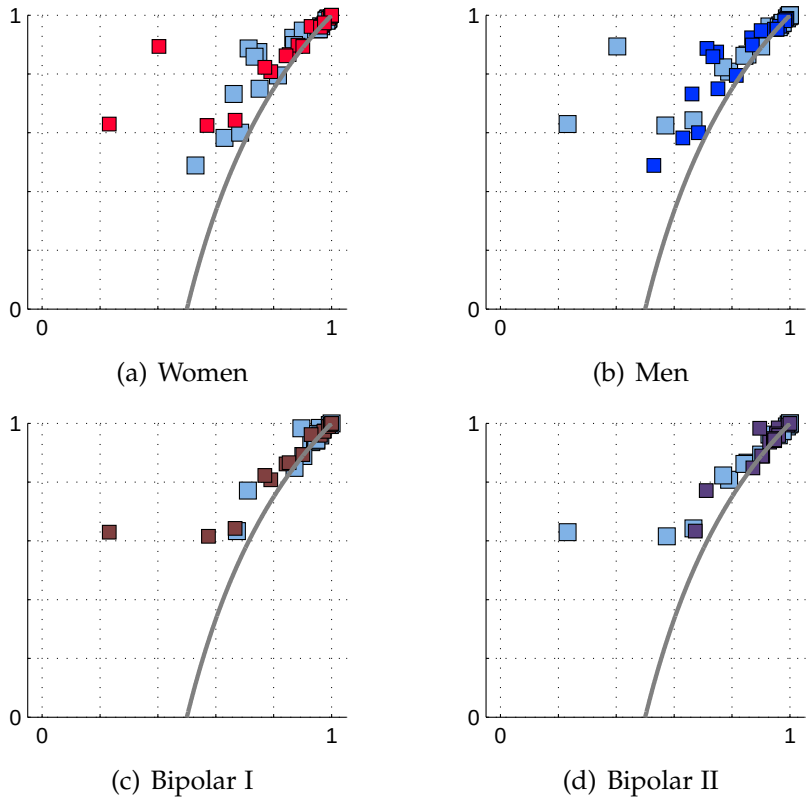


Figure 3.12: Continuity versus compliance for Set G (a) Women, (b) Men and Set D (c) Bipolar I patients, (d) Bipolar II patients.

3.2.2.1 Message latency

The foregoing analysis has assumed that any text message latency is small in comparison with the patient's delay in responding to a prompt from the monitoring system. Since we do not know when the prompt message is received by the patient we cannot distinguish total network latency from the patient's response delay. However since the prompt messages are dispatched at weekly intervals, we can judge the scale of the overall delays by examining the time between prompt and receipt.

Figure 3.13 shows the mean weekly delay in response including both the time taken by the patient to complete the questionnaire and any network latency. A prompt message is sent to the patient on the same day each week and the patient returns ratings in response. For each patient, the distribution of weekly delay times is summarised by its mean, and the mean delays are shown for the set of patients in Figure 3.13. The analysis shows that most patients have a mean delay of half of one day or more. This result is expected because the questionnaire

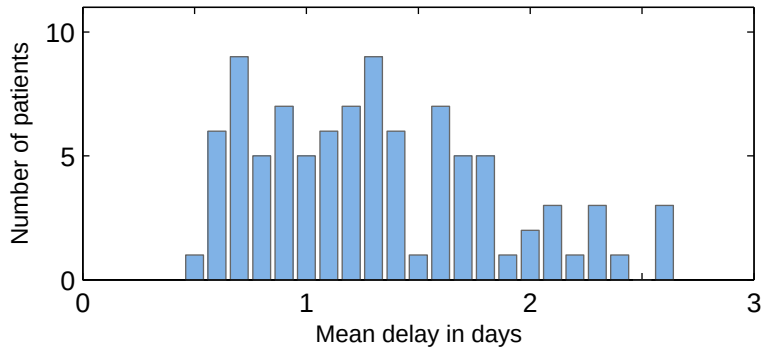


Figure 3.13: The mean time between the prompt message, which is sent each week and the receipt of the patient's ratings, measured in days. For each patient's time series, the distributions of the delay time each week is summarised by the mean. The mean delay for the set of patients is then presented as a distribution.

relates to a weekly period rather than an instant in time: patients do not have to reply to the prompt immediately. However, the network delay remains unknown and a quantitative study of the monitoring infrastructure would be valuable in determining the scale and nature of network latency.

3.2.3 Correlates of non-uniformity

Next we examine non-uniformity as a correlate of mood. There are 9 variables for depression corresponding to symptoms of *Sleep*, *Feeling sad*, *Appetite*, *Concentration*, *Self-view*, *Death/Suicide*, *Interest*, *Energy*, *Slowed down*. There are 5 variables for mania for symptoms of *Happiness*, *Self-confidence*, *Sleep*, *Talking* and *Activity*. Each symptom is summarised for each patient's time series by mean, standard deviation and the mean absolute difference between sequential values. We calculate the correlation for each symptom with continuity over the set of 93 patients in Set A. The rank correlations and p-values for correlation between continuity and depression symptoms are shown in Table 3.3.

No correlations were found between mean symptom levels and continuity. For the dispersion statistics only sleep in the depression questionnaire was found to be correlated with continuity. Variability of sleep correlates negatively with continuity when measured by standard deviation ($\rho = -0.26$, $p = 0.01$) and mean absolute difference between sequential sleep ratings ($\rho = -0.25$, $p = 0.02$). Scatter plots for both statistics are shown in Figure 3.14 which shows that for low continuity, sleep ratings tend to be variable. The converse is not true: that is, a high continuity does not imply less variable ratings. Repeating the investigation using

		Variability measure	
Domain	*Mean*	*Std. dev.*	*Mean abs. diff.*
Sleep	+0.14 (0.18)	-0.26 (0.01)	-0.25 (0.02)
Feeling sad	-0.13 (0.21)	-0.17 (0.10)	-0.09 (0.39)
Appetite/weight	-0.06 (0.59)	-0.04 (0.75)	-0.02 (0.88)
Concentration	-0.12 (0.24)	+0.01 (0.94)	-0.00 (0.96)
Self-view	-0.13 (0.22)	-0.15 (0.14)	-0.13 (0.23)
Death/suicide	-0.11 (0.27)	-0.15 (0.16)	-0.19 (0.06)
General interest	-0.11 (0.29)	-0.16 (0.12)	-0.19 (0.07)
Energy level	-0.08 (0.43)	-0.14 (0.19)	-0.05 (0.61)
Slowed down/Restless	-0.09 (0.39)	-0.08 (0.45)	-0.01 (0.91)

Table 3.3: Rank correlation (p-values) between depression symptoms and continuity for Set A. Correlations significant at the 5% level are highlighted.

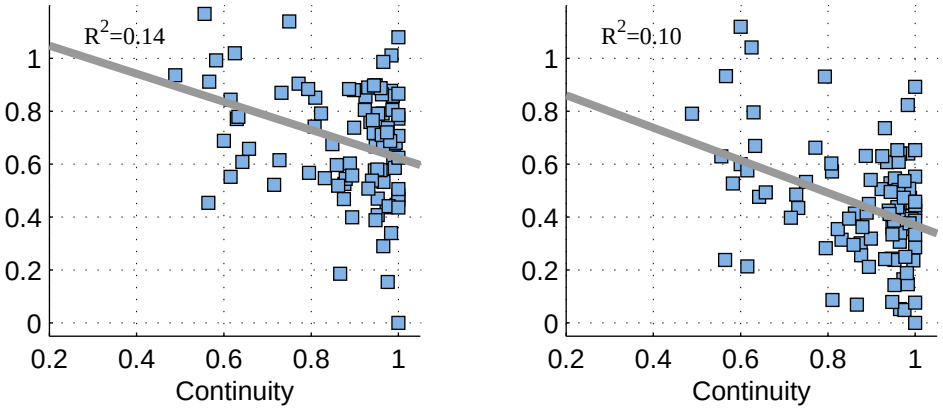


Figure 3.14: Scatter plots for variability of sleep against continuity. In (a) the standard deviation of all the resampled sleep values (excluding imputed points) for each patient are plotted against the continuity score for that patient. In (b) the mean of the absolute difference between sequential resampled values, again ignoring imputed points, is used. For both cases, patients with lower continuity show a higher variability in sleep responses on average. The linear least-squares fit is marked as a line. Sleep is rated on a scale from 0 to 3.

compliance as the non-uniformity measure yields the same results as with continuity: out of all the depression symptoms, only sleep variability correlates with compliance.

We note that there will be a sampling distribution for both the mean and

variability measures arising from the limited sample sizes, which would manifest in Figure 3.14 as a range for each point. For some symptoms, any correlation with non-uniformity might be hidden by this effect. However, since the same sampling limits apply to all symptoms we can distinguish sleep variability as having a particularly strong association with non-uniformity of response.

3.2.3.1 Interpretation

The relation of non-uniformity of response with sleep variability is an important finding from this analysis. The association is also interesting if response uniformity is taken as an indicator of general functioning. We would expect that delays in responding are caused by holidays, work commitments, physical illness, forgetting to reply, a low priority for replying or unpredictable behaviour. Psychological factors may have an influence and several of the symptoms explicitly measured on the QIDS scale are relevant, in particular severe lassitude or lack of energy, a lack of interest, poor concentration, and thoughts of death/suicide. As pointed out, it is quite possible that correlations with these variables exist but that they are below the noise threshold.

The relatively stronger effect of sleep points to a number of possibilities. Firstly, a strong association between sleep and mental illness is well established, if not well understood [105]. One possibility is that sleep is simply the strongest indicator of an underlying disorder which causes irregularity through the behavioural issues listed above. The causation might be more direct, for example sleep causing problems with memory or other functioning, leading to lost or delayed ratings. Given that patients may over- or under- report data depending on their mood, more direct telemonitoring monitoring approaches that do not depend on the patient's providing a response would be appropriate. An example of such as scheme is constant activity monitoring, which automatically returns data without the patient's intervention. We note that, it is a high variability of sleep ratings rather than a high mean rating that predicts non-uniformity of response. It may be that there is some adaptation to poor sleep, whereas inconsistent sleep leads to inconsistent behaviour. The data is too noisy and does not provide a strong enough effect to provide definitive evidence for these possibilities.

3.3 Correlates of depression

In this section we investigate the autocorrelation for depression time series and correlations between different time series. The *Edelson-Krolik* method [37] is used to compute correlation in the presence of non-uniformity, then use this method to investigate seasonality of depression. We then examine correlation between depression symptoms in individual patients, and then correlation between the time series of different patients.

3.3.1 Measuring correlation

The sample autocorrelation of a time series $\{y_1, y_2, ..., y_N\}$ is defined [20, p 23],

$$r(\tau) = \frac{\sum_{t=1}^{N-\tau}(y_t - \mu)(y_{t+\tau} - \mu)}{\sum_{t=1}^{N}(y_t - \mu)^2} \tag{3.21}$$

where N is the length of the time series, μ is its arithmetic mean and τ is the time difference or *lag* being measured. An informative way of representing the serial dependence in a time series is by a graph of autocorrelation coefficients $r(k)$ against the integer lag k. This sequence represents a sample autocorrelation function (a.c.f.) and is called the *correlogram* [35]. Since natural time series often have missing or irregular data, it is the applied sciences that have derived methods for their analysis. In astrophysics Edelson and Krolik [37] derived the discrete correlation function (DCF) for correlation estimation in non-uniform time series. It is defined for two discrete, centered time series a_i and b_j, first as a set of unbinned discrete correlation values

$$UDCF_{ij} = \frac{a_i b_j}{\sqrt{(\sigma_a^2 - e_a^2)(\sigma_b^2 - e_b^2)}} \tag{3.22}$$

for a measured pair of observations (a_i, b_j) whose time difference is Δt_{ij}. Here a_i and b_j are a concise notation for $a(t_i)$ and $b(t_j)$ respectively, σ_a, σ_b are the respective standard deviations and e_a, e_b are estimates of the measurement noise in each time series. The discrete correlation function is derived by averaging the set of M unbinned values

$$DCF(\tau) = \frac{1}{M} \sum_{i,j:|\Delta t_{ij} - \tau| < \frac{\Delta\tau}{2}} UDCF_{ij} \tag{3.23}$$

where $\Delta t_{ij} = |t_i - t_j|$, τ is the bin centre, $\Delta\tau$ the bin width. The standard error is,

$$\sigma_{DCF}(\tau) = \frac{1}{M''} \left(\sum (UDCF_{ij} - DCF(\tau))^2 \right)^{1/2} \tag{3.24}$$

recalling that $UDCF_{ij}$ is a set and $DCF(\tau)$ is a scalar for given τ. The summation is over $|\Delta t_{ij} - \tau| < \frac{\Delta\tau}{2}$ as before and the normalising constant M'' is equal to $((M-1)(M'-1))^{\frac{1}{2}}$ with M' the number of unique measurement times for a_i.

The Edelson-Krolik method is closely related to the *variogram*, an approach that is well known in geostatistics where it is used to model spatial correlations [27]. It was until recently rarely mentioned in texts on time series or in the statistical literature as a whole [58] with the exception of Chatfield [20] and Diggle [35], who defines the variogram as follows,

$$V(k) = \frac{1}{2}\mathcal{E}[\{Y(t) - Y(t-k)\}^2] \tag{3.25}$$

$$= \gamma(0)\,(1 - \rho(k)) \tag{3.26}$$

where terms are defined as before. A plot of the quantities $v_{ij} = \frac{1}{2}\{y(t_i) - y(t_j)\}^2$ for all delays $k_{ij} = t_i - t_j$ is called the sample variogram. As with the DCF, random scatter in the plot may arise from small sample sizes used in calculating v_{ij}. This scatter can be reduced by averaging v_{ij} over binned time values to give $\bar{v}(k)$.

The binned variogram and discrete correlation function are examples of a *slotting* approach that uses a rectangular kernel to bin pairs of observations. They belong to one of four categories identified by Broerson *et al.* [15] for handling non-uniform data. The other categories are direct transform approaches, such as the Lomb-Scargle periodogram [119], model-based estimators (which presuppose a knowledge of the time series dynamics) and resampling through interpolation. The Lomb-Scargle approach, kernel methods (though not slotting) and linear interpolation are compared in [114]. Since the data analyzed in this study has high relative noise and large gaps in the time indexes, we apply the Edelson-Krolik slotting approach for correlation estimation. It provides a sample correlogram directly and avoids the assumptions necessary for interpolation or model-based estimators.

3.3.2 Applying autocorrelation

We examine the autocorrelation function of the resampled depression time series in Set A using the Edelson-Krolik method[4] to find the autocorrelation at successive lags. Four examples of correlograms are shown in Figure 3.15 in comparison with a standard correlogram (lighter line) which has not been adjusted for non-uniform response times.

[4]The author's MATLAB implementation of the Edelson-Krolik method was used for this work.

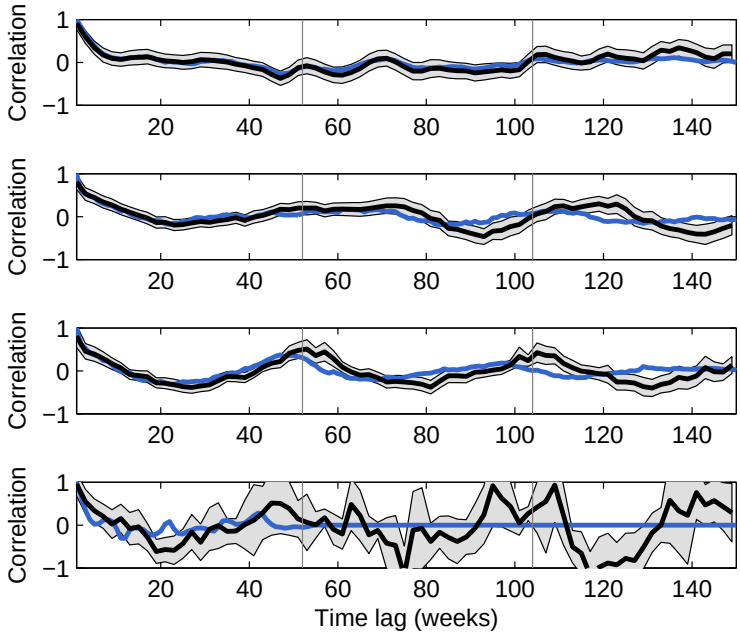


Figure 3.15: Correlograms for the depression time series from four patients. In each plot, the dark line is the correlogram estimated using the Edelson-Krolik method with a bin width of 2 weeks and showing two standard errors each side as a filled region. The lighter line is the autocorrelation calculated under the assumption of a uniform series. In the time plot third from the top there is clear evidence of yearly seasonality of depression. The continuity values for the time series are, from top to bottom: 0.99, 0.92, 0.87 and 0.30. Vertical lines are year markers corresponding to 52 and 104 weeks. Note that correlograms are defined only at integer lags or bin centres but are shown as continuous lines for clarity.

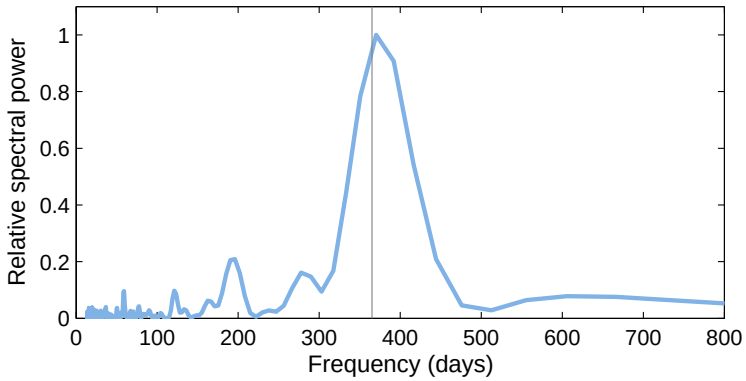


Figure 3.16: Lomb periodogram for a patient exhibiting seasonality of depression. The corresponding correlogram in Figure 3.15 is third from the top. The spectral power is normalised by the peak power and the periodicity of 365 days is marked as a vertical line. The peak is at a period of 370 days, the difference being due to a short sample size. In general the depression time series do not show such clear evidence of yearly periodicity, although some patients have a peak at or near this period.

Figure 3.16 is the Lomb-Scargle periodogram corresponding to this time plot. It shows a peak of spectral power at 370 days indicating a yearly seasonality. The depression time series do not in general show clear evidence of yearly periodicity, though some have a peak at or near this period. Most exhibit a rapid decrease in correlation with lag and some show evidence of a trend, indicated by the correlogram not tending to zero as the lag increases.

3.3.3 Applying correlation

In this section we apply the Edelson-Krolik method to correlations between time series. There are two applications, the first being to look at correlations between symptoms for individual patients and the second between depression time series for different patients.

3.3.3.1 Correlation between depression symptoms

For investigating the correlation between depression symptoms in individual patients we select uniform time series with at least 100 points. For each patient the correlation is then found between pairs of symptoms and those with non-negative correlations are selected. To find correlations, only the first 100 points are used and imputed values are excluded. The resampled time series are used because the fixed length then constrains the period of observation for each patient to be approximately one year. Of the 53 patients satisfying the minimum length criterion, 24 do not return valid correlations for all of the pairs, and 6 have some negative correlations, leaving 23 patients in Set E. The selection process is illustrated in Figure 3.17 and the statistical qualities are shown in Table 3.4, which includes Set A for comparison. The table shows that Set E is broadly representative of the larger Set A with the exception of time series length: the mean length for set E is 170 points compared with 95 points for Set A.

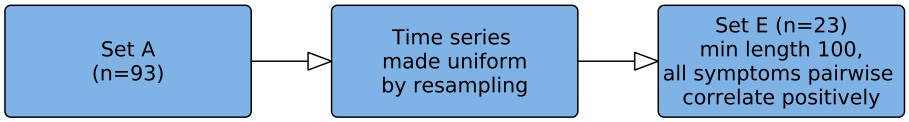


Figure 3.17: Flow chart for data selection. From Set A (n=93), a resampled set of time series is created. From it, set E (n=23) is created, having at least 100 data points in the resampled time series, and with all the symptom time series for each patient having valid pairwise correlations.

	n	*Age in years*	*Gender F/M*	*Subtype I/II/?*	*Length points*	*Depr.n QIDS*
Set A	93	45(20)	60/33	41/22/30	95(121)	6.3(5.2)
Set E	23	45(24)	17/6	12/7/4	170(68)	7.1(7.5)

Table 3.4: Median (interquartile range) of age, length and mean mood distributions with demographic summary for Sets A and E.

Using Set E, the correlation between pairs of symptoms is found the Edelson-Krolik method (3.23) and the scores for each pair is averaged over the set of patients. A heat map showing the relationship between symptom domains is shown in Figure 3.18. On average, the symptom domains *Sleep* and *Appetite/weight* correlate less than other domains. By contrast *Feeling sad* correlates strongly with other domains while *Slowed down/restless* shows less correlation with others.

An analysis of the autocorrelation structure for symptom time series explains why the symptoms of *Sleep* and *Appetite/weight* tend to correlate less when paired with other domains. We take the 23 time series in Set E and find the autocorrelations using the Edelson-Krolik method. The results for lags 1 to 4 are shown in Figure 3.19 where for each lag the average correlation over the set of 23 patients is shown. The symptoms *Sleep* and *Appetite/weight* have a lower autocorrelation than the other symptoms which explains their relatively low pairwise correlation in Figure 3.18. Although *Sleep* and *Appetite/weight* have a similar first order autocorrelation, Figure 3.18 shows that they do not themselves correlate as a pair, the reason being that their autocorrelation structure is somewhat different: the autocorrelation for *Sleep* remains higher than *Appetite/weight* as the lag increases.

We note that these two symptoms are the most amenable to objective measurement out of the nine symptoms in the QIDS rating scale and that *Slowed down/Restless*, which might also fall into this category also correlates less than the others. It may be that the other symptoms: *Feeling sad*, *Concentration*, *Self-view*, *Thoughts of death/suicide*, *Interest* and *Energy level* have a common factor which influences them more than it does the other three symptoms. This finding is similar to that in [116] which identified three factors in the IDS instrument: cognitive/mood, anxiety/arousal and sleep (or sleep/appetite for the self-rated instrument).

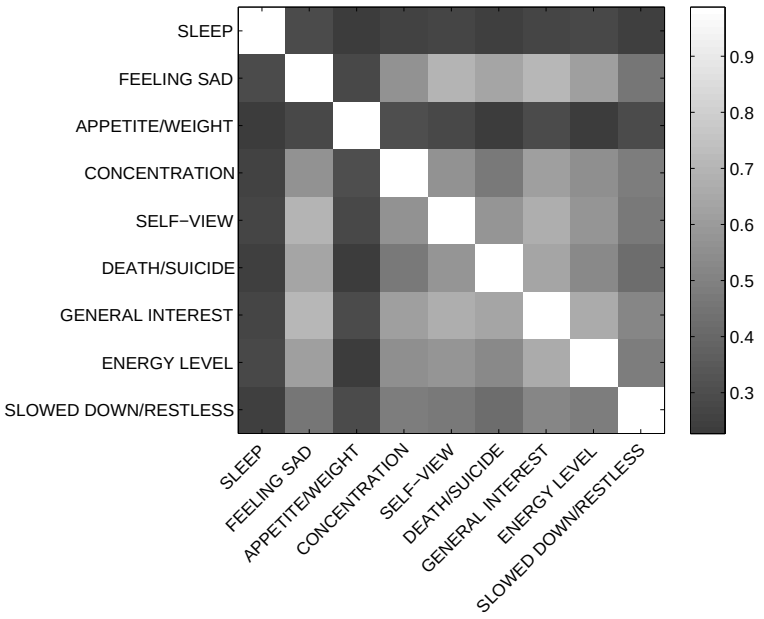
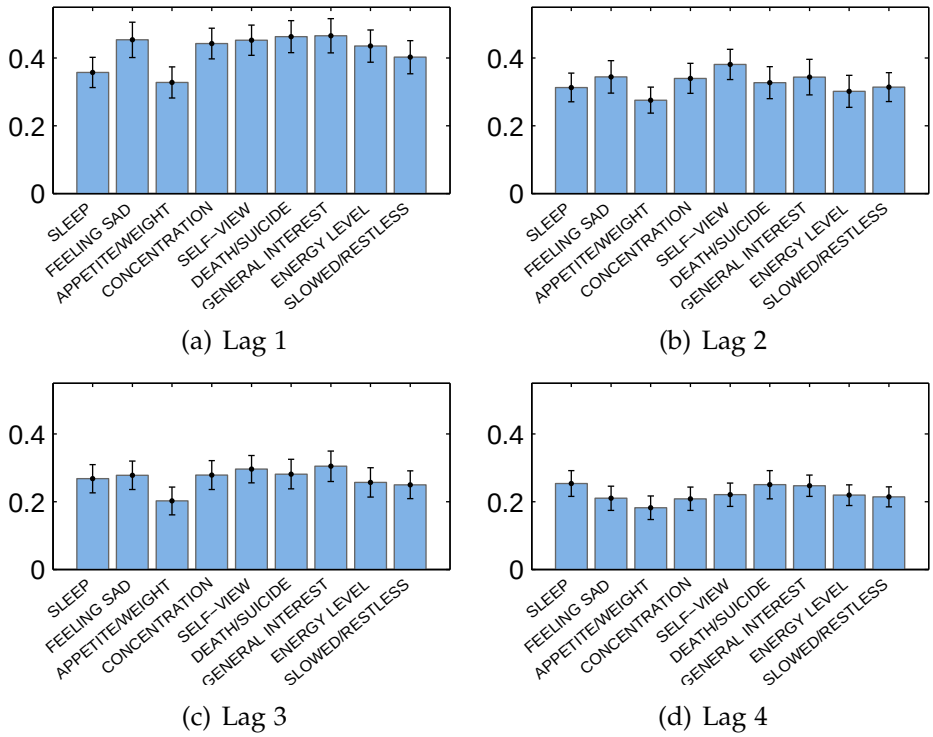


Figure 3.18: Matrix of mean correlation between pairs of depression symptoms. For each patient in a set of 23, we find the correlation between pairs of symptoms and present the average over whole set. Negative or non-significant correlations are excluded.



(a) Lag 1 (b) Lag 2

(c) Lag 3 (d) Lag 4

Figure 3.19: Autocorrelation for symptom time series. The chart represents the mean autocorrelation of a set of 23 patients, with error bars showing the standard error.

3.3.3.2 Correlation between patients' time series

We next look for similar mood changes in patients by examining correlations between their time series of depression ratings. For this analysis we select a set of 28 patients whose depression series cover the years 2009 and 2010 and which have fewer than 20% imputations in that period. The selection process is illustrated in Figure 3.20 and the statistical qualities are shown in Table 3.5, which includes Set A for comparison.

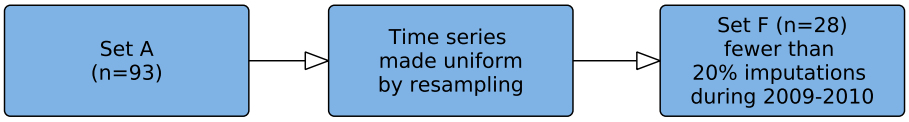


Figure 3.20: Flow chart for data selection. From Set A (n=93), a resampled set of time series is created. Set F (n=28) is comprised of time series which span the years 2009-2010 and have fewer than 20% of imputed points over that period.

	n	*Age in years*	*Gender F/M*	*Subtype I/II/?*	*Length points*	*Depr.n QIDS*
Set A	93	45(20)	60/33	41/22/30	95(121)	6.3(5.2)
Set F	28	50(22)	18/10	14/9/5	191(50)	5.9(7.9)

Table 3.5: Median (interquartile range) of age, length and mean mood distributions with demographic summary for Sets A and F

We compute the pairwise correlations between their time series and for each pair we create a surrogate set of 100 pairs of time series. We create surrogate time series by shuffling the time order of existing series while maintaining their mean, variance and autocorrelation function[5]. The original correlation between the pair is then compared with the distribution of 100 pairwise correlations and the null hypothesis is rejected for those that lie in the top 1%.

In the set of 28 patients, 8 pairs are found to be correlated with an associated significance level of 1%. However, since there are 378 possible pairs, we apply the more stringent test of $p < \frac{0.01}{378}$, computed using the Fisher transformation applied to the correlation. The two pairs that correlate with a significance at this level are shown in Figure 3.21. Both examples show some examples of similar mood change but no clear pattern such as yearly seasonality.

[5]The algorithm used for this process is described in [73] and is implemented using the *TISEAN* function `surrogates` [61]. Further explanation about surrogates is given in Chapter 5.

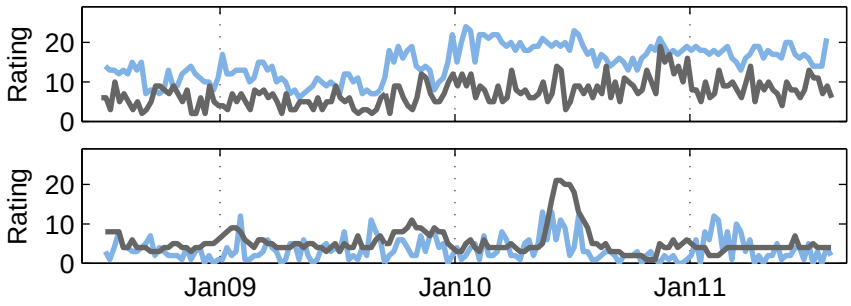


Figure 3.21: Pairs of time plots which correlate. The upper plot has a correlation of 0.50 over the period 2009-2010 and the lower plot a correlation of 0.43.

To examine correlations more generally over the set of patients we compare the distribution of pairwise correlations in the original set of time series with that of a surrogate set. As before we use Set F but in this case the surrogate data set is formed by creating one surrogate time series from each of the original time series. The distribution of the pairwise correlations for both the original and surrogate data sets is shown in Figure 3.22.

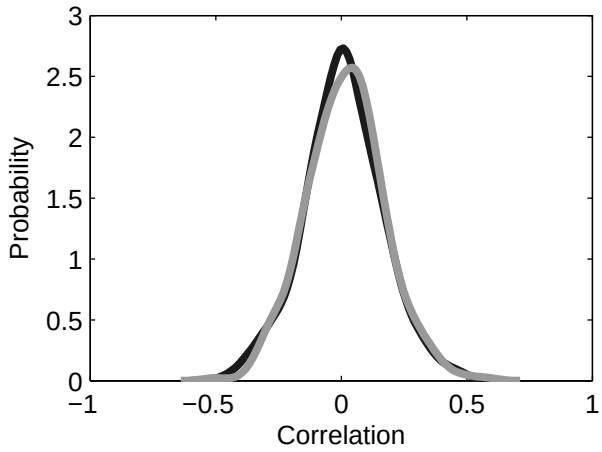


Figure 3.22: Kernel density estimate of pairwise correlations between time series. The dark line is the density estimate for the original set of time series and the light line for the surrogate data. Each surrogate time series is derived from its original counterpart by taking the Fourier transform and randomizing the phases to obtain a time series with the same power spectrum. The method removes any correlation between pairs of time series that arises from a common source rather than by chance. The similarity of the distributions shows that in general there is little correlation present among pairs of the original time series.

The correlations between time series for original and surrogate data sets appear to have the same distribution and a two sample Kolmogorov-Smirnov test

returns a value of $p = 0.32$. This result suggest that external factors do not have a strong influence on depression over the set of patients. It does not however preclude the possibility that there are strong environmental effects in individual cases such as yearly mood variation.

3.4 Conclusions

We have described the data set used in this thesis and presented some of its statistical properties. The data is not uniformly sampled so we introduced two new measures for quantifying missing and non-uniform data. The quantification of non-uniformity can be useful in 1) investigation of non-uniformity as a correlate of other variables 2) selecting subsets of data where uniformity is a requirement 3) use as supplementary information for a clinician. We found that time series uniformity does not correlate with either gender or diagnostic subtype. However, variability of sleep correlates with continuity. This finding has implications for selecting time series according to their uniformity since it may exclude patients with more variable sleep ratings.

The Edelson-Krolik method uses relative distances rather than fixed lags to determine time series correlation and so it is robust to non-uniform sampling intervals. We used the method to generate correlograms of depression ratings and showed that one patient exhibited mood with yearly seasonality. Most patients do not show evidence of seasonality, but instead show a rapid decrease in autocorrelation. We next examined correlations between depression symptoms and found that *sleep* and *appetite/weight* show a lower average correlation than other symptoms. Examination of the autocorrelation structure for these domains showed that it was different from that of the others.

Finally, we examined correlations between patients' depression time series but found no evidence of correlation in general. We note that for some patients, the weekly sampling will be below the Nyquist frequency for depression, so information might be lost. A study identifying the range of frequencies in depression in bipolar would help in choosing an optimal sample rate, consistent with practical considerations.

4

Forecasting mood

Introduction

In Chapter 1, we discussed Bonsall's approach [11] to characterizing mood time series, and asked if it is possible to forecast mood in bipolar disorder using weekly time series. Since effective mood prediction could support management of the disorder and provide an early warning for relapse, the issue is of clinical importance. This chapter aims to answer the question by applying a range of time series forecasting methods to the OXTEXT depression data, including Bonsall's forecasting method, described in Section 4.1. In addition we use baseline forecasts to assess the skill of methods: without a baseline it is hard to evaluate the results from more sophisticated methods. We use in-sample forecasting to examine the dynamics of the time series in comparison with two baseline methods: unconditional mean and persistence. To find out if forecasting could be useful in practice we estimate the expected prediction error using out-of-sample forecasting.

The structure of the chapter is as follows. Section 1 provides a review and critique of the Bonsall *et al.* [11] paper. Section 2 describes how the data has been selected from an initial set, and provides some descriptive statistics of the selected time series. Section 3 describes the in-sample and out-of-sample forecasting experiments. Finally, Section 4 draws conclusions both in terms of the data qualities and the potential for effective forecasting.

4.1 Analysis by Bonsall *et al.*

Bonsall and his co-authors [11] applied time series methods to self-rated depression data from patients with bipolar disorder. The focus of their study was mood stability: they noted that while treatment has often focused on understanding the more disruptive aspects of the disorder such as episodes of mania, the chronic week-to-week mood instability experienced by some patients is a major source of morbidity. The aim of their study was to use a time series approach to describe mood variability in two sets of patients, stable and unstable, whose members are classified by clinical judgement.

The time series data were obtained from the Department of Psychiatry in Oxford and was from 23 patients monitored over a period of up to 220 weeks. The patients were divided into two sets of stable and unstable mood based on a psychiatric evaluation of an initial six month period of mood score data. The classification of patients into two groups was made on the basis of mood score charts and non-parametric statistical analysis, described in [63]. The depression data for each group was then analysed using descriptive statistics, missing value analysis (including the attrition rate) and time series analysis. Descriptive statistics showed that average mood scores and variability were higher for the unstable group. The distributions of scores between the groups were different, with the stable group having a mode of 0 compared with 5 for the unstable group. It was found that missing values varied randomly through time, but there was also some variation among patients. There was no significant difference between the patient attrition rates between the groups.

The time series analysis was based on applying autoregressive models of orders 1 and 2 to the depression data for each patient. The error model used was the gamma distribution. Both conventional (AR) and threshold (TAR) autoregressive models were used with the threshold set to the mean score. Model fitting was based on maximum likelihood and took account of missing values using the expectation-maximization algorithm. In order to balance goodness of fit and model complexity the Akaike information criterion (AIC) was used to choose the model. The chosen model was found to depend on the group, with the stable group best described by TAR(1) and the unstable group by TAR(2). Within the groups themselves, they found that the AR processes differed (in their parameters but not in their order) above and below the patient's mean score.

The authors concluded that the existence of nonlinear mood variability suggested a range of underlying deterministic patterns. They suggested that the difference between the two models could be used to determine whether a patient would occupy a stable or unstable clinical course during their illness and that the ability to characterize mood variability might lead to treatment innovation.

4.1.1 Critique

The Bonsall *et al.* study has a well-founded motivation. There is evidence that symptoms of bipolar disorder fluctuate over time, and it is common for patients to experience problems with mood between episodes [71]. In approaching time series with unknown dynamics, the use of an autoregressive model is an obvious starting point. However there are signs that the model fit is poor in this case. The distribution of RMSE values for the stable patient group is reported to have a median of 5.7 (0.21 when normalised by the maximum scale value) and the distribution for the unstable group a median of 4.1 (0.15 normalised) [11, Data supplement]. These are in-sample errors rather than expected prediction errors estimated by out-of-sample forecasting. As such, they are rather high compared with the reported standard deviations (3.4 for the stable group and 6.5 for the unstable group) suggesting that the unconditional mean would be a better model for the stable group.

The reason for the poor model fit is not clear. One contributing factor might be that the time series are non-stationary: a visual inspection of Figures 4 and 5 in [11] shows a variation in mean for some of the time plots. To mitigate non-stationarity, a common technique is to difference the time series, as in ARIMA modelling, or to use a technique that has the equivalent effect, such as simple exponential smoothing. Also the use of a baseline model is missing from the analysis. Using a baseline gives a reference of model skill and a means of identifying implementation errors. In any case, the comparison of threshold AR with conventional AR models does not adequately identify model dynamics and nor does it justify their suggestion of chaotic dynamics. Although the authors cite Gottschalk *et al.*'s [55] claims of chaos, they did not mention the criticisms given in [79] nor Gottschalk *et al.*'s response in [56].

4.2 Time series analysis

4.2.1 Data selection

The data used in this chapter is the non-uniform *OXTEXT* time series which were described in the previous chapter. In this section a subset of patients whose time series have a minimum length of 64 points is selected for testing stationarity and self-affinity. Only the depression ratings are used, because for some patients the mania scores are always zero. The selection process is illustrated in Figure 4.1 and the statistical qualities of the subset are shown in Table 4.1. The data was collected using the same telemonitoring techniques and the same rating scales as Bonsall *et al.* [11], but the studies are independent and the two data sets are not known to have any shared patients.

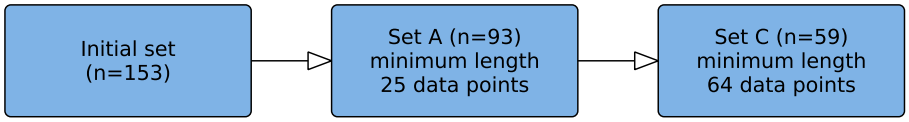


Figure 4.1: The Department of Psychiatry has provided a data set of 153 patients, each having time series of mania and depression. From this initial data set, Set C (n=59) of time series having a minimum length of 64 data points is selected.

	n	*Age in years*	*Gender F/M*	*Subtype I/II/?*	*Length points*	*Depr.n QIDS*
Set A	93	45(20)	60/33	41/22/30	95(121)	6.3(5.2)
Set C	59	50(15)	38/21	31/14/14	156(80)	5.9(4.5)

Table 4.1: Median (interquartile range) of age, length and mean mood for the time series selected for analysis. Set A with a minimum time series length of 25 is shown for comparison.

4.2.2 Stationarity

We examine stationarity for the depression time series in Set C using descriptive and inferential statistics. Figure 4.2 shows the distributions of the first and second halves of 25 time series from Set C. The median is marked as a square, with error bars showing the 25^{th} and 75^{th} percentiles of each distribution. There is variation among the patients' time series, with some showing a marked change in median,

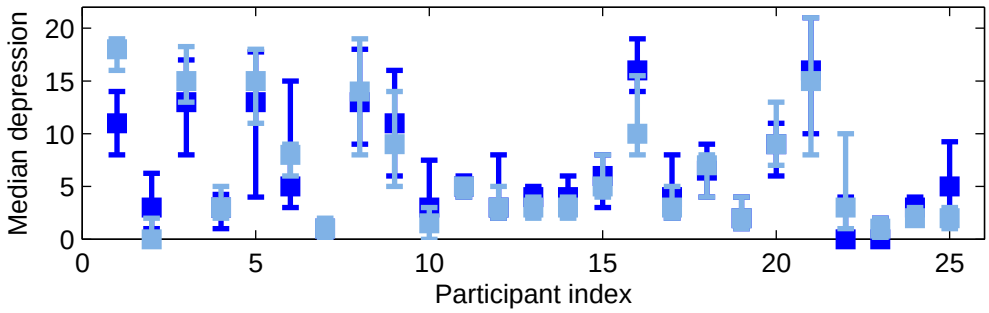


Figure 4.2: Change in median depression over the observation period for the first 25 patients in Set C. The figure shows the change in median rating of the first half (dark) of a patient's depression time series compared with the second half (light). Error bars show the 25^{th} and 75^{th} percentile in each distribution. For patients with indices 7 and 19 the distributions of the first and second half of the time series are identical.

for example patient 1, and others showing an identical distribution, for example patients 7 and 19.

Figure 4.3 shows the trend in the first and second halves of nine time series chosen randomly from Set C. A trend is a long-term variation up or down and in some cases it is a result of medication. It implies nonstationarity, which is a change in the distribution of response values over time. Most time series in Figure 4.3 show little marked difference in trend but in two cases (middle-left and bottom right panes of the figure) there are clear changes.

A quantitative test for nonstationarity is made by comparing the distributions for the first and second half of each patient's time series using the two sample Kolmogorov-Smirnov test, described in Section 2.1.3.1. The test is applied to each of the 59 patients in Set C and we find significance for 28 of these at the 1% level. The result implies that a large proportion of time series undergoes a change in distribution between the first and second halves of the observation period.

4.2.3 Detrended fluctuation analysis

Another quality that appears to vary among patients is the apparent roughness of the time series, a quality related to its autocorrelation function. In this context, roughness represents the degree to which mood is changing erratically over time. Detrended fluctuation analysis, introduced by Peng *et al.*[108], measures roughness by estimating the statistical self-affinity of a time series. It also detects long range correlations in the presence of nonstationarity. The algorithm finds the change in fluctuation with window size over the integrated series giving the

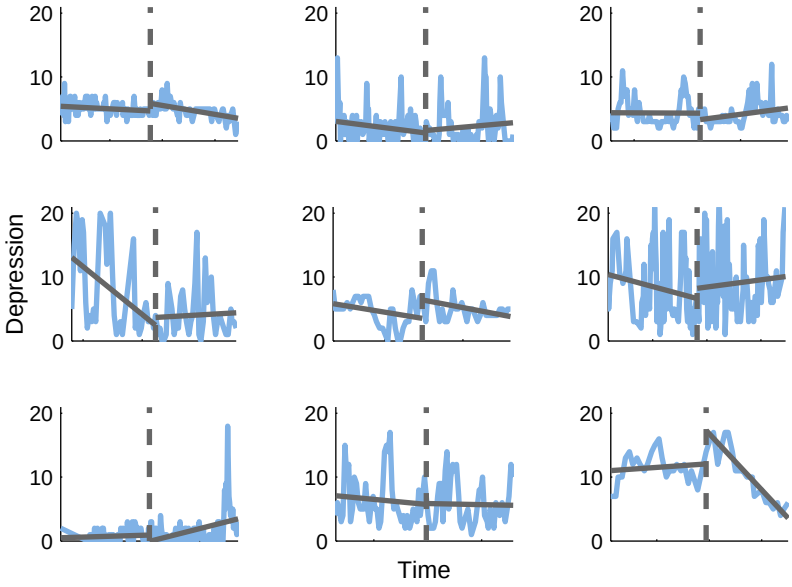


Figure 4.3: Illustration of nonstationarity using nine time plots selected randomly from Set C. A linear least squares line is fit to the first and second half of each plot. In some cases the trend differs markedly from the first half to the next, possibly because treatment has been changed. Time periods vary between each plot.

scaling exponent s which measures the self-affinity or smoothness of a time series. The fluctuation L is found by first fitting a trend line within a given window length and deriving the slope m and intercept b by least squares. The fluctuation for the window size is then found as

$$F(L) = \left[\frac{1}{L} \sum_{n=1}^{L} (y(n) - mn - b)^2 \right]^{\frac{1}{2}} \tag{4.1}$$

Self similarity is manifested as $F(L) \propto L^s$, and the scaling exponent s is found as the slope of the log-log graph of L against $F(L)$. A scaling exponent of $s = 0.5$ represents white noise, a value of 1.0 corresponds to $1/f$ noise, and a value of 1.5 corresponds to a random walk.

The performance of the technique has been studied under different conditions of nonstationarity and time series length [21] and an algorithm for computing the scaling exponent is described in [83]. The algorithm was applied to each of the time series in Set C using the implementation associated with [83]. The results are shown in Figure 4.4. Inspection of the time plots corresponding to large s shows that those patients with $s > 1$ are highly nonstationary. For those time series with small s, the time plots show much less of a change of mean rating, and have an

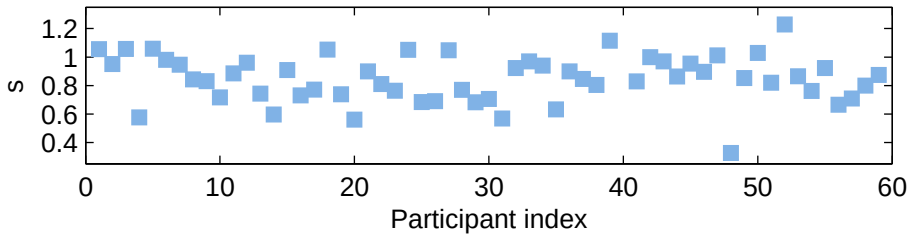


Figure 4.4: Variation in smoothness over the set of time series in Set C. Smoothness is measured by the scaling exponent s derived from detrended fluctuation analysis (DFA). Higher values of the scaling exponent correspond to smoother time series.

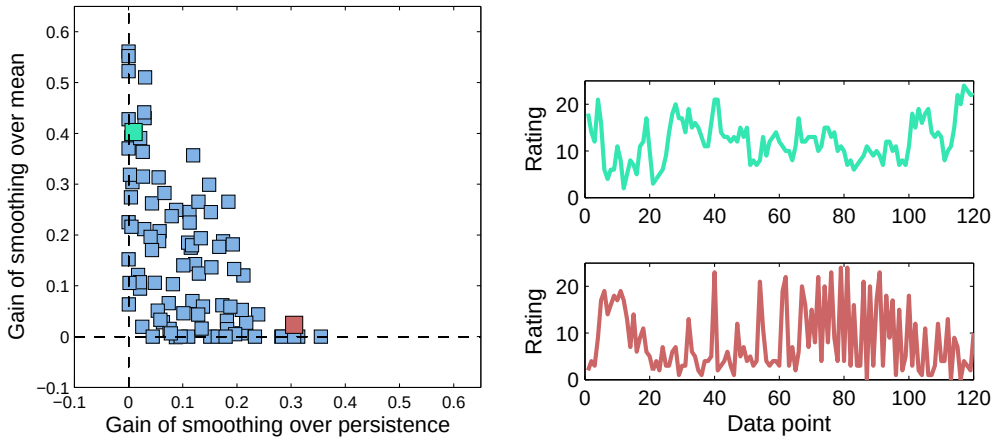
erratic, almost random variation. We find that few time series are periodic, many show a global trend and there is a variation from noise to random walk dynamics.

4.3 Forecasting

The time series analysis shows that there is variation in the stationarity and smoothness qualities of the time series. This section describes some experiments with forecasting depression. The purpose of forecasting is twofold, firstly to examine the dynamics of individual time series and secondly to estimate the expected prediction error. Time series dynamics are investigated by using in-sample forecasting to compare exponential smoothing with baseline methods. Out-of-sample forecasting is used to estimate the expected prediction error. For both in-sample and out-of-sample approaches we forecast the next mood rating starting from a small margin at the beginning of the time series and continuing to the end. The error for the patient is then summarized by applying a loss function to the residuals and these errors are examined for the set of patients.

4.3.1 In-sample forecasting

For in-sample forecasting, we compare simple exponential smoothing forecasts with baseline predictors of unconditional mean and persistence. Exponential smoothing takes a forecast $\hat{y}_t$ at time t and adjusts it to give a next step forecast of $\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t)$, where y_t is the actual value at time t and α is a constant between zero and one. In order to measure performance against both baselines simultaneously we set the seed forecast value $\hat{y}_1$ to the unconditional mean. A value of zero for the smoothing parameter then uses the unconditional mean predictor while a value of one gives a persistence forecast. The smoothing parameter is found for each patient by searching for the minimum forecast error



(a) (b)

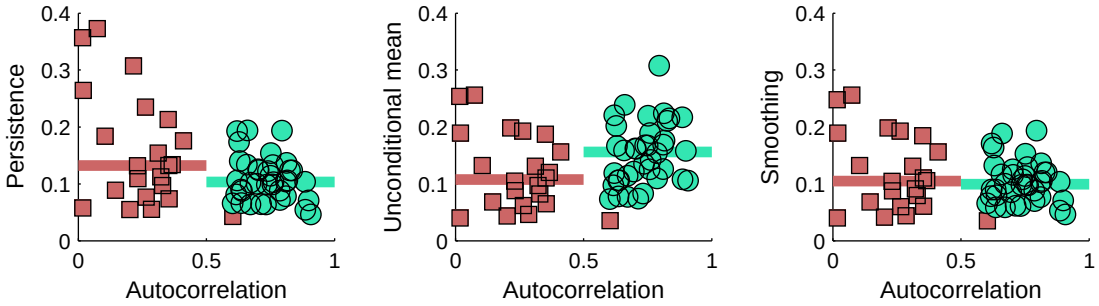
Figure 4.5: The scatter plot shows the relative error reduction of simple exponential smoothing compared with persistence (x-axis) and unconditional mean (y-axis) where each square corresponds to a patient's time series. The gain on each axis is represented as the proportionate decrease in patient error compared with the baseline. Points in the top left region have a smoothing parameter close to 1, corresponding to a persistence forecast. Points in the bottom right region have a smoothing parameter close to 0, corresponding to a mean forecast. Simple exponential smoothing always performs at least as well as the baseline methods, but for many time series there is no improvement over the relevant baseline. Only 20 of the series improve by more than 10% over both persistence and unconditional mean forecasts. The time plots correspond to the highlighted points on the scatter plot, and they illustrate smooth and rough series respectively.

over the parameter range of zero to one. In this way we can get some insight into the relative accuracy gains over these two baselines.

Figure 4.5(a) shows the results for Set A (n=93) using the non-uniform time series. The scatter plot shows the relative gains over the baseline methods on each axis where every square corresponds to a depression time series. Gains of exponential smoothing over unconditional mean are as much as 0.56, that is, a 56% reduction in error, and 0.36 over persistence. The time plots in Figure 4.5(b) shows examples where no gain is found over persistence (top right panel) and none over unconditional mean (bottom right panel). The lower plot varies considerably from week to week and has a DFA scaling exponent of 0.8 while the top plot has a scaling exponent of 1.1 and it shows less variation.

4.3.1.1 Effect of smoothing

We next consider forecast results for two groups which represent these examples. One group comprises 23 patients with 'rough' time series for which unconditional



(a) Persistence (b) Unconditional mean (c) Exponential smoothing

Figure 4.6: Forecast error (normalised RMSE) against first order correlation for persistence, unconditional mean and exponential smoothing methods. Two groups of patients are shown for each method: squares correspond to rough time series and the circles to smooth time series. A horizontal line is drawn through each cluster to represent its median. The left panel shows that errors for persistence forecasting are higher and more dispersed for the rough set than for the smooth set. The situation is reversed for unconditional mean forecasting, shown in the middle panel, where the rough set shows lower errors than the smooth set. Simple exponential smoothing, shown in the right panel, adapts to the qualities of both sets and preserves the accuracy of the respective baselines.

mean forecasting improves by 10% or more over persistence forecasting. The second group is of 41 patients with 'smooth' time series where persistence improves by 10% or more over unconditional mean forecasts. The difference in smoothness between the sets is reflected in the DFA scaling exponents: for the smooth set the median exponent is 0.98 and for the rough set it is 0.73. Figure 4.6 shows the forecast error plotted against first order correlation for two groups of patients. Forecast accuracy for the groups is seen to be promoted by their relevant baseline method and penalized by the opposite method. Simple exponential smoothing preserves accuracy in both groups by adapting to the correlation structure of each time series. Most of the gain in accuracy from exponential smoothing arises from its ability to appropriately smooth the highly variable time series.

4.3.2 Out-of-sample forecasting

The in-sample forecasting approaches have revealed some of the dynamic properties of the time series. Out-of-sample forecasting estimates the expected prediction error and shows how accurate forecasting would be in practice. Parameter values are trained on the first 40 data points of the individual time series, and retrained at intervals of 20 points as the forecasts are made through the series. Each patient's error is estimated by applying one of three loss functions to the

next step forecasts: root mean square error (RMSE), mean absolute error (MAE) and median absolute error (MdAE), and errors are normalized by the maximum of the rating scale (0-27).

4.3.2.1 Implementation

For this work we use a range of forecasting methods to Set C which has 59 patients each having at least 64 data points. The methods are summarised in Table 4.2. For the autoregressive models (AR1,AR2) the ARfit package [120] is used to choose the model order using the Bayesian Information Criterion (BIC), which is described in Section 2.2.3.1. For the self-exciting autoregressive models (SR1,SR2) an implementation originally written by Siddharth Arora [5] was adapted by the author to create models with two regimes using autoregressive orders 1 and 2 respectively. The threshold variable takes on sample values between the 15th and 85th percentiles of the training data and its value is monitored using a unit delay. The simple exponential smoothing (SES) method was coded by the author using an implementation originally written by Siddharth Arora [5] as a basis.

For Gaussian process regression (GPR), the implementation used for hyperparameter optimisation and prediction is the GPML software dated July 2007 [112]. The covariance function applied is a rational quadratic which has the form $k(r) = \left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha}$ where r is the difference between time indices, l is a length scale and α determines how the covariance changes with r. The k-nearest neighbour forecasting method (KNN) is the local linear model implemented using the *TISEAN* software [61]. The threshold autoregressive model (BON) used by Bonsall *et al.* [11] is implemented by fitting autoregressive models separately above and below the mean of the training data, the model orders being chosen as either 1 or 2 depending on the Bayesian Information Criterion.

4.3.2.2 Results

Figure 4.7 shows the results for RMSE loss as error distributions over the set of patients. The highest errors are generated by persistence and k-nearest neighbour forecasts. Large errors are mitigated by the smoothing methods, in particular by unconditional mean, which has no outliers although it has relatively poor accuracy. This observation matches the finding in Section 4.3.1 that some of the time series are better forecast by mean than by persistence.

The median forecast error for the unconditional mean forecast method is

Abbreviation	*Method*
UCM	Unconditional mean
PST	Persistence
AR1	Autoregressive order 1
AR2	Autoregressive order 2
SR1	SETAR with 2 regimes and order 1
SR2	SETAR with 2 regimes and order 2
SES	Simple exponential smoothing
GPR	Gaussian process regression
KNN	K-nearest neighbour regression
BON	Mean threshold AR by Bonsall *et al.*

Table 4.2: Out-of-sample forecasting methods

higher than that for the other methods, reflecting its inability to model recent mood and exploit any autocorrelation that is present. The distributions for methods other than the unconditional mean are similar: all have a median error of approximately 0.1. The poor performance of the SETAR(2,1) and SETAR(2,2) methods may be attributed to a poor model fit due to the absence of clear regimes. It might also, as with k-NN regression, be due to the short training sets.

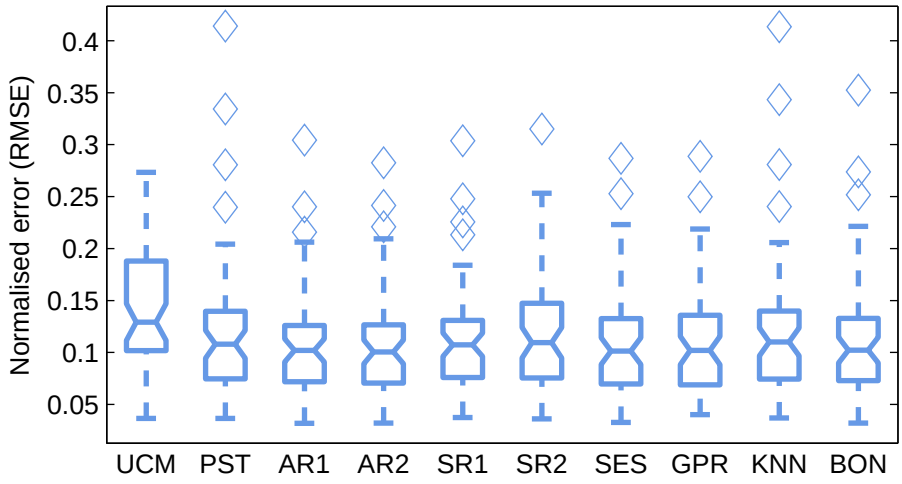


Figure 4.7: Distributions of out-of-sample errors for different forecasting methods. The errors are those of the 59 patients in Set C where the error for each patient is summarized by RMSE. This loss function is applied to the individual next step forecasts for each patient (the number of forecasts varies from patient to patient because the time series length varies over the set). Points are drawn as outliers (diamonds) if they are larger than $q3 + 1.5 * (q3 - q1)$ or smaller than $q1 - 1.5(q3 - q1)$, where $q1$ and $q3$ are the 25^{th} and 75^{th} percentiles respectively. The whiskers extend to the adjacent value, which is the most extreme data value that is not an outlier.

4.3.2.3 Loss functions

The forecast errors for different loss functions are shown in Table 4.3. The choice of loss function makes little difference to the relative performance of the methods, except for MdAE which gives a better score to unconditional mean forecasts. In

Loss fn.	*UCM*	*PST*	*AR1*	*AR2*	*SR1*	*SR2*	*SES*	*GPR*	*KNN*	*BON*
RMSE	0.13	0.11	0.10	0.10	0.11	0.11	0.10	0.10	0.11	0.10
MAE	0.10	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
MdAE	0.07	0.07	0.06	0.07	0.07	0.07	0.06	0.07	0.06	0.06

Table 4.3: Out-of-sample next step forecasting results showing errors for different loss functions.

this case, the very high residuals that characterise this baseline are not penalised as much as with the other loss functions. We note that the median RMSE value for the implementation of the *BON* method is 0.1, which compares with in-sample values reported by Bonsall *et al.* [11, electronic supplementary material] of 0.21 for their stable patient group and 0.15 for the unstable group. There are differences in the data and experiments: the set of patients used by Bonsall *et al.* is smaller, time series lengths are different and the Bonsall *et al.* errors from are in-sample forecasts. However the figure of 0.21 for the stable set is large compared with those presented here and also, as noted in section 4.1, with the standard deviation reported in that paper.

4.3.3 Non-uniformity, gender and diagnosis

This section reports the effects of non-uniformity, gender and diagnosis on the out-of-sample forecast error. In this section, we use the time series in set C, and their uniform equivalents which we denote Set C-R. Where forecasting methods are applied to the uniform time series, we denote the results by adding –R to the method, as for example PST-R for persistence forecasting applied to the uniform series. A similar notation is used for distinguishing male from female and BPI from BPII patients.

4.3.3.1 Effect of non-uniformity

The forecasting methods have so far been applied to time series which are non-uniformly sampled in time. However, all the methods, apart from Gaussian pro-

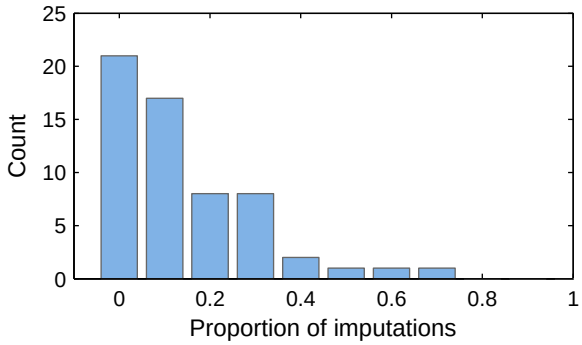


Figure 4.8: Histogram of the proportion of imputed points for time series in Set C-R. Out of 59 time series, 38 (the first two bars) have less than 15% of values that are imputed. The proportion of imputations in the resampled series is $1 - C_p$ where C_p is compliance as defined in Chapter 3.

cess regression, assume a uniform sampling period. We examine the effect of this assumption by applying the methods to uniformly sampled time series. The advantage in using uniform series is that the methods assume uniformity, but a disadvantage is that most of the series have imputed points. The proportion of imputations in the time series is shown in Figure 4.8.

The method of resampling to create a uniform time series is described in Section 3.2.1.1. Imputation is made by random selection from the previous 4 responses. The rationale for this choice is that random selection from the entire series would be a poor estimate in most cases, while using the last value would fill gaps with a constant value. We chose 4 weeks as a compromise because for most time series, the autocorrelation has decayed by this time, as shown by Figure 3.19.

We apply three of the forecasting methods (PST,SES,GPR) to both the original series in Set C and their resampled counterparts in Set C-R. The results are shown in Figure 4.9. The forecast errors for the resampled series are slightly lower, but in general terms there is little difference. Imputations in the resampled series can increase or decrease the forecast error. A next-step forecast error might be reduced or increased depending on the range of the previous four ratings and the dynamics of the time series, and the overall RMSE will depend on the proportion of imputations. Figure 4.10 shows the ratio of errors of the resampled and the original series against the non-uniformity measures which were developed in Chapter 3. As the compliance decreases, the relative error can either increase or decrease, the limit determined by the proportion of imputations. For continuity,

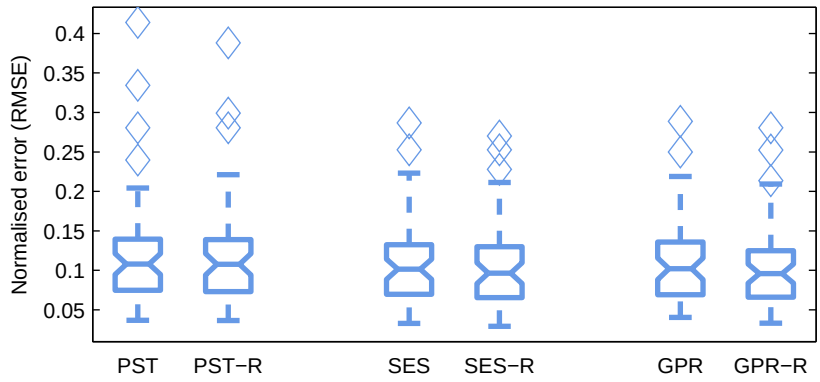
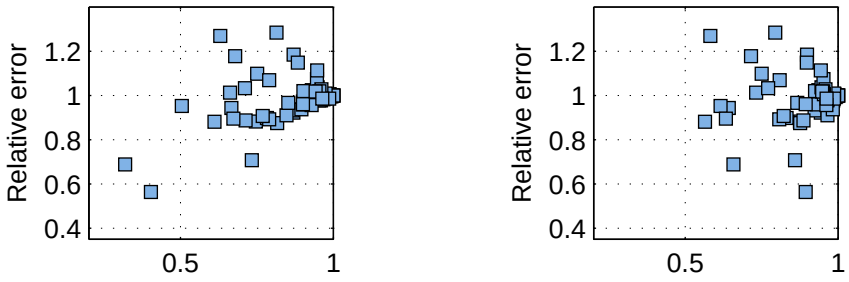


Figure 4.9: Distributions of out-of-sample errors for non-uniform and uniform time series. The errors are those of 59 time series in Set C and Set C-R and the error for each patient is summarized by RMSE. Errors are shown for persistence (PST), simple exponential smoothing (SES) and Gaussian process regression (GPR), first applied to the original (eg. PST) then to the resampled (eg. PST-R) time series.



(a) Compliance (b) Continuity

Figure 4.10: Ratio of errors of the resampled and original time series against (a) compliance and (b) continuity. The forecasting technique used is persistence. The spread of relative errors depends on compliance, with the errors approaching each other as the compliance approaches unity. For continuity the relation is less clear. The figure shows that the effect of imputation can be either to increase or decrease the forecast error.

the relation is less clear because when there is a long gap in the time series, a high continuity can co-occur with a low compliance.

4.3.3.2 Effect of gender

We apply out-of-sample forecasting to time series in Set G which comprises 20 male and 20 female patients with bipolar I disorder. The selection process for set G and its statistical properties are given in Chapter 3. As before, next step forecasts are made and the error is found by taking the RMSE over the residuals. The distribution of errors for three forecasting methods are shown in Figure 4.11 The forecast error distributions between males and females cannot be distinguished.

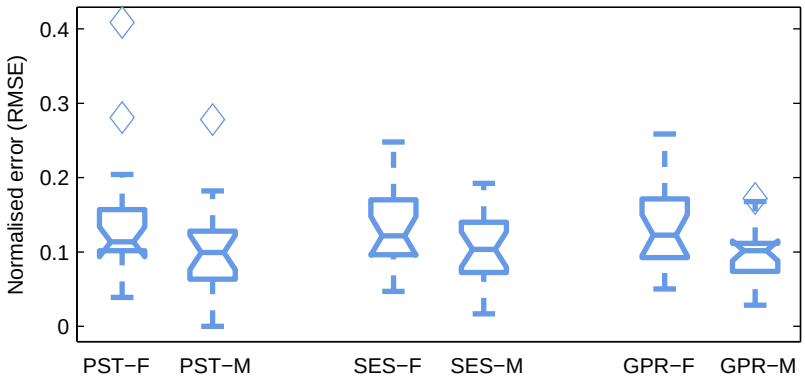


Figure 4.11: Distributions of out-of-sample errors for male and female patients in Set G. The error for each patient is summarized by RMSE. This loss function is applied to the individual next step forecasts for each patient. Errors are shown for persistence (PST), simple exponential smoothing (SES) and Gaussian process regression (GPR), first applied to the females (eg. PST-F) then to males (eg. PST-M). For the distribution PST-F, the notched comparison interval is below the 25^{th} percentile and for GPR-M, it is above the 75^{th} percentile. In each case this is because the sample size is small.

This is seen in Figure 4.11 where the notched 95% confidence intervals of the median overlap. A two-sample Kolmogorov-Smirnov test on the distributions confirms that they are not significantly different.

4.3.3.3 Effect of diagnosis

We apply out-of-sample forecasting to time series in Set D which comprises 16 patients with bipolar I disorder and 16 patients with bipolar II disorder. The selection process for set D and its statistical properties are given in Chapter 3. The distribution of errors for three forecasting methods are shown in Figure 4.12. The forecast error distributions between sets of patients with BPI and BPII cannot be distinguished. This is seen in Figure 4.12 where the notched 95% confidence intervals of the median overlap.

4.4 Conclusions

We have examined the stationarity and smoothness qualities of the depression time series and found a large variation among patients. By using exponential smoothing for in-sample forecasting we saw a range of dynamics from smooth to rapidly changing, random dynamics. In-sample forecasting reveals that many time series are forecast better by unconditional mean than by a persistence predictor. These time series show a rapid change in rating from week to week either

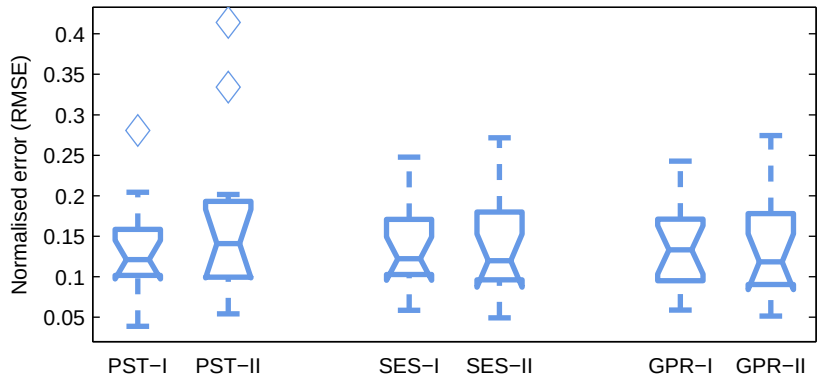


Figure 4.12: Distributions of out-of-sample errors for participants in Set G (n=32) with equal numbers of patients having bipolar I disorder and bipolar II disorder. Errors are shown for persistence (PST), simple exponential smoothing (SES) and Gaussian process regression (GPR), first applied to the 16 patients in the BPI set (eg. PST-I) then to the BPII set (eg. PST-II). For some of the distributions the notched comparison interval is below the 25^{th} percentile because the sample size is small.

because of a high signal to noise ratio or under-sampling: the mood is changing by day, rather than by week.

We applied many forecasting methods to determine the out-of-sample errors, but found the results to be the same for different methods. Using uniform time series did not change the results very much compared with those for the original, non-uniformly sampled series. We could find no difference between forecast errors for males and females, or between BPI and BPII patients.

Time series returned by the patients differ markedly in length, response interval, smoothness and stationarity. For in-sample forecasting simple exponential smoothing will always perform at least as well as unconditional mean or persistence forecasts. However, only 20 of the original 153 series improve by more than 10% over both persistence and unconditional mean forecasts. It is not surprising then, that it is hard to improve over the baselines for out-of-sample forecasting. Most of the time series are noisy or lack serial correlation, and gains from the minority of time series that benefit from smoothing will be masked by errors in other patients.

An increase in accuracy might be gained by partitioning the dataset and forecasting different groups separately but it seems unlikely that the improvement would be more than marginal when averaged over the whole set. For some patients, effective depression forecasts are not possible because the time series exhibit little serial correlation. Other patients can be forecast to a degree of accuracy,

but the clinical benefit is not clear because they form a minority of the whole set. We conclude, then, that effective forecasting is not possible for weekly depression data.

5

Mood dynamics

Introduction

In the previous chapter we concluded that effective forecasting was not possible using weekly depression data. The question remains whether depression forecasting is effective when using a more highly selected group of patients. This point is of less interest clinically because there seems to be little merit in being able to forecast mood only for a minority of patients. But while a highly selected set is less representative of the population it may provide some insight into the disorder that can be applied more generally.

This chapter examines longer time series for both dynamics and forecast error. The structure of the chapter is as follows. In section 1 we review the paper by Gottschalk *et al.* [55] and give a critique of its claims of chaotic dynamics in bipolar disorder. In section 2 we apply surrogate data tests for both randomness and nonlinearity to eight selected time series. Section 3 reviews Gaussian process regression and explains the choice of inference methods and likelihood function. Finally, section 4 applies out-of-sample forecasting to six selected time series. The forecast errors for linear and nonlinear forecasting methods are examined and related to the claims of stochastic nonlinearity made by Bonsall *et al.* [11] which were reviewed in Chapter 4.

5.1 Analysis by Gottschalk *et al*

Gottschalk *et al*. [55] analysed daily mood records from 7 patients with bipolar disorder and 28 normal controls. The 7 patients with bipolar disorder all had a rapid cycling course; that is, they had all experienced at least 4 affective episodes in the previous 12 months. Patients kept mood records on a daily basis (controls on a twice-daily basis) by marking mood on an analogue scale each evening to reflect average mood over the previous 24 hours. The selected participants kept records for a period of 1 to 2.5 years.

The authors used three kinds of surrogate data to generate reference time series 1) shuffle surrogates, where the original time series are randomly permuted, 2) spectrally similar surrogates, using phase randomisation in the frequency domain, 3) a nonlinear transformation of a linear stochastic process that corresponds to a permutation of the original. Three surrogates of each type were generated for each original time series.

The power spectra of the mood charts were analysed by plotting on a log-log scale, and the negative slope α was found to lie between 1.0 and 1.5 for most patients, but approximately 0.69 ± 0.16 for controls. The correlation dimension was estimated using the Grassberger-Procaccia method [57] and its convergence with embedding dimension, if any, was noted. Out of the 7 patients, 6 had correlation dimensions which converged at a value less than 5, while for controls, the convergence occurred no lower than 8. The surrogate time series did not show a convergence with embedding dimension.

5.1.1 Critique

From these results the authors inferred the presence of chaotic dynamics in the time series from bipolar patients. They noted the unreliability of correlation dimension as an indicator of chaos, but adduced the results from time plots, spectral analysis and phase-space reconstruction to demonstrate the difference from controls. Their claims were challenged by Krystal *et al.* [79] who pointed out that the power-law behaviour is not consistent with chaotic dynamics. In their reply [56], Gottschalk *et al.* said in reply that the spectra could equally be fitted by an exponential model. Further, the authors did not investigate the Lyapunov spectrum, which can provide evidence of chaotic dynamics. Their claim of deterministic chaos rested mainly on the convergence of correlation dimension and, as they acknowledged, this is not definitive [89]. Their change of model for spectral decay

weakened the original claims further – the evidence does not support nor deny them. The claims of stochastic nonlinearity which were reviewed in Chapter 4 show that identifying mood dynamics is challenging. In the case of Gottschalk *et al.*, a claim of chaotic dynamics was based on the convergence of correlation dimension, but this is a unreliable indicator.

5.2 Surrogate data analysis

In this section we review the theory of surrogate data analysis, then apply an analysis to uniform time series that are selected for length and stationarity. The analysis is in two parts, firstly to detect time series autocorrelation and secondly to detect nonlinearity.

5.2.1 Theory

The method of surrogate data [130] [122] [73] compares an original time series with a set of artificial surrogates, generated by a defined process and which share some property with the original. For example, to test for any serial correlation, we permute the original series n times to obtain a surrogate set having the same amplitude as the original but without any correlation structure. A test statistic is then applied and the results displayed graphically to demonstrate any difference between the original and surrogate time series. For the null hypothesis of white noise, we use the autocorrelation at varying lags as a test statistic.

We consider the null hypothesis of a linear stochastic model with Gaussian inputs. If this cannot be rejected then there is a question over the use of more complex, nonlinear models for forecasting. For this analysis the surrogate data must be correlated random numbers with the same power spectrum as the original data. This is a property of data which has the same amplitudes as the original data but different phases. Following [73, p112], we take the FFT of the original series,

$$\tilde{s}_k = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} s_n e^{i2\pi nk/N} \tag{5.1}$$

We then multiply the complex components $\tilde{s}_k$ by uniformly distributed random phases, $\tilde{s}'_k = \tilde{s}_k \, e^{i\phi_k}$, $0 < k < N$, then take the inverse FFT,

$$s'_n = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \tilde{s}'_k e^{-i2\pi nk/N} \tag{5.2}$$

We then assume that the stationary Gaussian linear process is distorted by a time independent function $s_n = s(x_n)$ where x_n is the linear process. We invert s by rescaling the values to follow a Gaussian distribution $\hat{x}_n = \hat{s}^{-1}(s_n)$. The phase randomisation is done on the $\hat{x}_n$ giving surrogates $\hat{x}'_n$ which are then rescaled to the original distribution of the s_n. These *amplitude adjusted Fourier transform* (AAFT) surrogates have a slightly different power spectrum from the original series because we have to estimate the untransformed linear process. To make the surrogates match the original spectrum more closely, an iterative filter can be applied [121]. This method (IAAFT) works by approximating the original linear correlation at the first step and the original distribution at the second step. We summarise the kinds of surrogate data generation methods and corresponding null hypothesis as follows,

- Algorithm 0 or shuffle surrogates test the null hypothesis that the data is independent and identically distributed.
- Algorithm 1 or shuffled Fourier transform surrogates test for linearly filtered noise.
- Algorithm 2 or amplitude adjusted Fourier transform surrogates (AAFT) test for a monotonic nonlinear transformation of linearly filtered noise. The IAAFT algorithm matches the original power spectrum more closely.

We note that Algorithm 1 surrogates are not appropriate for our data, which is discrete. Another variant of Algorithm 2, which does not assume that the static transform is monotonic and which can result in less bias in the correlation is the corrected AAFT (CAAFT) [80]. This algorithm uses an AR model to generate a time series which is transformed through inverse gaussianisation to match the original distribution exactly and the original linear correlations on average. Other methods for surrogate generation exist, for example for a null hypothesis of a periodic orbit with uncorrelated noise [125]. However, the level of seasonality in the time series is not sufficient to justify their use in this analysis.

5.2.2 Data selection

As in previous chapters, we use data collected as part of the *OXTEXT* programme which uses the *True Colours* self-management system for mood monitoring. A description of this data is given in Chapter 3 along with descriptive statistics. The process of data selection differs from in previous Chapters and is as follows:

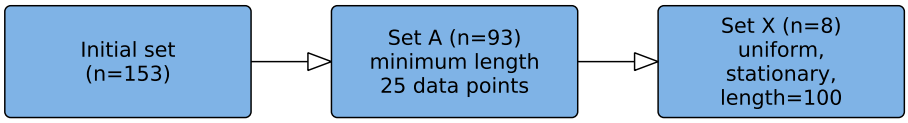


Figure 5.1: The initial set is 153 time series provided by the Department of Psychiatry. From this data eight stationary, uniform time series with a minimum length of 100 points are selected for use in the surrogate data analysis.

1. From the initial set of 153 patients, we removing repeated response values by taking the most recent in the sequence. We then select those members whose time series have at least 25 data points, or approximately six months duration to create Set A (n=93). This process is the same as in Chapter 3, which also gives descriptive statistics for Set A. We then resample the time series in Set A to an exact weekly sampling interval using the method described in Figure 3.7 in Chapter 3.

2. We select time series with a minimum resampled length of 100 data points having 5 or fewer imputations in their first 100 points. We truncate the time series to 100 points and test for stationarity using a Kolmogorov-Smirnov test on the first and second halves of the time series. The eight patients which show no significant difference at the 5% level constitute Set X. Figure 5.1 summarizes the data selection process.

The statistical qualities of the eight selected patients are shown in Table 5.1, which includes Set A for comparison. Time plots of depression time series used in this section are shown in Figure 5.2.

Set (n)	*Gender F/M*	*Subtype BP I/II/?*	*Age in years*	*Length points*	*Depression QIDS*
A (n=93)	60/33	41/22/30	45 (20)	136 (161)	6.4 (5.3)
X (n=8)	6/2	4/2/2	48 (28)	181 (67)	5.1 (2.5)

Table 5.1: Median (interquartile range) of age, length and mean depression with demographic summary for the eight selected patients. Set A is shown for comparison.

5.2.3 Autocorrelation

We first examine the sample autocorrelation function for the eight selected time series and compare it with that for surrogates which have the same distribution.

Figure 5.3 shows the results, with the sample acf shown for lags 1 to 6. Seven of the eight time series are distinct from their shuffle surrogates, showing that they have some serial correlation. However the time series in the top left panel of Figure 5.3 cannot be distinguished from its permutations and thus is unsuitable for the purposes of forecasting. We keep this time series in the experimental set for reference.

Before testing for nonlinearity we compare the autocorrelation of CAAFT surrogates with the eight original time series to ensure that the surrogates reproduce the original autocorrelation. The software from [80] is used to generate the CAAFT surrogates. Figure 5.4 shows the sample autocorrelation of both the original time series and the surrogates. It can be seen that the surrogates reproduce the original autocorrelation well apart from those in the lower right panel which exhibit a negative bias.

5.2.4 Nonlinearity

For the test of nonlinearity we compare test statistics on the original and CAAFT surrogates. We use two kinds of test statistic. The first is the ratio of linear *vs.* nonlinear in-sample forecast error. If there is nonlinearity present in the original time series we would expect its ratio of nonlinear to linear forecast error to be

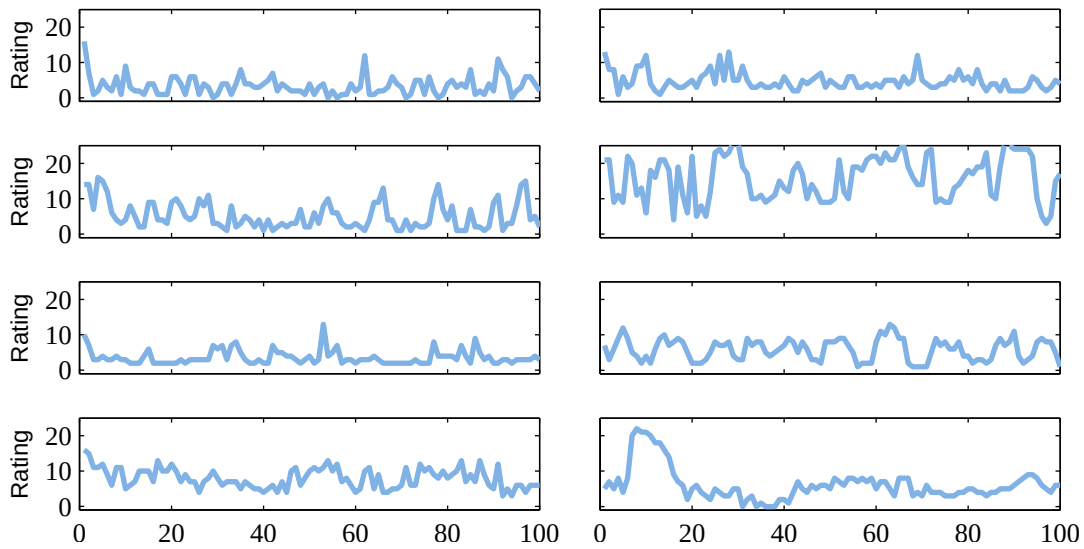


Figure 5.2: Depression time series from the eight patients used in the study. Patients are numbered 1–8 starting top-left and scanning across and down. Each plot shows depression ratings sampled on an approximately weekly basis.

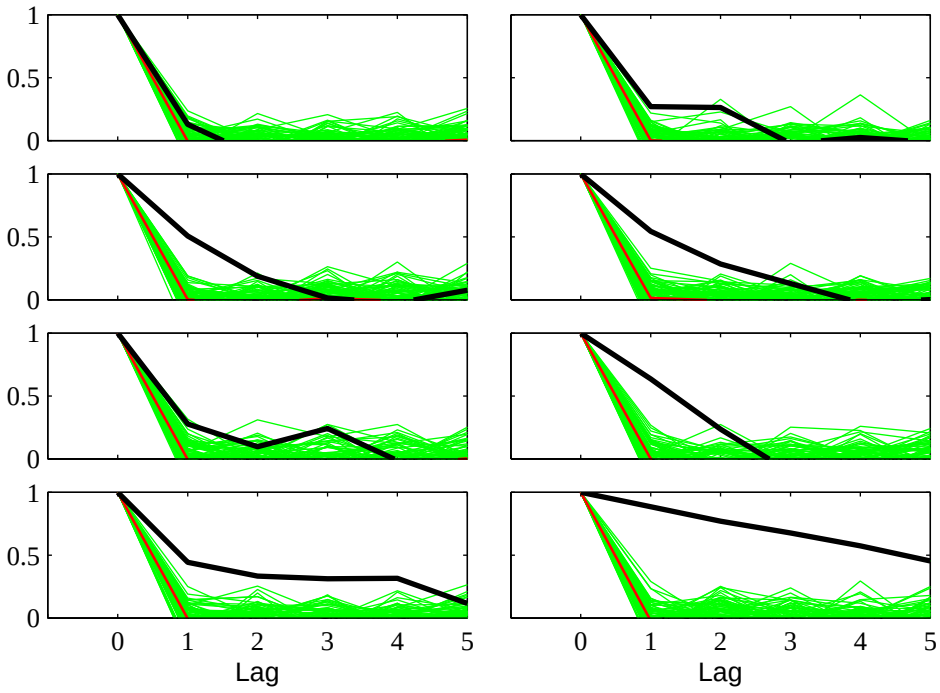


Figure 5.3: Sample autocorrelation statistic for the eight original time series and their shuffle surrogates. For each of time series, surrogates are created by randomly permuting the original time series. The sample autocorrelation function is shown for the surrogates and that for the original series is shown in a dark line. The thin red line shows the median of the surrogates. The top left plot shows a time series that is indistinguishable from white noise while the other time series show serial correlation.

distinct from that in its surrogates. The second test statistic is a time reversal asymmetry statistic which detects time asymmetry.

5.2.4.1 Forecast error statistic

We look at the ratio between linear and nonlinear in-sample forecast errors. The linear methods used are persistence and autoregressive models order 1 and 2. The nonlinear methods are the simple nonlinear predictor (SNL) and the local linear predictor (LLP). These methods are explored further in Chapter 6 in the context of k-nearest neighbour forecasting. They are implemented with the *TISEAN* functions `lzo-test` and `lfo-test` [61]. For each time series we forecast each point using the model and return the rms error divided by the standard deviation of the data. We then plot the ratio of linear to nonlinear error for both original and surrogate time series. Figures 5.5 – 5.8 show the results of these comparisons.

For each plot the results for the surrogate time series are displayed as a histogram and that for original series shown as a vertical line. Figure 5.5 shows the

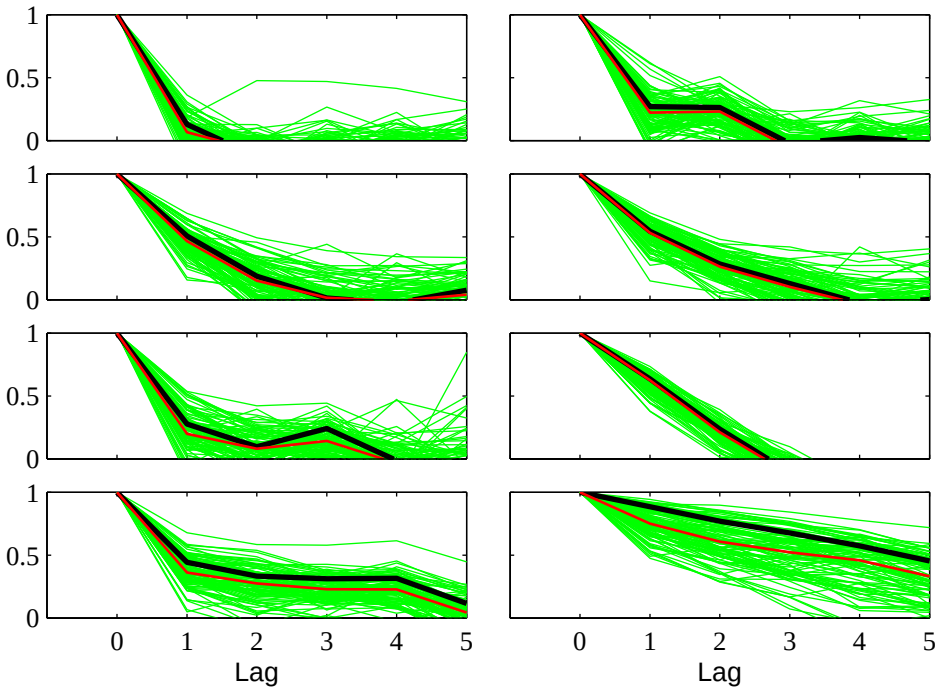


Figure 5.4: Sample autocorrelation statistic for the eight original time series and their CAAFT surrogates. The autocorrelation of the original time series is shown as a thick black line and the median of the surrogates is shown as a thin red line. In most cases the original time series is typical among the surrogates with the exception of the plot on the lower right which shows a downward bias in the surrogates.

distribution of the persistence/SNL error ratios. None of the original time series appears to benefit from nonlinear forecasting and the bottom-right histogram shows the time series is better predicted by persistence than is the case for its surrogates. This is a result of the lower average correlation among the surrogates for this time series which was shown earlier in Figure 5.4. The broad distribution of the test statistics for this time series is a result of the transient at its start. If the high values in the transient appear as spikes in the surrogate, there is a high linear forecast error, but if they are more contiguous, the linear error is lower.

Figure 5.6 shows the distribution of AR(1)/SNL errors. In the bottom left histogram the original time series shows a lower error ratio than its surrogates. This result is repeated in Figure 5.7 which uses AR(1)/LLP errors. Finally, Figure 5.8 shows the ratio of AR(2)/LLP and again shows no evidence of nonlinearity for any of the time series in the sample. In summary none of the time series is consistently distinct from its surrogates and there is no general evidence of nonlinearity.

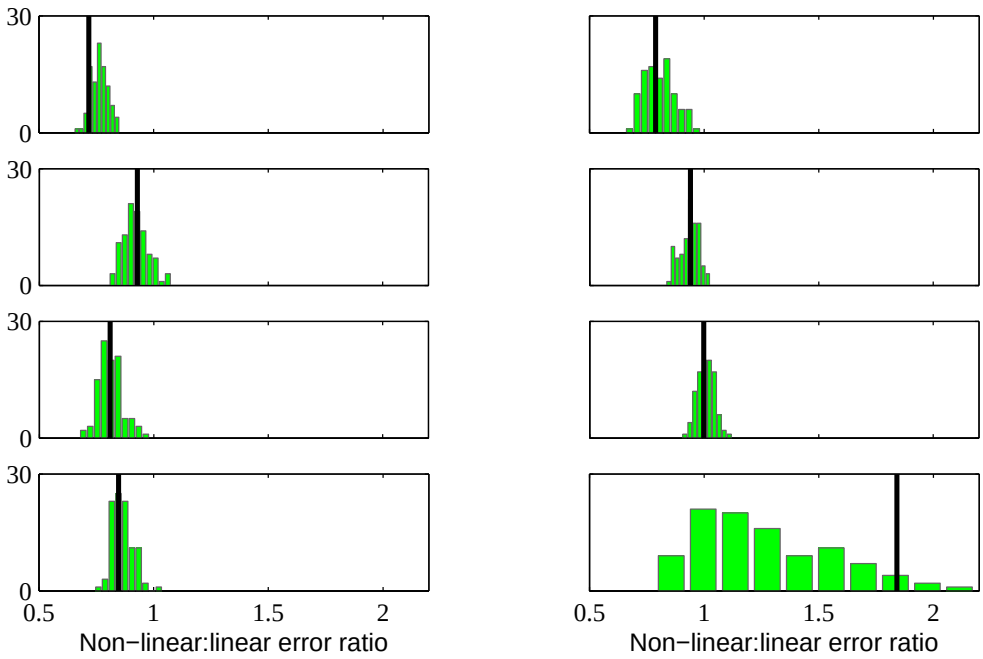


Figure 5.5: In-sample error ratio for a nonlinear method (SNL) vs persistence forecasting. The surrogates are shown as a histogram and the original time series as a dark line. The panel on the bottom-right shows that the original series is relatively better forecast by the linear than the nonlinear method. This anomaly is caused by the surrogates having an artificially lower average correlation, shown in the equivalent panel in Figure 5.4.

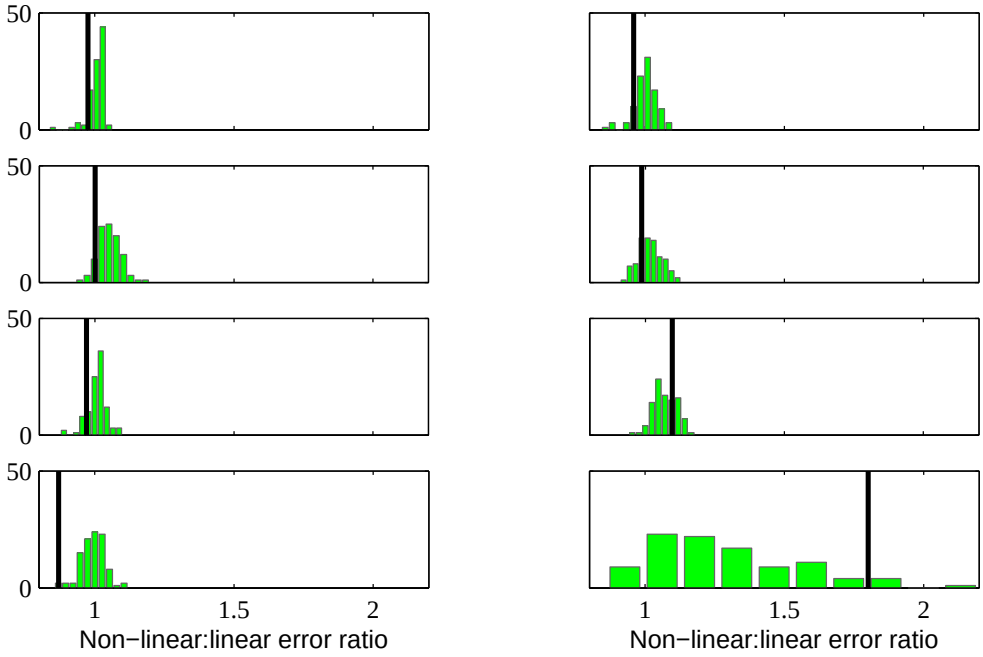


Figure 5.6: In-sample error ratio for a nonlinear method (SNL) vs AR1 forecasting. The time series on the bottom left is forecast better by a nonlinear method than its surrogates.

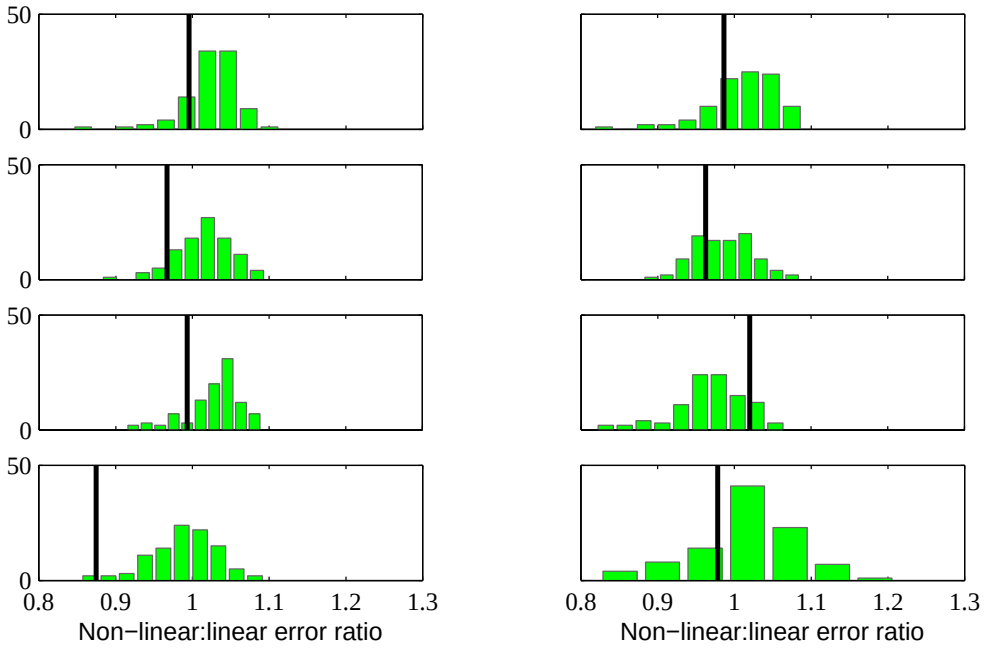


Figure 5.7: In-sample error ratio for local linear vs AR1 forecasts. The original time series in the bottom left panel is better forecast by a nonlinear method than its surrogates.

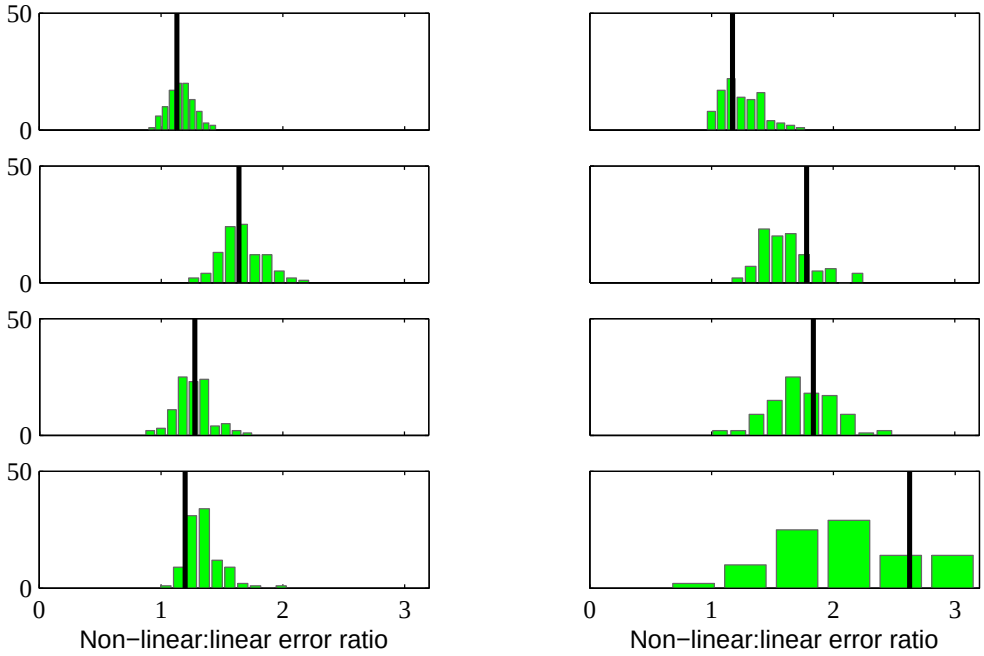


Figure 5.8: In-sample error ratio for local linear vs AR2 forecasts. None of the time series is distinguished from its surrogates.

5.2.4.2 Time asymmetry statistic

Finally, we compare the original and surrogate time series using a time reversal asymmetry statistic. Asymmetry of the time series when reversed in time can be a signature of nonlinearity [130]. A measure of time reversibility is the ratio of the mean cubed to the mean squared differences,

$$\mathcal{Q} = \frac{\mathcal{E}[(y_{i+1} - y_i)^3]}{\mathcal{E}[(y_{i+1} - y_i)^2]} \tag{5.3}$$

Figure 5.9 shows the $\mathcal{Q}$ statistic values for the surrogates shown as a histogram and the statistic for the original series shown as a vertical line. The only time series which appears distinct is the one on the bottom right, but in general there is no evidence of time asymmetry in the eight time series under consideration.

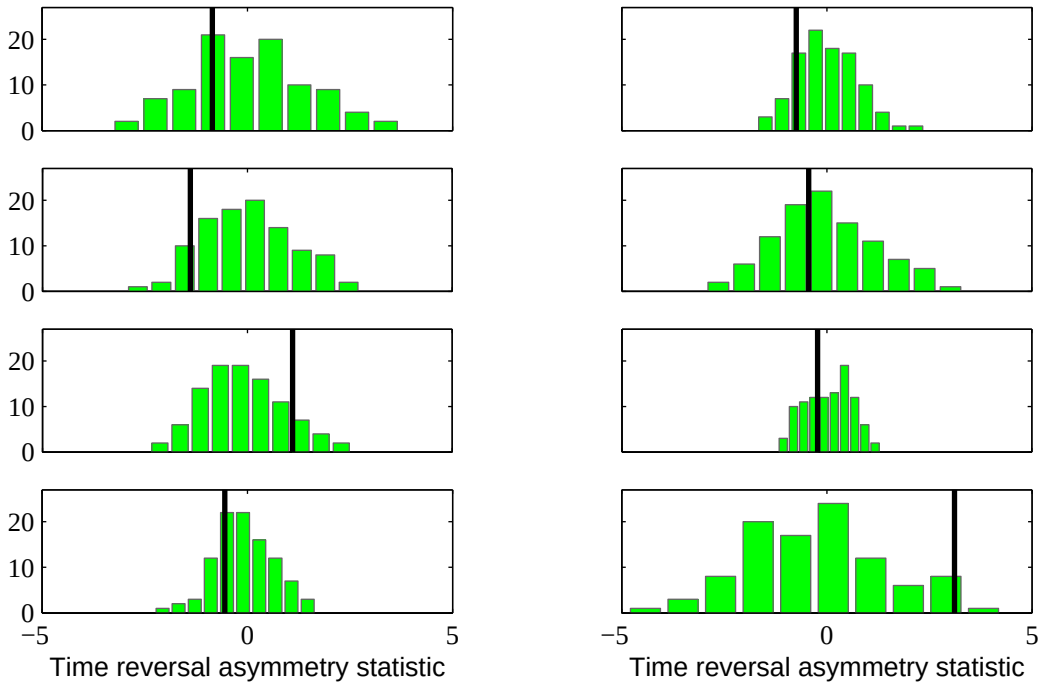


Figure 5.9: Time reversal asymmetry statistic. The statistic for the original time series is represented as a vertical line and that for the surrogates is shown as a histogram. Again there is no evidence of nonlinearity except for the bottom right panel.

5.3 Forecasting

In this section, we perform out-of-sample forecasting using just six uniform time series which are stationarity and have a minimum length of 150 points. The co-

variance function and inference method for Gaussian process forecasting is carefully selected, and the results are compared with those of other forecasting methods.

5.3.1 Data selection

The process for data selection is different from that in previous sections. Here we select stationary, uniform time series with no more than 10 imputations in the first 150 points. The test for stationarity is between the first 100 points and the remaining 50 points. The six time series that fulfil these criteria constitute Set Y. Figure 5.10 summarizes the selection process.

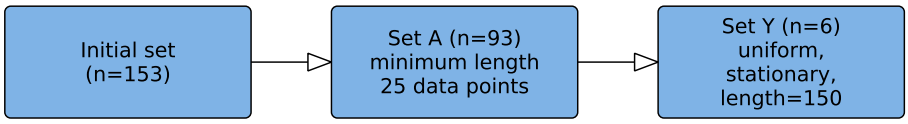


Figure 5.10: The initial set is 153 time series provided by the Department of Psychiatry. From this data six stationary, uniform time series with a minimum length of 150 points are selected.

The statistical qualities of the six selected time series are shown in Table 5.2, which includes Set A for comparison. The time plots of depression time series used in this section are shown in Figure 5.11.

Set (n)	*Gender F/M*	*Subtype BP I/II/?*	*Age in years*	*Length points*	*Depression QIDS*
A (n=93)	60/33	41/22/30	45 (20)	136 (161)	6.4 (5.3)
Y (n=6)	4/2	2/4/0	45 (30)	210 (52)	5.2 (9.9)

Table 5.2: Median (interquartile range) of age, length and mean depression with demographic summary for the six selected patients. Set A is shown for comparison.

5.3.2 Gaussian process regression

The theory for Gaussian process models is described in Chapter 2 where it is applied to forecasting mood. The method assumes a Bayesian nonparametric model where the regression function itself has a prior distribution. The distribution is a Gaussian process which is specified by a covariance function $k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta})$ to define the correlation between latent function values at inputs $\mathbf{x}$ and $\mathbf{x}'$. Properties of

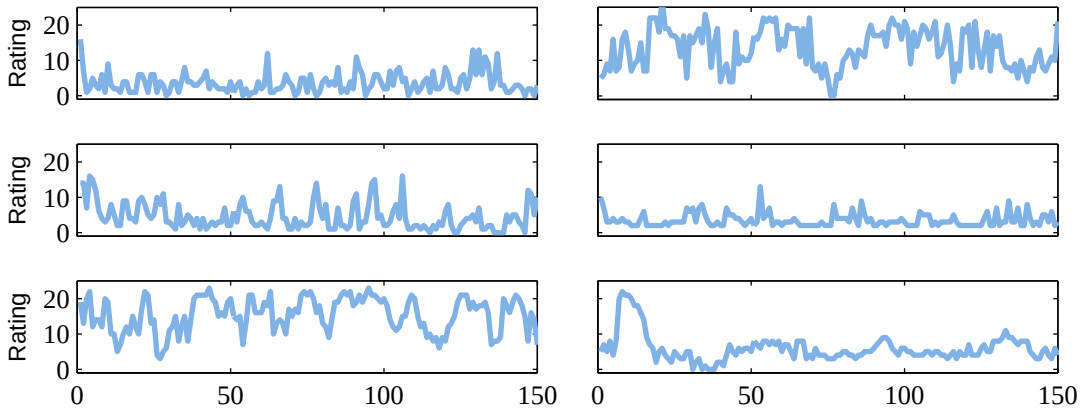


Figure 5.11: Time series from the 6 patients used in the study. Patients are numbered 1–6 starting top-left and scanning across and down. Each plot shows depression ratings sampled on an approximately weekly basis. The maximum possible score on the rating scale for depression (QIDS) is 27.

the prior process, such as the length scale, are determined by hyperparameters $\boldsymbol{\theta}$, which are estimated from the data by maximum likelihood along with a noise term σ_n. The predictive equation is then $\bar{f}_* = \mathbf{k}_*^{\mathbf{T}}(K + \sigma_n^2 I)^{-1}\mathbf{y}$ where $\mathbf{k}_*$ is the vector of covariances of the test point with the training points $\mathbf{y}$, and K is the covariance matrix of the training set.

An example of draws from a Gaussian process distribution is shown in the two upper panels of Figure 5.12. They are generated artificially using the method described in [113]. A Matérn covariance function of order 3 is chosen and the length

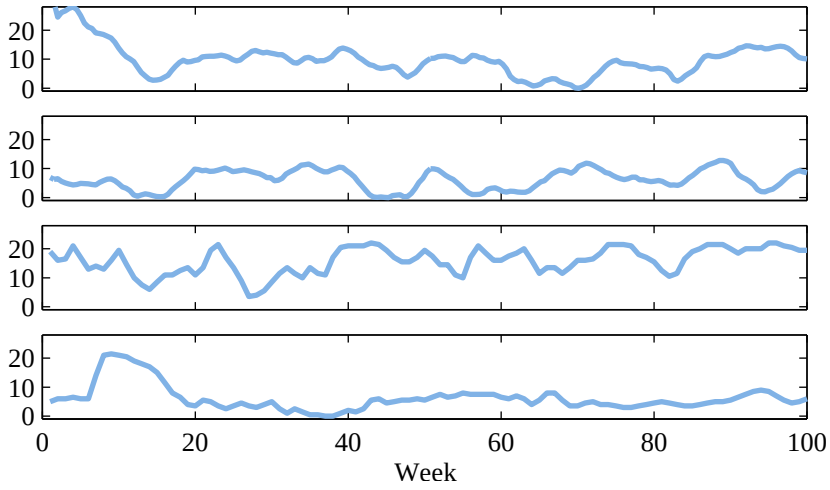


Figure 5.12: Sample draws from a Gaussian process. The upper two plots are realisations of a Gaussian process with a Matérn covariance function $k(r) = \sigma^2\left(1 + \frac{\sqrt{3}r}{l}\right)\exp\left(-\frac{\sqrt{3}r}{l}\right)$ with length scale $l = 10$ and standard deviation $\sigma = 5$. The two lower time plots are moving-average smoothed mood time series shown for comparison.

scale and signal variance are set to fixed values. By way of comparison, the two lower plots in Figure 5.12 are moving-average smoothed depression time series from the OXTEXT data. The artificial series match the smoothed mood series for time scale and variance because these hyperparameters were chosen accordingly. In forecasting, the training set is used to determine the hyperparameter values, which are then used in the forecast equation (2.58).

5.3.2.1 Inference methods and likelihood

The hyperparameters are chosen from the training data by optimising the marginal likelihood given in (2.60) and restated in a different form here,

$$p(\mathbf{y}|\mathbf{t}) = \int p(\mathbf{y}|\mathbf{f},\mathbf{t})\, p(\mathbf{f}|\mathbf{t})\, d\mathbf{f} \tag{5.4}$$

This is the integral of the product of the likelihood $p(\mathbf{y}|\mathbf{f},\mathbf{t})$ and a Gaussian prior $p(\mathbf{f}|\mathbf{t})$. The posterior over the parameters is the normalised product of these two terms. If the likelihood is a Gaussian function, then the marginal likelihood can be found analytically and is given by (2.62). This is an exact inference because there is no approximation of the posterior. When the likelihood is non-Gaussian, the marginal likelihood $p(\mathbf{y}|\mathbf{t})$ can be found by using a number of approximations to the posterior. We apply three approaches in this section all of which are implemented in the GPML software, version 3.2 [112]. The first is to use leave-out-one cross-validation (*LOO*), described in Section 2.2.3 and applied to Gaussian processes in [127]. The second method is to estimate the posterior using Laplace's method [136] and the third method to use an algorithmic approach, variational Bayes, to approximate the integral.

A likelihood function is a conditional density $\mathcal{P}_\rho(y|f)$ which relates the scalar latent function values f_i to the outputs y_i. It is parameterised by hyperparameters $\boldsymbol{\rho}$ which are learned by minimising (5.4). We apply two likelihood functions, the $sech^2$ and Student's t, both distributions having heavier tails than the Gaussian function. In the implementation used [112], the $sech^2$ likelihood is parameterised by the variance σ_n and the Student's t distribution uses two hyperparameters σ_n and ν.

Table 5.3 summarises the forecast methods for Gaussian process regression used in this chapter. Three kinds of covariance function are used, squared exponential (*SQE*), rational quadratic (*RQU*) and Matérn (*MAT*). For the Matérn covariance function, we use $sech^2$ and t likelihood distributions in addition to

Abbreviation	*Description*	*Covariance*	*Lik.*	*Inf.*
SQE	Sqr. exponential	$\sigma^2 \exp\left(-\frac{r^2}{2l^2}\right)$	Gauss	Exact
RQU	Rat. quadratic	$\sigma^2 \left(1+\frac{r^2}{2\alpha l^2}\right)^{-\alpha}$	Gauss	Exact
MAT1	Matérn	$\sigma^2 \left(1+\frac{\sqrt{3}r}{l}\right) \exp\left(-\frac{\sqrt{3}r}{l}\right)$	Gauss	Exact
MAT2	Matérn	"	$sech^2$	Laplace
MAT3	Matérn	"	$sech^2$	LOO
MAT4	Matérn	"	t	VB

Table 5.3: Gaussian process forecast methods. In the covariance functions r is the difference between time indices, l is the length scale and α is a scale factor.

a Gaussian likelihood. The non-Gaussian likelihood functions require approximate parameter inference, and cross-validation (*LOO*), Laplace approximation and variational Bayes are used.

5.3.2.2 Model selection

We select the covariance function, likelihood function and inference method by examining the likelihood of the Gaussian process on the training data, and by considering the out-of-sample error. The negative log likelihood from training the first 100 points of each of the six time is shown in Table 5.4. The *MAT3* method with $sech^2$ likelihood and using cross-validation inference has the lowest negative log-likelihood, but otherwise there is little to choose between the methods.

Table 5.5 shows the out-of-sample RMSE forecast error for the different covariance functions and inference methods. The Gaussian process is trained on the first 100 points of the time series and used to predict the next 50 points. Whereas in training, the *MAT3* method was distinguished from the others, it performs no better in forecasting. There is little to distinguish the other methods: *MAT2* has the lowest mean out-of-sample error, but by a very small margin. We therefore choose *MAT2*, which uses a Matérn covariance function with $sech^2$ likelihood and Laplace inference.

5.3.2.3 Retraining

Figure 5.13 shows the result of forecasting using a single time series chosen from the six selected for forecasting. The hyperparameters are trained on the first 100 points of the time series and next-step predictions made for the next 50 values.

	SQE	*RQU*	*MAT1*	*MAT2*	*MAT3*	*MAT4*
	Gauss	*Gauss*	*Gauss*	*sech*^{2}	*sech*^{2}	*t*
Time series	*Exact*	*Exact*	*Exact*	*Lpce*	*LOO*	*VB*
1	240.8	240.9	240.9	235.3	235.6	240.1
2	295.4	295.9	296.1	296.6	286.2	338.5
3	263.0	263.4	263.3	266.7	250.2	263.6
4	206.7	207.2	206.5	197.7	195.8	194.5
5	272.3	272.2	272.1	274.1	255.2	297.8
6	223.1	220.8	220.0	219.6	188.1	222.3
Mean	250.2	250.1	249.8	248.3	235.2	259.5

Table 5.4: Negative log likelihood for different covariance functions, likelihoods and inference methods applied to the six time series selected for forecasting.

	PST	*SQE*	*RQU*	*MAT1*	*MAT2*	*MAT3*	*MAT4*
		Gauss	*Gauss*	*Gauss*	*sech*^{2}	*sech*^{2}	*t*
Time series		*Exact*	*Exact*	*Exact*	*Lpce*	*LOO*	*VB*
1	3.89	3.30	3.39	3.32	3.32	3.33	3.13
2	5.82	4.81	4.77	4.76	4.86	4.89	5.38
3	3.79	3.37	3.36	3.39	3.21	3.33	3.40
4	2.55	1.90	1.91	1.87	1.88	2.06	1.91
5	3.50	3.30	3.34	3.31	3.26	3.36	3.27
6	1.41	1.41	1.39	1.40	1.44	1.44	1.42
Mean	3.49	3.02	3.02	3.01	3.00	3.07	3.09

Table 5.5: Out-of-sample forecast error (RMSE) for different Gaussian process methods. Persistence forecast error (PST) is shown for comparison.

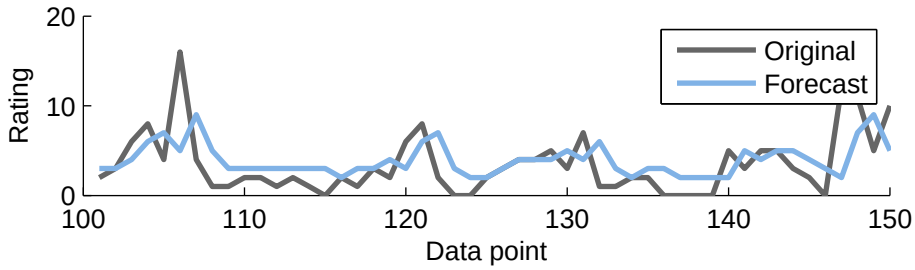


Figure 5.13: Gaussian process forecasting using a Matérn covariance function. The hyperparameters are trained on the first 100 data points to give a length scale $l = 1.8$ and a process standard deviation of $\sigma = 3.2$. The original series is shown as a dark line and the prediction as a light blue line.

As the forecast point moves through the test set, it is possible to retrain using a longer training set. For example, once the first point has been forecast, the training set can comprise 101 points. In this section, we retrain the hyperparameters periodically during forecasting to see if this improves accuracy.

Figure 5.14 shows the forecast error for each of the six time series using different retraining schedules. A retraining period of 10 means that the parameters are trained on the first 50 points, then 60 and so on. A period of 50 implies that the hyperparameters are trained only once because the test sets are 50 points in length.

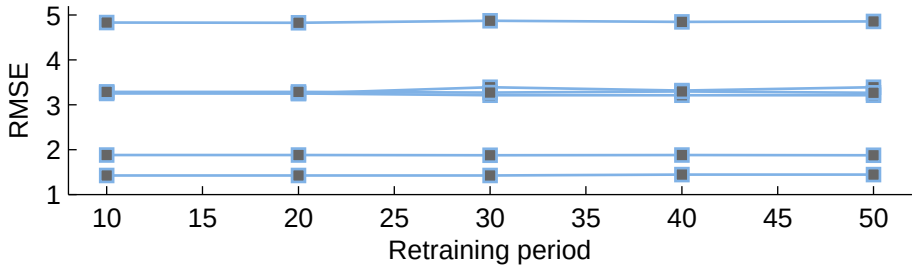


Figure 5.14: Next-step forecast error plotted against retraining period for Gaussian process regression for the six time series. In principle more frequent retraining could reduce error by better reflecting the properties of the time series. In practice this is not found to be so. A Matérn covariance function with $sech^2$ likelihood and Laplace approximation is used.

Three hyperparameters have to be learned in this case: the likelihood parameter, signal variance and length scale. The results show that retraining makes no difference as the potential training set length gets larger. The process length scale in particular is typically smaller than the 100 weeks initial training length, as shown in Figure 5.12, lower two panels. This suggests that it should optimise

during the 100 weeks initial training, and that more training will not substantially change its value.

5.3.3 Forecasting

In this section we examine out-of-sample forecast error for a range of forecasting methods, including Gaussian process regression. The methods used are similar to those applied in Chapter 3, but with some additions. As before persistence (*PST*), simple exponential smoothing (*SES*), and autoregression orders 1 (*AR1*), and 2 (*AR1*) are used. Gaussian process regression using a Matérn covariance function with $sech^2$ likelihood and Laplace inference (*MAT2*) is used. The simple nonlinear (*SNL*), and local linear (*LLP*) methods are used. Some variants of the Bonsall *et al.* regime switching methods [11] are applied. These methods use an autoregressive model with different regimes above and below the mean. In the earlier application, the model orders were trained from the data and could vary between patients and regimes. Here we use that approach (*BNN*) but also methods (*BN1*) and (*BN2*) which fix the orders at 1 and 2 respectively for all patients and regimes, training only the model parameters from the data. The methods are summarised in Table 5.6.

Abbreviation	*Method*	Definition
PST	Persistence baseline	
SES	Simple exponential smoothing	Section 2.3.3.2, p 42.
AR1	Autoregressive model order 1	Section 2.3.3.1, p 40.
AR2	Autoregressive model order 2	Section 2.3.3.1, p 40.
MAT2	Gaussian process regression, Matérn	Section 2.4.1, p 45.
SNL	Simple nonlinear predictor	Section 6.1, p 115.
LLP	Local linear predictor	Section 6.1, p 115.
BN1	Bonsall order 1	Section 5.3.3, p 112.
BN2	Bonsall order 2	Section 5.3.3, p 112.
BNN	Bonsall order n	Section 5.3.3, p 112.

Table 5.6: Forecast methods applied to the six selected time series

We apply all the methods to the six uniform, stationary time series of length 150 points. Forecast methods are trained on the first 100 data points of the individual time series and the next step forecasts made for each of the next 50 points. Each patient's error is estimated by applying a root mean square error (RMSE) loss function to the next step errors.

Time series	*PST*	*SES*	*AR1*	*AR2*	*MAT2*	*SNL*	*LLP*	*BN1*	*BN2*	*BNN*
1	3.89	3.34	3.26	3.43	3.32	3.32	3.34	3.17	3.40	3.40
2	5.82	4.93	4.93	4.82	4.86	4.90	4.91	5.44	5.20	5.43
3	3.79	3.39	3.28	3.32	3.21	3.42	3.36	3.44	3.55	3.43
4	2.55	1.80	1.86	1.85	1.88	1.86	1.89	1.92	1.91	1.91
5	3.50	3.59	3.32	3.33	3.26	3.70	3.51	3.34	3.37	3.34
6	1.41	1.41	1.36	1.37	1.44	1.52	1.45	1.33	1.37	1.32
Mean	3.49	3.08	3.00	3.02	3.00	3.12	3.08	3.11	3.13	3.14
Median	3.65	3.37	3.27	3.33	3.24	3.37	3.35	3.26	3.39	3.37

Table 5.7: Out-of-sample forecast errors (RMSE) for six stationary, uniform time series with a length of 150 points. Lowest values are highlighted.

Table 5.7 shows the out-of-sample forecast results. There is little difference in accuracy between the forecasting methods. The range in error between best *MAT2* and the worst method *PST* is less than 0.5 of a rating unit. So we find that the nonlinear forecast methods do not outperform linear methods for the selected time series and methods.

We apply a Diebold-Mariano test [34] to compare the predictive accuracy of the different forecasting methods with the persistence baseline. This test is applied to the forecast errors from each model, and is described in Section 2.1.3.3. The test statistic is shown in Table 5.8 for the out-of-sample forecast results using an identity loss function in (2.12). Time series 2 and 4 exhibit a significant difference between the accuracy of persistence forecasting compared with other methods: this difference can also be seen in Table 5.7 as a relatively high error for persistence forecasting. However for most patients, the other forecasting methods show no more accuracy than persistence.

5.4 Conclusions

We have found that the depression time series cannot be distinguished from their surrogates generated from a linear process when comparing the respective test statistics. These results can mean that either 1) the original series have linear dynamics, 2) the test statistics for distinguishing linear from nonlinear behaviour do not have the power to detect the kind of nonlinearity present, 3) the process is nonlinear but the sampling is inadequate or the time series are too short to

Time series	*SES*	*AR1*	*AR2*	*MAT2*	*SNL*	*LLP*	*BN1*	*BN2*	*BNN*
1	1.41	1.63	1.09	1.56	1.42	1.44	2.10	1.25	1.25
2	2.54	2.91	2.47	2.66	2.00	2.12	1.87	2.33	1.86
3	1.27	1.44	1.34	1.51	0.84	1.25	1.04	0.75	1.00
4	2.81	3.15	3.09	2.86	2.92	3.03	3.02	2.96	3.00
5	0.81	0.89	0.84	1.19	-0.77	-0.02	0.76	0.67	0.77
6	0.00	1.72	1.41	-0.85	-0.65	-0.41	2.04	0.72	2.17

Table 5.8: Diebold-Mariano test statistic for out-of-sample forecast results. Absolute values of greater than 1.96 show a difference in predictive accuracy at the 5% level. Time series 2 and 4 exhibit a significant difference for the majority of methods.

represent the dynamics. It is uncertain which hypothesis is to be preferred, but it would be worthwhile repeating the analysis on longer, more frequently sampled data. On the current evidence, though, there is no reason to claim for any nonlinearity in mood time series that we examined.

We find that nonlinear forecasting methods have no advantage over linear methods in out-of-sample forecasting even on a highly selected set of time series. This result contrasts with the claim in Bonsall *et al.* [11] that weekly time series from patients with bipolar disorder are described better by nonlinear than linear processes. The discrepancy between the results in this chapter and those reported by Bonsall *et al.* [11] may be attributed to differences between data and methods. Although the data in both cases is from the same source and is qualitatively the same, the sets of patients are different. Importantly, whereas the time series used in this chapter are uniform, those used by Bonsall *et al.* are not. The difference in methods is also important: Bonsall *et al.* [11] use only autoregressive and threshold autoregressive models, neither of which compensates for non-stationarity, whereas a wide variety of the methods is used here.

The question remains as to what kind of stochastic process best describes the weekly data. The relatively better performance of the linear methods suggests a low order autoregressive process or a random walk plus noise model [19, §2.5.5], for which simple exponential smoothing is optimal [19, §4.3.1]. However, the similarity of performance across methods and the poor forecast accuracy preclude the selection of any one process as describing the time series.

6

Nearest neighbour forecasting

Introduction

Previous chapters have covered data analysis, forecasting and mood dynamics. This chapter introduces k-nearest neighbour models for time series and presents a new forecasting algorithm. First we provide some theoretical background to the k-nearest neighbour method and its application to time series forecasting. Next we explain the details of the new algorithm and how it relates to existing methods. Finally we evaluate the algorithm on synthetic time series, electrocardiogram (ECG) data and the bipolar depression data.

6.1 K-nearest neighbour forecasting

The k-nearest neighbour method is a form of instance or memory based learning. Instead of using a training set to estimate model parameters, instances are selected from the set to use as predictors. The method is commonly used in classification as a low bias function approximator [59], and its use in regression is described in Section 2.2.1. In time series forecasting, the algorithm selects past sequences which are similar to the current sequence and uses their successor points to make a point prediction. A theoretical analysis of the method from a time series perspective is given in by Yakowitz [141] and practical examples of the method's application to finance, energy and hydrology are given by Arroyo[6]. Fernandez-Rodriguez *et al.* [43] apply nearest neighbour methods to forecasting to nine european currencies and find that it outperforms a random walk model.

Barkoulas *et al.* [7] apply similar methods (locally weighted regression) to forecasting US interest rates and find an improvement in prediction accuracy over linear benchmarks. Arora *et al.* [5] provide a review of nearest neighbour forecasting methods in the more general context of financial forecasts.

The method has frequently been used for hydrology applications such as, for example, Karlsson and Yakowitz's work on rainfall runoff forecasting [74]. Wu *et al.* [140] find that the k-nearest neighbour method outperforms both neural network and ARMA models for forecasting river flow. Conversely, Toth *et al.* [132] apply the method to flood prediction and find that neural networks outperform k-nearest neighbour prediction. The k-nearest neighbour forecasting method is also related to methods motivated by a dynamical systems approach and to compression-based forecasting.

Compression-based forecasting methods rely on the relation between the predictability of a sequence and its compressibility. In principle any technique based on information criteria such as AIC and BIC[1] or regularisation methods could be said to employ compression. However, forecasting based on compression algorithms such as Lempel-Ziv has not been widely used because time series are usually considered to be too noisy for the method to be effective. An unusual example of the use of compression software to forecast currency exchange rates is the study by Ryabko *et al.* [117].

In the context of dynamical systems theory, the equivalent forecasting method is sometimes called the *method of analogues*, and it is related to both probabilistic and non-probabilistic compression-based methods, and state space based regression methods. This approach embeds the time series in a state space using delay co-ordinates and learns the nonlinear function using local approximation [39]. Method of analogues approaches have been applied in finance [5][42] and they are the basis of the *TISEAN* methods used for forecasting depression in Chapter 2.3.3.3. Their theory is now described in more detail.

6.1.1 Method of analogues

The method of analogues was originally proposed as a nonlinear forecasting method by Lorenz [86]. Here we follow the argument in Kantz *et al.* [73, p50], for the development of the method, and describe some forecasting algorithms.

[1] AIC and BIC are explained in Section 2.2.3.

For a dynamical system, a set of vectors sampled at discrete times in its state space is described by,

$$\mathbf{x}_{t+1} = F(\mathbf{x}_t) \tag{6.1}$$

Assuming that the discrete time map $F : \mathbb{R}^m \to \mathbb{R}^m$, is continuous with respect to its argument, we can predict a future state $\mathbf{x}_{t+1}$ by searching past states for the one closest to the present state $\mathbf{x}_t$. If a state $\mathbf{x}_{t_0}$ is similar to $\mathbf{x}_t$, then the continuity assumption guarantees that $\mathbf{x}_{t_0+1}$ will be close to $\mathbf{x}_{t+1}$.

To construct a forecasting algorithm, access to the system state appears to be necessary in order to estimate future states, but there is no direct access to the state, just a series of measurements, $\mathbf{y} = (y_1, \ldots y_t)$. However we can approximate the system state by Taken's delay embedding theorem [129], which defines the conditions for reconstructing the state space. A clear introduction to state space reconstruction including its history is given in Casdagli *et al.* [18].

The delay vector is,

$$(s(\mathbf{x}_n), s(F(\mathbf{x}_n)), s(F \circ F(\mathbf{x}_n)), \ldots) \tag{6.2}$$

where s is a measurement function and F is the map representing the dynamics as before. With this representation it can be seen that knowing $s(\mathbf{x}_n)$ at successive times is equivalent to knowing a set of different coordinates at a single time if the mapping F couples the different coordinates. We can rewrite the delay vector in terms of the observed variable y, as $\mathbf{s}_n = (y_{n-(m-1)\tau}, \ldots y_{n-\tau}, y_n), m \geq 1$. Here m is called the embedding dimension and τ is the embedding delay.

The m-dimensional reconstruction of the system dynamics facilitates a method of analogues forecasting method. A simple predictor takes the average of the 'successor' points to the neighbouring delay vectors formed from the time series,

$$\hat{y}_{t+1} = \frac{1}{|\mathcal{U}_\epsilon(\mathbf{s_t})|} \sum_{n:\mathbf{s}_n \in \mathcal{U}_\epsilon(\mathbf{s_t})} y_{n+1} \tag{6.3}$$

where y_{n+1} is the successor point to the vector $\mathbf{s}_n$, $\mathcal{U}_\epsilon$ is a neighbourhood of radius ϵ and $|\mathcal{U}_\epsilon(\mathbf{s_t})| \geq k$ where k is an adjustable parameter. Figure 6.1 illustrates a reconstructed state space with neighbouring vectors and the neighbourhood boundary shown as circle.

The algorithm for simple nonlinear prediction is shown as Algorithm 1. For stable predictions to be made, at least k forecast vectors are needed, so the neighbourhood radius ϵ is progressively increased by a factor r until enough vectors are found.

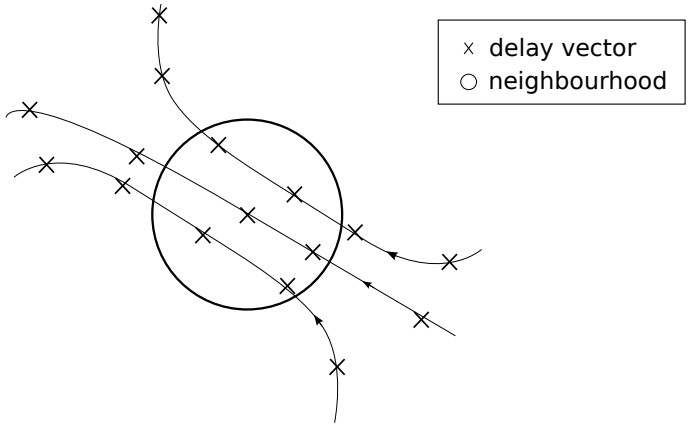


Figure 6.1: Representation of a reconstructed state space showing delay vectors and the neighbourhood of the current state.

Algorithm 1 Simple nonlinear prediction

1: Record the sequence $\mathbf{s}_t = (y_{t-(m-1)\tau}, \ldots y_{t-\tau}, y_t)$
2: Using a distance measure $\mathcal{L}$ and a neighbourhood radius ϵ, find at least k sequences $\mathbf{s}_n$ close to $\mathbf{s}_t$
3: If at least k sequences are not found, expand the neighbourhood ϵ by a factor r and repeat from 2:
4: Using the candidate sequences $\mathbf{s}_n$, the next step forecast at time $t+1$ is the average of their successor (image) points y_{n+1}.

6.1.1.1 Local linear forecasting

The simple nonlinear prediction method, described in the last section, approximates system dynamics by averaging over the successor points, and it relies on the assumption that the mapping F is continuous. If we also assume that the first derivative exists, a linear model can be fit locally, approximating the dynamics using a local Taylor expansion. Local linear regression is described in a statistical learning context in Hastie *et al.* [59, p194], and a summary is given in Section 2.2.1.2. The parameters of the local models are found by minimising,

$$\sigma^2 = \sum_{n:\mathbf{s}_n \in \mathcal{U}_\epsilon(\mathbf{s_t})} (y_{n+1} - \mathbf{a}_t \cdot \mathbf{s}_n - b_t)^2 \tag{6.4}$$

to solve for $\mathbf{a}_t$ and b_t. As before y_{n+1} is the successor point to the vector $\mathbf{s}_n$. The next step prediction is then given by,

$$\hat{y}_{t+1} = \mathbf{a}_t \cdot \mathbf{s}_t + b_t \tag{6.5}$$

The algorithm for local linear prediction is shown in Algorithm 2.

Algorithm 2 Local linear prediction

1: Record the sequence $\mathbf{s}_t = (y_{t-(m-1)\tau}, \ldots y_{t-\tau}, y_t)$
2: Using a distance measure $\mathcal{L}$ and a neighbourhood radius ϵ, find at least k sequences $\mathbf{s}_n$ close to $\mathbf{s}_t$
3: If at least k sequences are not found, expand the neighbourhood ϵ by a factor r and repeat from 2:
4: Using the candidate sequences $\mathbf{s}_n$, find the local linear parameters $\mathbf{a}_t$ and b_t by minimising (6.4)
5: The next step forecast at time $t+1$ is given by (6.5)

Local linear forecasting constructs a tangent plane to the surface $F(\mathbf{s})$ at a point $\mathbf{s}_t$. The neighbourhood has to be large enough to encompass enough vectors for learning the parameters $\mathbf{a}_t$ and b_t without a large statistical error. Further, the neighbourhood diameter must be larger than the noise amplitude to avoid fitting the model to the noise, which is an additive vector term on the underlying local linear model. Since the algorithm functions by expanding the neighbourhood to gain enough vectors, there is a risk of it becoming too large to approximate the curvature of the surface [59, p242]. The effect of this is to incorporate vectors whose state evolution may differ widely resulting in reduced forecast accuracy.

6.1.2 Non-parametric regression

The forecasting methods described so far in this chapter have been based on a non-stochastic model, such as a local linear fit to a reconstructed state space. A similar method can be derived using a non-parametric regression approach. Following the argument in Hastie *et al.* [59, p18], we define $X \in \mathbb{R}^p$ as an input vector and $Y \in \mathbb{R}$ as a real-valued output. We assume the model,

$$Y = f(X) + \zeta \tag{6.6}$$

where ζ is a noise term. We want to find the regression function $f(X) : \mathbb{R}^p \to \mathbb{R}$ for predicting Y given X. Assuming a squared error loss function $L(Y, f(X)) = (Y - f(X))^2$, we solve by minimising the expected prediction error[2] with respect to f,

$$EPE(f) = \mathcal{E}[(Y - f(X))^2] \tag{6.7}$$

[2]Expected prediction error is defined in Section 2.2.3.1.

which has the solution,

$$f(x) = \mathcal{E}[(Y|X = x)] \tag{6.8}$$

The nearest neighbour estimation of the regression function is,

$$\hat{f}(x) = \frac{1}{k} \sum_{i:x_i \in N_k(x)} (y_i|x_i) \tag{6.9}$$

where $N_k(x)$ is the neighbourhood containing the k points in the training set which are closest to x. Averaging approximates the expectation $\mathcal{E}(f)$, and the neighbourhood is used to select the input data x.

The foregoing description is given in a general regression context. In the context of time series, the regression function (6.8) becomes,

$$f(y) = \mathcal{E}[Y_{t+1} \,|\, y = Y_{t-m+1}, ..., Y_t], \qquad (m \leq t) \tag{6.10}$$

Here, Y_t, $t = 1, 2 ...$ is the tth random variable from the stationary and ergodic process $\{Y\}$. The assumption of stationarity ensures that $f(y)$ is not a function of time. The assumption of ergodicity ensures that the sample average converges to the expectation, a requirement for the nearest neighbour estimator.

If we assume that $m < t$, then $\{Y\}$ is generated by a Markov chain[3] whose states are formed by delay vectors $\mathbf{s}_n = y_{n-m+1}, ..., y_n$, which have dimension m and unit time $\tau = 1$. We form a distribution of state transitions by defining the neighbourhood $\mathcal{U}_\epsilon(\mathbf{s_n})$ to give a sample of the transitions $p(s_{n+1}|\mathbf{s}_n)$. The best predictor for the Markov chain, assuming squared error loss, is approximated by the mean of this conditional distribution [73, p261],

$$\hat{y}_{t+1} = \frac{1}{|\mathcal{U}_\epsilon(\mathbf{s_t})|} \sum_{n:\mathbf{s}_n \in \mathcal{U}_\epsilon(\mathbf{s_t})} y_{n+1} \tag{6.11}$$

which is also the simple nonlinear predictor in the deterministic case, given in (6.3). A discussion of the relation between the stochastic and deterministic cases is given in [106].

6.1.3 Kernel regression

The simple nonlinear forecast of (6.11) gives each instance vector equal weight in computing the predictor. Kernel regression gives greater weight to instance vectors which are closer to the current state. In effect it extends the neighbourhood

[3] A Markov chain is a univariate Markov model in continuous space and discrete time.

radius to infinity and introduces a kernel W such that the prediction instance vectors $\mathbf{s}_n$ are weighted according to their closeness $d = ||\mathbf{s}_t - \mathbf{s}_n||$ to the system state $\mathbf{s}_t$. In this case the simple nonlinear forecast (6.11) becomes,

$$\hat{y}_{t+1} = \frac{\sum_{n \neq t} W(\mathbf{s}_t, \mathbf{s}_n)\, y_{n+1}}{\sum_{n \neq t} W(\mathbf{s}_t, \mathbf{s}_n)} \tag{6.12}$$

where as before $\mathbf{s}_n = (y_{n-(m-1)\tau}, \ldots y_{n-\tau}, y_n), m \geq 1$ and the summations are taken over all the possible delay vectors, that is $n = N, N - m\tau, N - 2m\tau$.. where N is the length of the time series.

For a kernel of the form $W(d) = \Theta(\epsilon - d)$, where Θ is the Heaviside step function, (6.12) simplifies to the simple nonlinear prediction of (6.11). A popular form of kernel is the squared exponential $W(d) = \exp\left(-\gamma d^2\right)$ in which the parameter γ determines the kernel width. Yu *et al.* [143] add a second kernel whose argument is the time difference between states, similar to the covariance functions used in defining a Gaussian process. Their *double kernel* is proposed in the context of quantile regression, while Arora *et al.* [5] apply it to time series. A discussion of kernel smoothing in a general statistical context is given in Section 2.2.1.2.

6.2 Current approaches

This section describes some current implementations of k-nearest neighbour forecasting for time series, and introduces a new approach called PPMD. The name *PPMD* is derived from *P*rediction by *P*artial *M*atch version *D* after a data compression technique that motivated the development of the approach [24]. In implementing k-nearest neighbour methods, there are choices to be made in selecting the embedding delay and neighbourhood radius, and in choosing the instance vectors. First we describe how parameters have commonly been selected and some of the problems that this can lead to. We then discuss how these problems have been addressed by recent work and finally introduce the PPMD forecasting method.

6.2.1 Parameter selection

The simple nonlinear prediction method based on (6.3) depends on the embedding delay τ, the embedding dimension m and the neighbourhood $\mathcal{U}_\epsilon$. In the deterministic case m must be greater than twice the dimension of the system to

guarantee that there will be no self-intersection in the reconstructed space, although it can be as small as the dimension [18, p59]. The embedding delay τ can be determined from the attractor geometry if that is known: the attractor should be unfolded so that its extension in all dimensions is roughly the same. A simpler heuristic approach advocated by Kantz *et al.* [73, p150] is to set τ to the autocorrelation length[4] of the time series. A discussion of optimal embedding parameter selection, with a criterion for selection is given in [124].

In the stochastic case the embedding delay is set to $\tau = 1$ in order to find the states whose transition probabilities we are to estimate. The best predictor for a Markov process of order p is formed by using delay vectors of length m where $m > p$. Kantz *et al.* [73, p263] suggest using (6.11) to compute the forecast error as a function of m and choosing the value that minimises this error. One example of this approach applied to the deterministic Lorenz system is given in Meng *et al.* [90].

6.2.2 Instance vector selection

The choice of instance vectors is determined by the neighbourhood radius ϵ and cardinality $|\mathcal{U}_\epsilon(\mathbf{s_t})| \geq k$, which is the number of instance vectors used for prediction. The value of k has to be high enough for the sample $p(s_{n+1}|\mathbf{s}_n)$ to be representative of the true distribution. The simple nonlinear and local linear prediction methods expand the neighbourhood radius until at least k vectors are found. For low values of the embedding dimension (short prediction vectors) and a large training set, it should be possible to find k or more vectors. However, for a high embedding dimension or a short training set, the neighbourhood may expand to include spurious prediction vectors. This is a manifestation of the curse of dimensionality in which the training data becomes sparse as the dimension increases. The problem may also arise for individual states when the density of state vectors close to the current state is low.

Arora *et al.* [5] address this problem by adjusting the neighbourhood radius according to the density of states. They propose the *fraction-Nearest Neighbor (f-NN)* method in which only the top fraction f of states are used for prediction, effectively varying the neighbourhood radius with state density. In this case k = $f \times N_t$, where $N_t = t - 2m + 1$ denotes the total number of previous state vectors at time t. The optimum value for f and the embedding dimension m are

[4]The autocorrelation length is the time taken for the autocorrelation to decay to e^{-1}.

estimated by minimising the in-sample forecast error. An evaluation on gross national product (GNP) time series shows that *f-NN* outperforms other nonlinear models such as Markov switching-AR (MS-AR) models and SETAR models.

The kernel methods described in Section 6.1.3 also mitigate the effect of instance vectors which do not represent the current trajectory by giving more weight to vectors that are close to the current state. Arora *et al.* [5] investigate a kernel method applied to their *f-NN* approach, and note slightly better results compared with *f-NN*, though with the penalty of an additional parameter. They find that a double kernel approach, in which the weighting is also a function of the time between states, does not significantly improve forecast accuracy.

6.2.3 PPMD Forecasting

Here we propose a modified form of the simple nonlinear forecasting model which addresses two issues related to spurious instance vectors. The first change reduces the effect of errors from spurious vectors by using the median rather than the mean to average successor points. In (6.11) it is assumed that the prediction vectors $\mathbf{s_n}$ are close to the current state $\mathbf{s_t}$. However when spurious vectors are included in the neighbourhood, this assumption is violated. Under these circumstances the mean of the successor points y_{n+1} may not be the best predictor of the next state. We mitigate the effect of a minority of spurious vectors by using the median, so that (6.11) becomes,

$$\hat{y}_{t+1} = \underset{n:\mathbf{s}_n \in \mathcal{U}_\epsilon(\mathbf{s_t})}{\text{median}} \{y_{n+1}\} \tag{6.13}$$

The second change attempts to reduce the number of spurious vectors that are selected by modifying the forecasting algorithm. For time series where the system dynamics are fixed and known, the values for the embedding delay τ and embedding dimension m can be estimated. In this case, fixing m and using the minimum neighbourhood radius to select prediction vectors from the training set provides the closest approximation to the present state. However, when the optimal embedding dimension is unknown, setting it to an arbitrary value and then expanding the neighbourhood is likely to select spurious vectors for prediction. A further problem arises in predicting time series with mixed dynamics in which case a single value of embedding dimension may not be appropriate. The PPMD algorithm addresses these issues by using a fixed neighbourhood size and progressively decrementing the embedding dimension to search for instance vectors.

6.2.3.1 Algorithm

The PPMD algorithm searches the training set for sequences which resemble the sequence of length m preceding the point y_{t+1} as shown in Figure 6.2. If no sequences are found, rather than increasing the neighbourhood size, it reduces m and repeats until a sequence is found. So when no instances of the sequence $y_{t-m+1}, \ldots y_t$ are found it uses the next shortest sequence $y_{t-m}, \ldots y_t$.

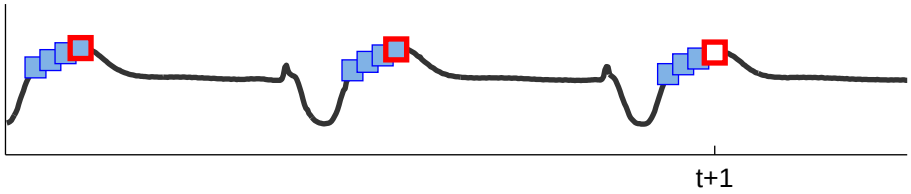


Figure 6.2: K-nearest neighbour forecasting for time series. The training set is searched for sequences that are similar to the candidate sequence comprising the three filled squares preceding $t+1$. The successor points (shown as blue with red borders) following the instance sequences are then used to estimate the next step forecast (white with red border).

Algorithm 3 PPMD

1: Record the sequence $\mathbf{s}_t = (y_{t-m+1}, \ldots y_{t-1}, y_t)$
2: Using a distance measure $\mathcal{L}$ and a neighbourhood radius ϵ, find at least k sequences $\mathbf{s}_n$ close to $\mathbf{s}_t$
3: If at least k sequences are not found, set $m = m - 1$ and repeat from 2:
4: Using the candidate sequences $\mathbf{s}_n$, the next step forecast at time $t+1$ is the median of the successor (image) points y_{n+1}.

The algorithm is given in Algorithm 3. It differs from simple nonlinear forecasting, Algorithm 1, in the search method in step 3: and the use of the median in step 4: to derive the forecast from the distribution of successor points.

6.2.3.2 Kernel variants

The use of the median of the distribution of successor points y_{t+1} is intended to reduce the contribution of spurious instance vectors to the forecast estimate. Another approach is the use of kernels, described in Section 6.1.3. A variant of the standard PPMD algorithm uses a kernel predictor in place of the median. The algorithm is modified so that in step 4: the kernel predictor defined in (6.12) replaces the median. Two variants are tried,

- For the first variant, a squared exponential $W(d) = \exp(-(2d/l_d)^2)$ is used as a kernel, where d is the distance of the instance vector $\mathbf{s}_n$ from the current state $\mathbf{s}_t$ and l_d is the length scale. This kernel variant is denoted PPK.
- A second variant uses a time dependent kernel $W(\delta t) = \exp(-(2\delta t/l_t)^2)$, where $\delta t = t - n$, the time displacement of the instance vector $\mathbf{s}_n$ from the current state $\mathbf{s}_t$ and l_t is the length scale. This kernel variant is denoted PPT.

In practice the length scale of the distance kernel is set to a proportion of the neighbourhood size: a schematic representing the effect of kernel weighting on different instance vectors is shown in Figure 6.3. For the time kernel, the length scale is set as multiples of the delay between the current state and the most recent instance vector.

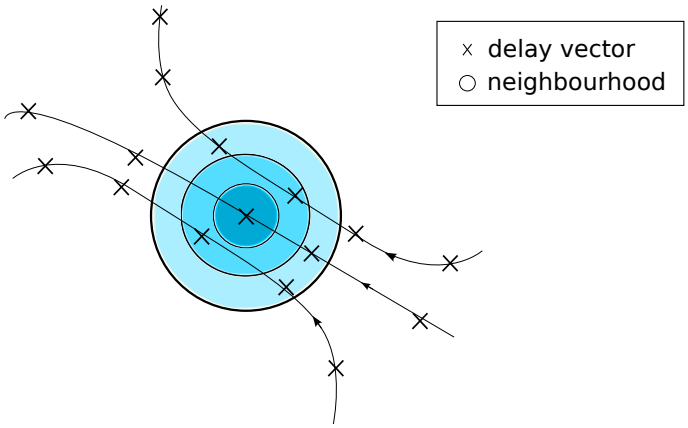


Figure 6.3: Representation of a reconstructed state space showing delay vectors at different distances from the current state. The colours represent weightings given to the vectors within the neighbourhood.

6.3 Evaluation

We compare the performance of the PPMD algorithm with the simple nonlinear and local linear forecasting methods implemented in the *TISEAN* software [61]. Several different time series are used for evaluation. The first is derived from the Lorenz equations [85] which define a well-known deterministic system with chaotic dynamics. Secondly we use the electrocardiogram (ECG) time series `qtdb/sel102` data from Physiobank [53]. Finally we apply the PPMD forecasting method and its kernel variants to the depression time series used in Chapter 5.

6.3.1 Lorenz time series

The Lorenz time series is derived from the equations,

$$\dot{x} = \sigma(y - x) \tag{6.14}$$
$$\dot{y} = x(r - z) - y \tag{6.15}$$
$$\dot{z} = xy - bx \tag{6.16}$$

We use the common setting for the parameters as $\sigma = 10, r = 28$, and $b = 8/3$. The equations are solved using MATLAB `ode45` over the time interval $t = [0, 1000]$ with a time step of 0.01 and an initial condition of $x = 1, y = 0, z = 0$. Figure 6.4 shows the solution.

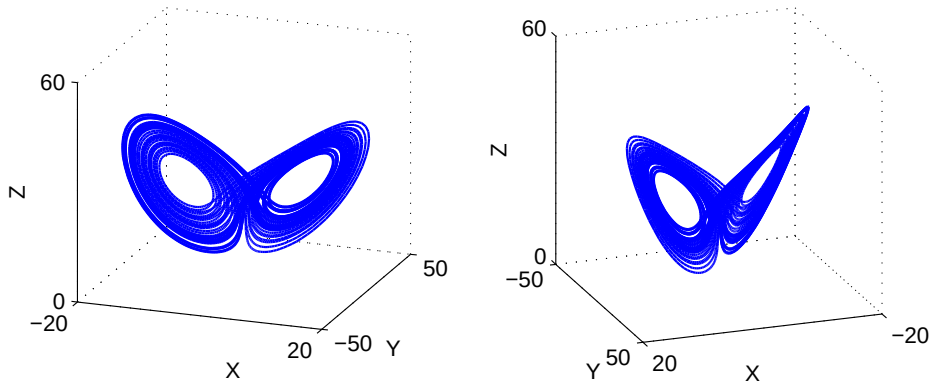


Figure 6.4: Solution for Lorenz equations [85] (6.14) to (6.16) with parameters $\sigma = 10, r = 28$, and $b = 8/3$. The first 1000 points, which contain the transient, are discarded.

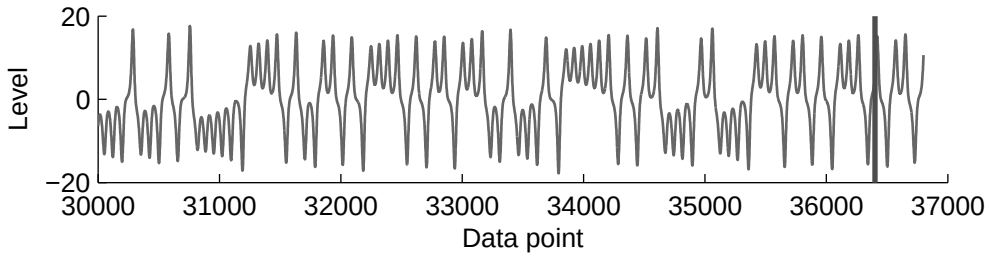


Figure 6.5: A segment from the Lorenz time series which is used for PPMD evaluation. The end of the training set at 36400 is marked by a vertical line. The subsequent 400 points are used as a validation set.

We use the x values in forecasting. The training set is a series of 36400 points, including the candidate series, and this set is used for both the PPMD and the

		Embedding dimension					
	k	*3*	*4*	*5*	*6*	*7*	*8*
	10	0.14	0.10	0.06	0.04	0.33	0.74
	20	0.10	0.10	0.05	0.01	0.01	0.03
Local linear	30	0.12	0.06	0.03	0.05	0.06	0.06
$\tau = 3$	40	0.12	0.13	0.05	0.11	0.10	0.08
	50	0.13	0.11	0.07	0.15	0.12	0.10
		Embedding dimension					
	k	*7*	*8*	*9*	*10*	*11*	*12*
	5	0.33	0.39	0.28	0.30	0.33	0.29
	8	0.30	0.28	0.26	0.28	0.30	0.28
Simple nonlinear	10	0.29	0.22	0.21	0.24	0.37	0.37
$\tau = 20$	15	1.77	0.24	0.20	0.21	0.35	0.33
	20	0.29	1.80	0.21	1.82	0.36	0.47
		Maximum Embedding dimension					
	k		*200*	*250*	*300*	*350*	
	5		0.17	0.18	0.18	0.23	
PPMD	10		0.22	0.12	0.22	0.12	
$\epsilon = 0.14$	15		0.50	0.88	0.50	0.88	

Table 6.1: RMS validation error for three different forecast methods applied to the Lorenz time series data. The matrix shows the validation error for different values of embedding dimension m and the number of instance vectors k. In each case the minimum error is highlighted.

TISEAN forecasting methods. A time plot showing the final segment from the training set and the validation set is shown in Figure 6.5.

We predict points in the validation set using the three forecasting methods: simple nonlinear forecasting based on (6.3), local linear forecasting based on (6.5) and PPMD based on (6.13). We estimate the parameters for the forecasting methods by minimising the RMS prediction error on the validation set. For the *TISEAN* methods a grid search is made over the embedding delay τ, the embedding dimension m and the minimum number of instance vectors k. For PPMD, the grid search is made over the maximum embedding dimension m, the minimum number of instance vectors k and the neighbourhood radius ϵ. The RMSE validation error for the different parameters is shown in Table 6.1 with the minimum error highlighted for each forecasting method.

A time plot of the validation set prediction using optimum parameters is

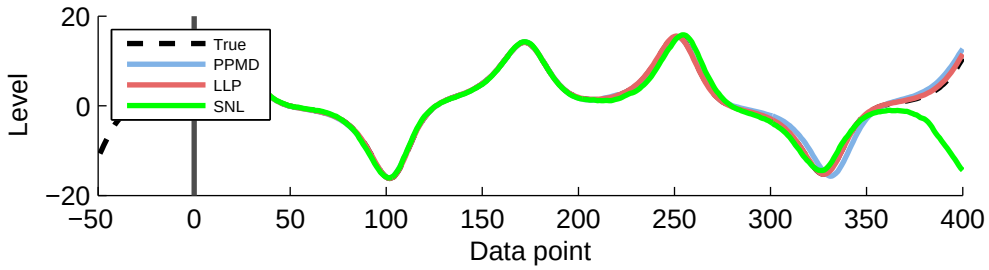


Figure 6.6: PPMD forecast method applied to the Lorenz time series. The plot shows PPMD forecast performance compared with local linear (LLP) and simple nonlinear (SNL) forecasts. The validation set and an additional segment is shown.

shown in Figure 6.6. The evaluation shows that local linear forecasting is most accurate for the chosen validation set, probably because the local linear model provides a better approximation to dynamics than averaging successor points of the instance vectors. The chosen parameters for PPMD ($k = 10, \tau_{max} = 250$) are similar to those found for simple nonlinear forecasting ($k = 15, m\tau = 180$). However, PPMD has about half the validation error of the *TISEAN* implementation because of differences between Algorithms 1 and 3.

6.3.2 ECG DATA

Having compared PPMD and *TISEAN* algorithms on the synthetic Lorenz time series, we apply them to the `qtdb/sel102` ECG data from Physiobank [53]. A time plot of the final segment from the training set is shown in Figure 6.7.

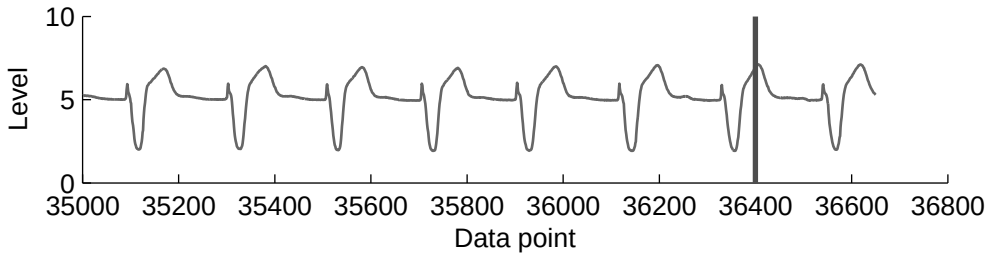


Figure 6.7: A segment from the `qtdb/sel102` ECG time series used for PPMD evaluation. The end of the training set at 36400 is marked by a vertical line. The subsequent 250 points are used as a validation set.

As before we optimise both the *TISEAN* and PPMD forecasting methods using a grid search over the parameter space. For the *TISEAN* methods the autocorrelation length is used to guide the choice of τ and we examine validation errors for $\tau = 1, 3, 10, 20$ and chose the value for which the errors are lowest. The root

		Embedding dimension			
	k	*3*	*6*	*9*	*12*
	5	2.55	–	–	–
Local linear	10	0.56	0.13	–	–
$\tau = 20$	20	0.75	0.31	0.39	0.34
	30	1.18	0.38	0.23	0.35
		Embedding dimension			
	k	*3*	*6*	*9*	*12*
	5	0.57	0.50	0.37	0.25
Simple nonlinear	10	0.54	0.34	0.11	0.34
$\tau = 20$	20	0.69	0.34	0.11	0.14
	30	0.87	0.27	0.27	0.19
		Maximum Embedding dimension			
	k	*150*	*200*	*250*	*300*
	5	0.47	0.19	0.35	0.58
PPMD	10	0.19	0.14	0.12	0.62
$\epsilon = 0.035$	15	0.34	0.26	0.64	0.55

Table 6.2: RMS validation error for three different forecast methods applied to an ECG time series. The matrix shows the validation error for different values of embedding dimension m and the number of instance vectors k. In each case the minimum error is highlighted. Some parameter values for local linear forecasting did not allow forecasting of the entire validation set and so are marked –.

mean square validation error for different values of embedding dimension m and number of instance vectors k is shown in Table 6.2. The findings can be compared with those in Kantz *et al.* [73, p271] who report parameter values of $m = 8, \tau = 5$ and $k = 10$ for predicting 250ms of a different ECG time series using simple non-linear prediction. The corresponding values from Table 6.2 are $m = 9, \tau = 20$ and $k = 10$.

Time plots of the validation set prediction using the optimum parameters for the three methods are shown in Figure 6.8. This shows that the local linear method reproduces the peak close to 0.6 seconds whereas the other two methods do not. The better resolution is likely to be due to a smaller neighbourhood, which for the *TISEAN* methods is the result of using a lower embedding dimension. A more detailed study of ECG prediction, along with a discussion of deterministic and stochastic aspects of the ECG is given in [72].

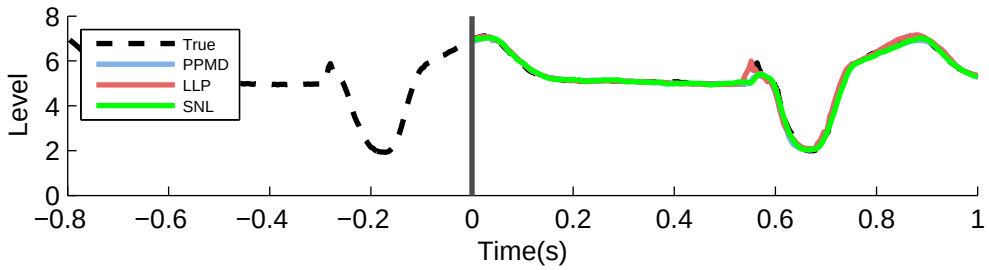


Figure 6.8: PPMD forecast method applied to the ECG validation set. The plot shows PPMD forecast performance compared with local linear (LLP) and simple nonlinear (SNL) forecasts.

6.3.2.1 Out-of-sample results

Using the parameter values found from the validation set, we make predictions on separate test sets, starting at the point 40,000 and making forecasts of lengths 200, 500 and 1000. The test is repeated to make 5 different trials on different segments of the time series for each of the three test lengths. For a test length of 200 the test set starts at points 40,000, 40,200, 40,400, 40,600 and 40,800. For a test set length of 500, the test set starts at points 40,000, 40,500 and so on.

		Trial						
Test set length		*1*	*2*	*3*	*4*	*5*	*Mean*	*Median*
200	*SNL*	0.68	0.56	0.14	0.13	1.02	0.51	0.56
	PPMD	0.76	0.35	0.55	0.34	0.89	0.58	0.55
500	*SNL*	1.06	0.35	0.55	1.15	0.34	0.69	0.55
	PPMD	0.98	0.50	0.25	0.36	0.35	0.49	0.36
1000	*SNL*	1.32	0.71	0.80	1.37	0.93	1.02	0.93
	PPMD	1.19	0.67	0.80	1.24	0.76	0.93	0.80

Table 6.3: Out-of-sample errors for ECG data. The table shows the root mean square error for simple nonlinear and PPMD forecasts on ECG data. Trial predictions are made starting at five different points in the time series, which for a test set length of 200 are 40,000, 40,200 etc. Local linear forecasting failed for several of the trials so the results are not included in the table.

The results are shown in Table 6.3. For the five test sets of length 200, PPMD is comparable in accuracy to simple nonlinear forecasts. However, for test lengths of 500 and 1000, PPMD is slightly more accurate. The better performance of simple nonlinear forecasting for short time horizons is likely to be the result of using a smaller neighbourhood and so obtaining better resolution of time series features.

6.3.2.2 Kernel variants

We repeat the validation set forecasts using the kernel variants described in section 6.2.3.2. We use the same forecast parameter values as for the PPMD forecast, and examine the validation error for a range of kernel widths. The validation errors are shown in table 6.4. Neither of the kernel methods improves on the basic PPMD algorithm, which uses the median. However in this case the forecasting parameter values are fixed at those found for PPMD forecasting, rather than for the kernel variants. The full parameter space is not searched because the compute time is too long. A full search is performed for the mood time series described in section 6.3.3.

PPK	×	1	2	3	4	5	6	7	8	9	10
	RMSE	0.32	0.31	0.34	0.35	0.31	0.37	0.38	0.38	0.38	0.38
PPT	×	2	4	6	8	10	12	14	16	18	20
	RMSE	0.37	0.17	0.16	0.16	0.39	0.32	0.40	0.18	0.18	0.22

Table 6.4: Validation error for the distance kernel method PPK and the time kernel variant PPT on ECG data. The length scale for the distance kernel method PPK is the neighbourhood radius 0.035 multiplied by the multiplier shown in the table. The length scale for the time kernel method PPT is the time between the current state and the most recent instance vector multiplied by the multiplier shown.

6.3.3 Mood data

We next evaluate the PPMD algorithm on a depression time series using a similar procedure as applied to the Lorenz and ECG time series. The time plot of the depression data is shown in Figure 6.9.

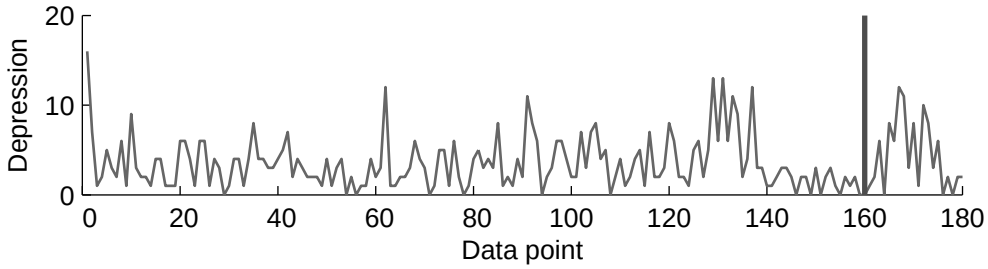


Figure 6.9: A segment from the depression time series used for PPMD evaluation. A segment showing point 1 to the end of the training set at point 160 is shown. The subsequent 20 points are used as a validation set.

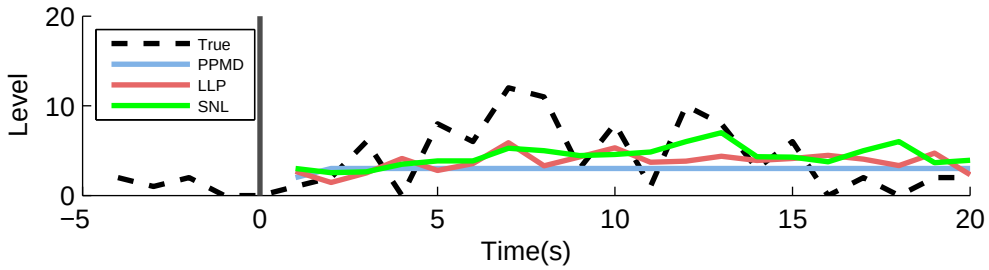


Figure 6.10: PPMD forecast method applied to a depression time series. The plot shows PPMD forecast performance compared with local linear (LLP) and simple nonlinear (SNL) forecasts on the validation set.

We optimise the forecast method parameters on a validation set of 20 points and show the in-sample forecast results in Figure 6.10. None of the methods performs especially well. This result is unsurprising because the results from Chapter 5 show that next-step forecasting is difficult for this data, so a horizon of 20 is unlikely to be successful using any forecast method. Further, nearest neighbour forecasting works best with a long training set and low noise, while the depression data is short and noisy.

6.3.3.1 Next-step forecasting

To complete the evaluation, we compare the next step performance of PPMD with that of other methods on the 6 time series of length 150 points that were used in Chapter 5. The forecast parameters are trained using the first 80 points as a training set and minimising the RMS error on the subsequent 20 points. The out-of-sample errors for the next 50 points are shown in Table 6.5 with persistence (*PST*), simple exponential smoothing (*SES*) and the simple nonlinear method (*SNL*). We include results for the PPMD variants with a distance kernel (*PPK*) and a time kernel (*PPT*).

The results show that the k-nearest neighbour methods do not perform well compared with those based on short term autocorrelation such as exponential smoothing or autoregression. From the perspective of accurate forecasting, k-nearest neighbour methods are not well suited to the time series. However their application has revealed that patterns of mood change tend not to recur in the time series.

ID	*PST*	*SES*	*SNL*	*PPD*	*PPK*	*PPT*
1	3.89	3.34	3.32	3.47	3.50	3.56
2	5.82	4.93	4.90	5.42	5.76	5.79
3	3.79	3.39	3.42	3.37	3.39	3.64
4	2.55	1.80	1.86	1.82	1.80	1.79
5	3.50	3.59	3.70	4.13	3.77	5.00
6	1.41	1.41	1.52	1.67	1.53	2.43
Mean	3.49	3.08	3.12	3.31	3.29	3.70
Median	3.65	3.36	3.37	3.42	3.45	3.60

Table 6.5: Next step forecast errors for PPMD applied to six stationary, uniform time series with a length of 150 points. PPMD (tagged *PPD*) uses the median of the successor points as the forecast. *PPK* is a kernel variant of PPMD which uses a weighted mean based on the distance of the instance vector from the current state. *PPT* is a variant of PPMD which uses a weighted mean based on the time difference between the instance vector and the current state.

6.4 Conclusions

This chapter has introduced a nearest neighbour forecasting method (PPMD) which searches for instance vectors using a fixed neighbourhood radius and a varying instance vector length. Instead of using the mean of the image points as a predictor, it uses the median. We have compared its forecasting performance with simple nonlinear and local linear forecasting using the Lorenz and ECG time series. In the case of the Lorenz time series PPMD performed well compared with simple nonlinear forecasting, but both methods were outperformed by the local linear method. In the case of the ECG time series PPMD performed well in comparison with both the other methods. PPMD captures dynamics by its ability to adjust the vector length depending on the position in phase space. This may help in modelling the Lorenz attractor whose wings may be modelled in 2 dimensions while the crossover needs 3. Similarly, the ECG's phase portrait can be only partly represented in a low-dimensional embedding space.

None of the k-nearest neighbour methods performed well when used for next-step forecasting of the mood data used in Chapter 5. We note that the time series are short and that other nonlinear forecasting methods do not perform well in this case. However it does appear from these results that patterns of mood change tend not to recur in the time series. We next conclude the thesis with general conclusions and a proposal for future work.

7

General conclusions

Introduction

This chapter concludes the thesis. First conclusions from the time series analysis, forecasting and dynamics are summarised. Next future work on mood data, techniques and narrative modelling are suggested. Finally, the thesis is summed up with a position statement.

7.1 Conclusions

7.1.1 Time series properties

The mood time series from patients with bipolar disorder are a new resource and we have made an analysis of their properties. We constructed two measures for quantifying non-uniformity: compliance, which measures missing data and continuity, which measures irregularity in the data. We applied them to the time series and found that non-uniformity of response does not correlate with either gender or diagnostic subtype but it does correlate with sleep variability. This finding has implications for selecting time series according to their uniformity because selecting only the uniform series would exclude patients with more variable sleep ratings.

As part of the analysis, data from 93 patients who had returned at least 25 questionnaire ratings was selected as a subset denoted Set A. This set comprised 60 female and 33 male patients, and they returned a median of 95 ratings. The

set was further divided into Set G which had equal genders, and Set D, which had equal numbers of bipolar I and bipolar II patients. Male and female patients could not be distinguished by their uniformity of response, nor could bipolar I or bipolar II patients be distinguished in this way.

Most of the depression time series have a short term autocorrelation structure rather than a seasonal form. However, using the Edelson-Krolik correlogram we found one patient who exhibited mood with yearly seasonality. When pairwise correlations between symptoms were examined we found that a patient's time series for *sleep* and *appetite/weight* show a lower average correlation. In addition, the autocorrelation structure for these two symptoms is distinct from those of the other symptoms. Finally, we examined correlations between patients' depression time series but found no evidence of correlation in general.

7.1.2 Mood forecasting

Psychiatrists have asked whether mood can be forecast using the self-rated data from telemonitoring. The issue is of clinical importance because effective mood prediction could help in managing of the disorder. We concluded that mood cannot be forecast using weekly self-rated data. We tried forecasting the set of male and female patients separately, and found that male and female patients could not be distinguished by their forecast error, nor could bipolar I or bipolar II patients be distinguished in this way.

With in-sample forecasting only 41 from a set of 93 patients show a 10% reduction in error when using persistence rather than an unconditional mean predictor. We tried a battery of forecast methods for out-of-sample forecasting but only 20 patients showed more than a 10% reduction in error relative to a baseline forecasts. Half of the set of 93 patients have fewer than 95 data points, and such a short length has repercussions for training the more complex forecasting methods. Most of the time series are noisy or lack serial correlation, and gains from a minority of time series are masked by large errors in other patients.

The fact that there is a measurable improvement over the persistence baseline shows that there is predictability in at least some of the time series. However the question addressed in Chapter 3 is a practical one: whether forecasting is effective for the set of patients as a whole, and this is shown not to be the case. The out-of-sample forecasting results in Chapter 3 were obtained using a minimum time series length of 64 ratings. The further question as to whether the length of the time series was limiting forecast accuracy was addressed by using a set of

patients with minimum time series length of 150 ratings. The poor forecast results for these patients suggest that length is not the main cause of poor accuracy. We also note that with weekly ratings the periods involved, 3 years in the case of 150 ratings, are too long for practical use in a clinical setting.

Forecasting in this study was based on the aggregate scores rather than on the separate survey components. Some early work on forecasting depression using the component scores was done but this was not pursued because the results were little better than those using aggregate scores, and at the cost of greater complexity. So although information is lost in aggregating the scores before forecasting, it appears not to affect accuracy greatly.

7.1.3 Mood dynamics

An earlier paper by Gottschalk *et al.* [55] claimed evidence of chaotic dynamics for depression in bipolar disorder with daily data. A more recent claim by Bonsall *et al.* [11] is that weekly mood variability is nonlinear. We investigated the dynamics of weekly depression data using surrogate data techniques and nonlinear forecasting methods. We found that the depression time series cannot be distinguished from their surrogates when the respective test statistics are compared. These results can mean that either 1) the original series have linear dynamics, 2) the test statistics for distinguishing linear from nonlinear behaviour do not have the power to detect the kind of nonlinearity present, 3) the process is nonlinear but the sampling is inadequate or the time series are too short to represent the dynamics. We also found that nonlinear forecasting methods have no advantage over linear methods in out-of-sample forecasting. This result contrasts with the report in [11] that non-linear models are the best fit for the mood data.

The time series analysed by Gottschalk *et al.* [55] used daily ratings whereas the set used in this study used weekly sampling. As such the results cannot be directly compared, but we can say that whereas Gottschalk *et al.* claimed to find nonlinear behaviour, we did not. The claim by Gottschalk was contested by Krystal *et al.* [79] on the grounds that the power spectra of the time series were not consistent with chaotic dynamics. We also note that the claim rested on the convergence of correlation dimension, which is not a definitive indication of chaos. It is not possible to say what Gottschalk *et al.* were detecting in their analysis because a finite correlation dimension can occur with a stochastic system [104]. So calculating the correlation dimension for the weekly series used in this study would result in similar questions rather than any more definitive answer.

7.2 Future work

7.2.1 Mood data

The most striking limitation of the current data set is that it is sampled weekly. Efforts were made to obtain more frequently sampled data but these were unfruitful. The existing data uses standard rating scales which apply to the previous week, and these scales have been carefully evaluated under that assumption. The spread of new technologies now allows more rapid sampling and there is a need for psychometric evaluation of scales for more rapidly sampled, self-rated data. A study of the range of frequencies in depression in bipolar disorder would also help in choosing an optimal sample rate, consistent with practical considerations. If daily data from a reliable and valid scale were available the following research projects would be worth pursuing,

- Mania episode forecasting is potentially very helpful for managing the disorder. It was tried for this project but abandoned because results were poor compared with a simple predictor based on a threshold value. A range of classification techniques was tried including decision trees and random forests (an ensemble of decision trees) but none was especially accurate. However, the use daily data and the collection of other variables, especially activity data, might make mania prediction more feasible.
- The analysis of dynamics for daily data would provide a useful replication of the Gottschalk *et al.* paper [55].
- Additional forecasting types and methods could be tried. For example categorical forecasting, distributional forecasts and horizons greater than next-step. Other forecasting methods could include multivariate input Gaussian processes using automatic relevance determination, and exponential smoothing methods incorporating trend and seasonality.

7.2.2 Techniques

- The new measures of non-uniformity address two situations which result in low quality time series: too many missing data points (low compliance) or insufficient contiguity among the existing data points (low continuity).

As continuity approaches unity, the error in measuring autocorrelation approaches zero. The theoretical relationship between continuity and this error could be investigated for different processes.

- The PPMD k-nearest neighbour method was developed for this application and a full evaluation on different data sets would be appropriate for a separate project. Some work has already been done on applying the technique to tidal data forecasts. The current technique for forecasting tides uses harmonic analysis and is accurate for long time horizons. Sea levels are also determined by atmospheric and local geographical variables, and these are modelled separately to determine the sea level at a particular location. Since tidal data has relatively low noise, the algorithm can effectively work as a 'function recorder' which selects the longest period of similar behaviour in the past. Initial work on applying PPMD to forecasting tides showed that it could reproduce the baseline harmonic analysis for short time horizons.

7.2.3 System identification

The identification of system dynamics, which might be high dimensional and include unobserved environmental influences would be difficult using just the data in this study. Without observational data, it is futile to try to describe the system dynamics [123, p66], but an approach focused only on data, especially univariate time series, is challenging without a general theory of the disorder.

The difficulty in applying fundamental models to mental illness is that existing models of brain function do not provide the level of description needed to explain symptoms. There exist electrophysical models which describe behaviour at the neural level and there are psychological explanations of behaviour, but these modes of explanation have tended to be disjoint. There is some work that attempts to close this 'explanatory gap', in particular the *Bayesian brain* hypothesis [77] [46] which has been applied to both neuroscientific observations [45] and mental illness [44] [47].

Fletcher and Frith [44] explain hallucinations and delusions in schizophrenia using a hierarchical Bayesian model of brain function. If mood episodes are assumed to be caused by aberrant perceptions, then the Bayesian brain approach is directly applicable to bipolar disorder. However, rather than making that assumption, here we propose a theoretical model which might subsequently be validated against observational data. Other fundamental models were discussed earlier in

Section 1.4.1 and Section 1.4.2. The proposal here extends the use of generative models which are used in the Bayesian brain explanations of psychopathology. In such explanations, generative models are used for learning representations of external causes of observations. We suggest that the generative capacity of these models can be further exploited to explain symptoms of mental illness.

7.2.3.1 Perception

An important function for any animal is to know its environment by using observations from its senses. We define an *external state* X as a cause or condition in the environment that fully determines the *observations* U. We define an *internal state* W as the cognitive representation of an external state based on the observations U. The task for the cognitive apparatus is to find the distribution of internal states conditioned on observations, $p(W|U)$ in order to model the external state.

Higher level inference is accomplished by using a hierarchy in which states inferred by one level are used as observations for the next. Figure 7.1 is a pictorial representation of the hierarchy of inference levels. Observations arrive from sensory neurons as spike trains which are processed into features, such as edges in a visual scene. The features themselves become observations for a second level of inference which in turn uses them to infer states that are meaningful the next highest level. Discussion of hierarchical inference is to be found in [100] [50] and [62] and a history of the ideas is given in [46].

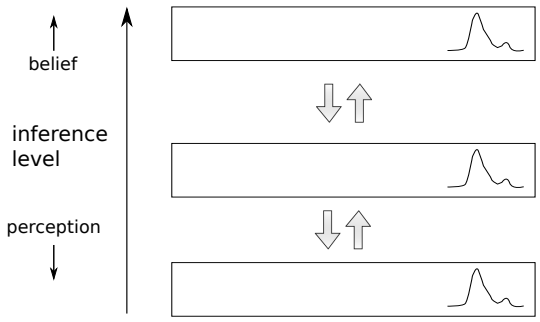


Figure 7.1: A pictorial representation of cognition as multi-level statistical inference. The graphs represent the conditional distribution $p(W|U)$ which has a peak at the most likely internal state $\hat{W}$. Inferences from the lowest level may consist of *features* such as edges in the visual scene which become observations for the next level in the hierarchy. High level cognition comprising beliefs, thoughts and imagination are represented at the top of the hierarchy.

7.2.3.2 Formal narratives

A generative model can generate artificial observations which have the same distribution as those already experienced: this leads to the possibility of simulation. A process that can capture how states change over time is likely to be useful to an organism because it models behaviour and allows future scenarios to be simulated and their likelihood predicted. So we define a *formal narrative* as a chain of states, or a simulation, which has an associated probability. A set of such narratives can be used to represent prospection, retrospection and mental imagery.[1]. Both prospective and retrospective narratives are seen as estimates of an uncertain external state. Under this model, prospection, memory and as mental imagery are seen as similar processes of construction. This view contrasts with models which represent memory as a more deterministic process.

7.2.3.3 Application

Currently there are a number of psychiatric symptoms for which a compelling explanation remains elusive. For example, there is no definitive account of obsessive-compulsive disorder (OCD) [67, p 165]. In bipolar disorder, some symptoms are addressed by mental imagery [63], but the lack of a formal description for mental imagery limits its explanatory power. Recent explanations for delusions and hallucinations in schizophrenia, for example [25] [44], use a formalism at too high a level to give a detailed account of what is going wrong.

The idea of a formal narrative does not in itself provide an account of symptoms either. However, it seems to have potential as a formalism on which explanations could be based. Disorganised speech can be understood as a failure of language generation resulting from incorrect transition probabilities between generated language elements. This failure might be related to known cognitive deficits in schizophrenia. Similarly, mania and flight of ideas in bipolar disorder could be seen in terms of over-generation (or under-pruning) of narratives. The quality of such explanations should be judged by whether they lead to successful new predictions which other theories cannot explain [67, p154]. In general, narratives might be a formalism that can combine evidence from psychology or

[1]Prospection is the mental function of anticipation or looking ahead. Retrospection is looking back in time. Mental imagery is the experience of accessing perceptual information from memory.

neuroscience with psychiatric symptoms. The detailed construction of explanations based on formal narratives is the task of a new research programme.

7.3 Summary

We return to the discussion at the beginning of Chapter 1. In order to make progress in understanding bipolar disorder, observational data should be combined with realistic models of brain function. However, we are experiencing an unparalleled increase in data at the same time that psychiatry has no coherent theoretical basis.

Most of this thesis has dealt with modelling mood using statistical time series techniques, but since prediction is an important aspect of cognition, mood might itself be dependent on neural prediction. It follows that statistical models could help in modelling brain dysfunction. In neuroscience, at least one theory of global brain function, Friston's free energy principle, has already been applied to imaging and electrophysiological data. The next challenge is to close the explanatory gap and provide more compelling explanations for mental illnesses.

Appendix A

Introduction

This appendix provides further statistics for the non-uniform time series in Sets A, G and D. Set A is the main set and comprises 93 patients whose time series have 25 points or more. Set G (gender) has 20 male and 20 female patients. Set D (diagnosis) has 16 patients with bipolar I and 16 with bipolar II disorder. In the statistics to follow the histogram bin width is chosen to give an accurate profile of each distribution. In the case of age, the bin width is 5 years, starting with a bin centred at age 20. For time series length, the bin width is 10 data points starting at a length of 20 data points. For mood, the bin width is 0.5 of a rating unit.

A.1 Statistics for time series and patient data

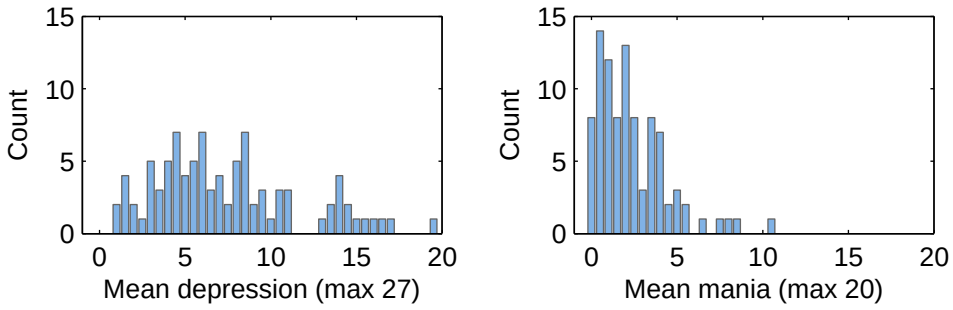
Figure A.1 shows the distribution of mean mood ratings for Set A whose length statistics are shown in Figure 3.3. No patients have a mean depression less than 1 whereas 29 patients return mean mania scores of less than 1, meaning that the patient is hardly more cheerful, self-confident etc. than usual. The standard deviation for mood ratings is shown in Figure A.2, which shows that mania scores are typically less dispersed than those for depression. Mean values for individual depression symptoms are shown by the box plot in Figure A.3. Many symptoms have a median for the set of patients between 0.5 and 1, where the maximum value for a single symptom domain is 3. Exceptions are *Thoughts of death or suicide* whose median lies closer to 0 and *Sleep*, whose median lies closer to 2.

A.2 Statistics split by gender

Figure A.4 describes the age and time series length for Set G, split by gender. The distributions for males and females are not markedly different. The mean and standard deviation of mood, split by gender, are shown in Figures A.5 and Figure A.6. For mean depression scores, the median for the set of males is 4.9 and for females it is 7.1. The scores for individual symptoms of depression are shown in Figures A.7 for women and A.8 for men. These figures show that women have a higher median rating for individual symptoms than men. Both sexes follow the same pattern as that for Set A with the symptom *Thoughts of death or suicide* having generally low scores and *Sleep*, high scores.

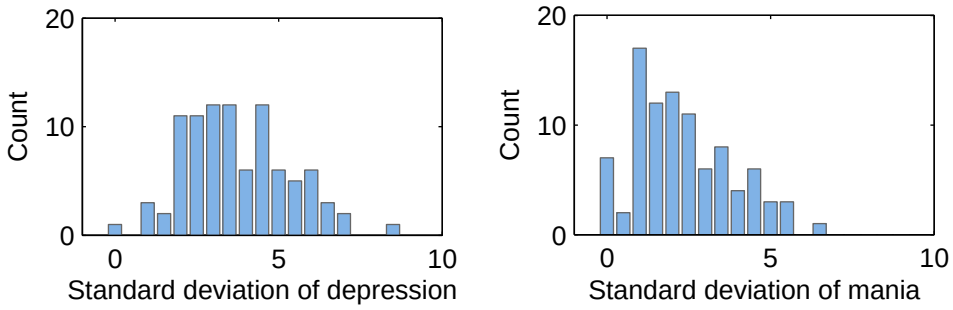
A.3 Statistics split by diagnostic subtype

Figure A.9 describes the age and time series length for Set D, split by diagnostic subcategory. The distributions for Bipolar I and Bipolar II patients are not markedly different. The mean and standard deviation of mood, split by diagnostic subcategory, are shown in Figures A.10 and Figure A.11. For mean depression scores, the median for the set of Bipolar I patients is 7.1 and for Bipolar II patients it is 8.7. The distribution of mean ratings for Bipolar I disorder is less dispersed than for Bipolar II disorder with an interquartile range of 2.7 for BPI compared with 8.2 for BPII. For mania ratings the median (iqr) for Bipolar I patients is 2.1 (2.9) and for Bipolar II patients it is 1.9 (2.3). The scores for individual symptoms of depression are shown in Figures A.12 and A.13. These figures show that patients with Bipolar II disorder have higher median ratings for individual symptoms than patients with Bipolar I disorder apart from *Concentration*. The distributions overall follow the same pattern as that for Set A with the symptom *Thoughts of death or suicide* having generally low scores and *Sleep*, high scores.



(a) Depression (b) Mania

Figure A.1: Distribution of the mean mood ratings for members of Set A (n=93).



(a) Depression (b) Mania

Figure A.2: Distribution of the standard deviation of mood ratings for members of Set A.

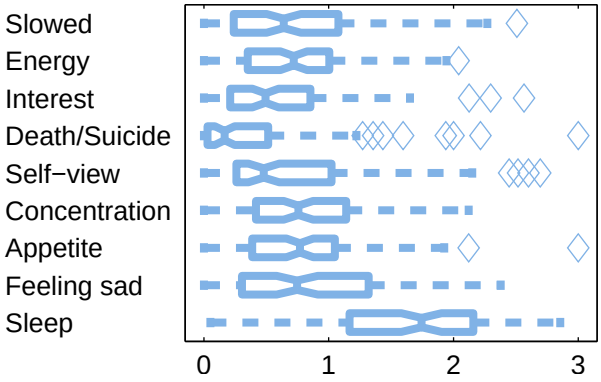
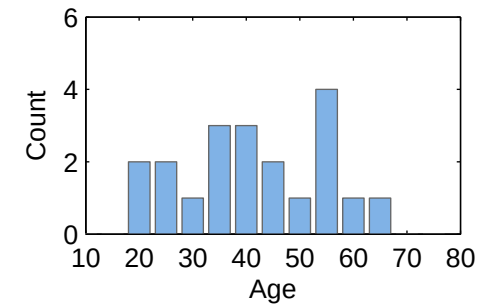
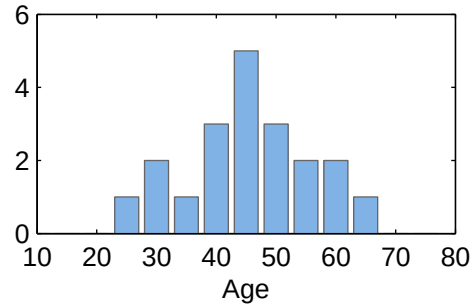


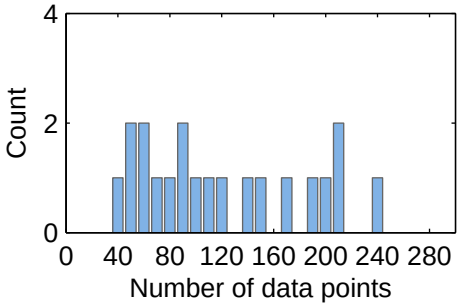
Figure A.3: Mean ratings for individual symptoms of depression. Points are drawn as outliers (diamonds) if they are larger than $q3 + 1.5 * (q3 - q1)$ or smaller than $q1 - 1.5(q3 - q1)$, where $q1$ and $q3$ are the 25^{th} and 75^{th} percentiles respectively. The whiskers extend to the adjacent value, which is the most extreme data value that is not an outlier.



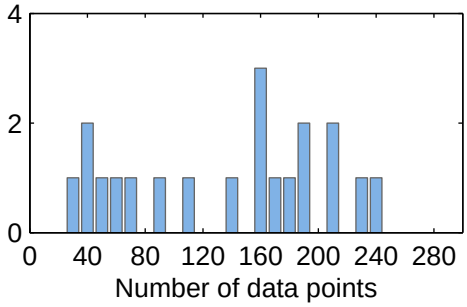
(a) Women



(b) Men

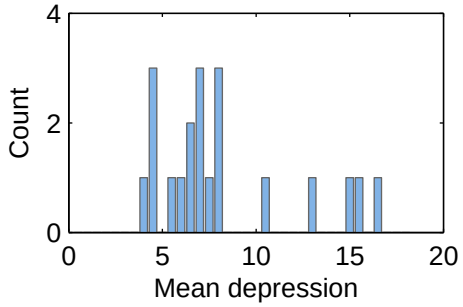


(c) Women

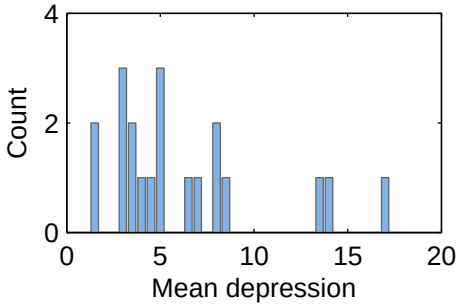


(d) Men

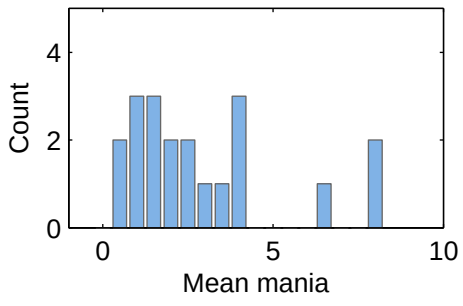
Figure A.4: Distribution of age and time series length for patients in Set G split by gender.



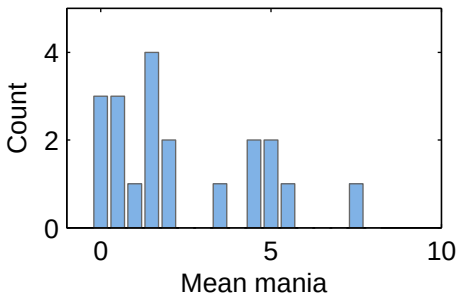
(a) Women



(b) Men

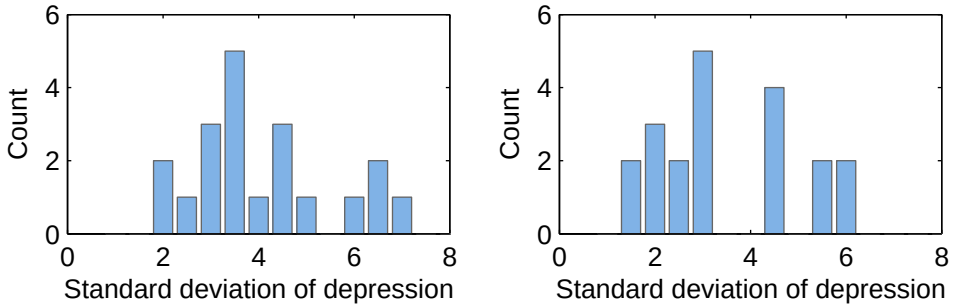


(c) Women



(d) Men

Figure A.5: Distribution of mean mood ratings for patients in Set G split by gender.



(a) Women (b) Men

Figure A.6: Distribution of the standard deviation of depression ratings for Set G.

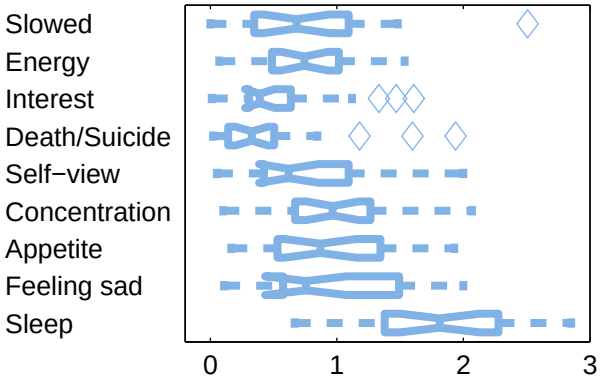


Figure A.7: Distribution of mean values for symptoms of depression in women in Set G.

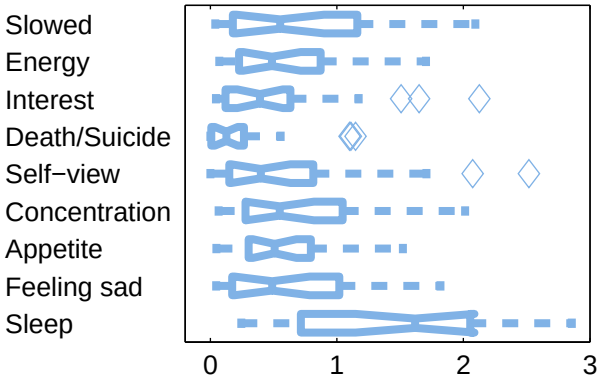
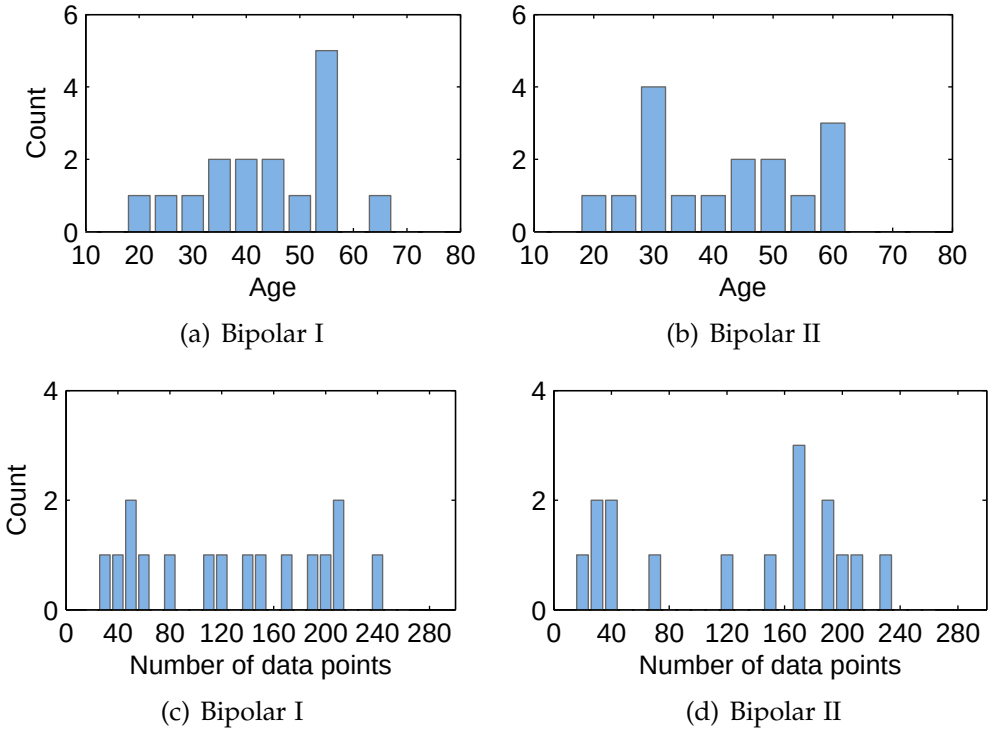


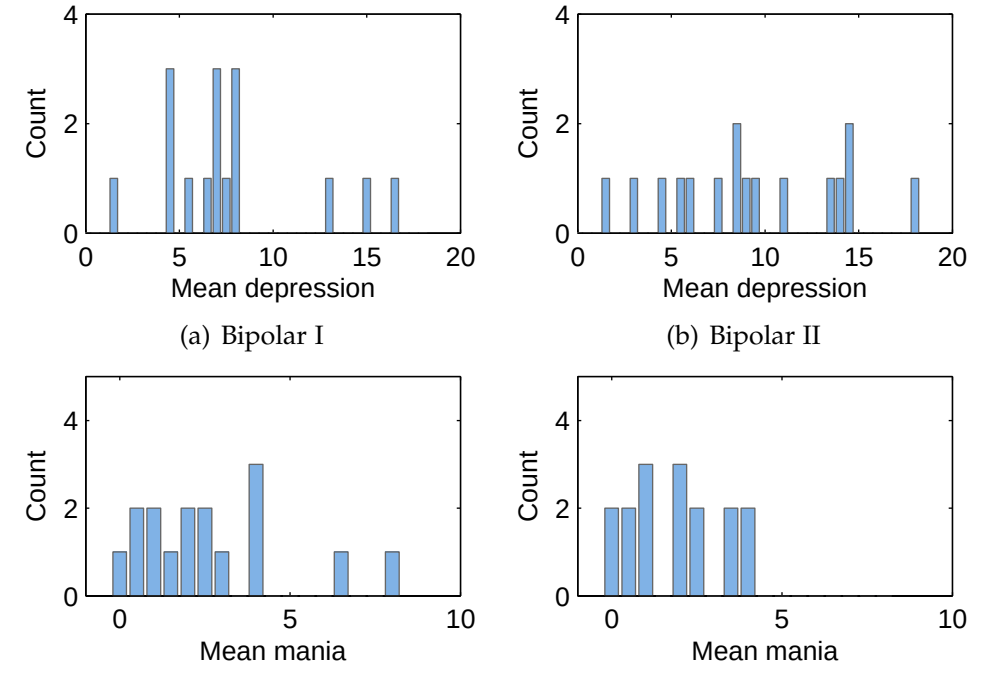
Figure A.8: Distribution of mean values for symptoms of depression in men in Set G.



(a) Bipolar I (b) Bipolar II

(c) Bipolar I (d) Bipolar II

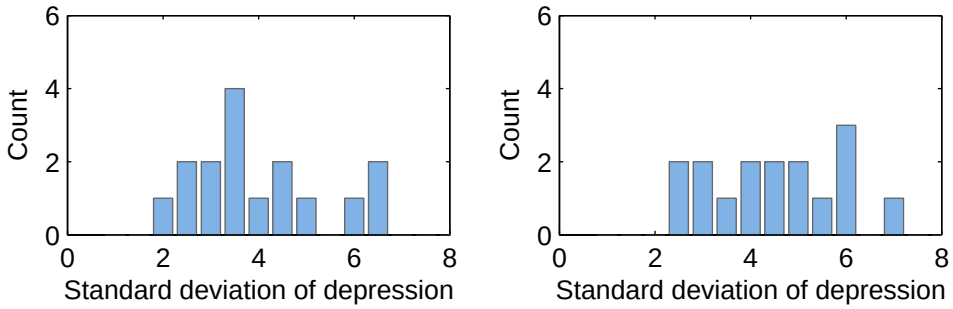
Figure A.9: Distribution of age and time series length for patients in Set D.



(a) Bipolar I (b) Bipolar II

(c) Bipolar I (d) Bipolar II

Figure A.10: Distribution of mean depression and mania ratings for patients in Set D.



(a) Bipolar I (b) Bipolar II

Figure A.11: Distribution of the standard deviation of depression ratings for Set D.

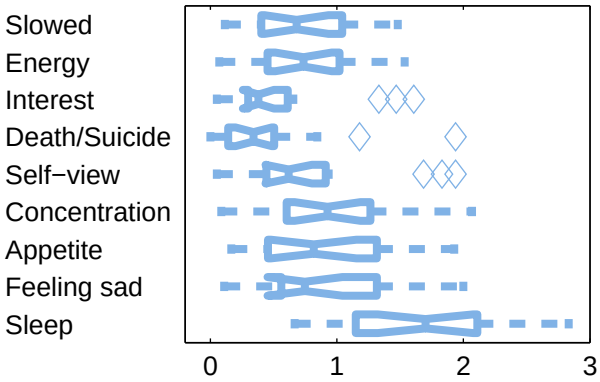


Figure A.12: Distribution of mean values for symptoms of depression for patients with Bipolar I disorder.

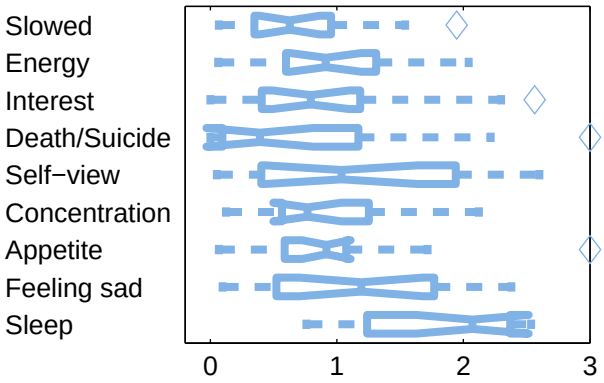
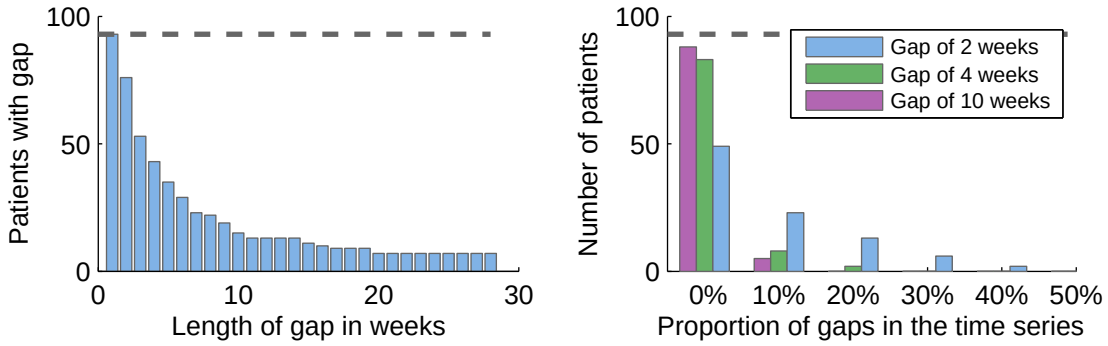


Figure A.13: Distribution of mean values for symptoms of depression for patients with Bipolar II disorder.

A.4 Interval analysis

Figure A.14(a) shows the prevalence of gaps of different length in the time series. Out of the 93 patients in Set A, 76 patients have at least one gap of two weeks or more. Longer gaps are less common in the set of time series: only 15 patients have at least one gap of 10 weeks or more. Figure A.14(b) shows the proportion of gaps in the individual time series. Again, gaps of 10 weeks are either not present, or form a small proportion of the intervals in a series: there are only 5 patients where the gaps form more than 5% of the time series.



(a) Presence of gaps

(b) Extent of gaps

Figure A.14: Analysis of gaps in the time series of patients in Set A. The left hand figure (a) shows the number of patients with at least one interval greater than or equal to the corresponding length on the x-axis. The dotted horizontal line indicates 93 patients, the number in Set A. All patients have at least one gap of one week or longer, 76 patients have at least one gap of two weeks or longer and 15 patients a gap of 10 weeks or longer. The right hand figure (b) shows the proportion of gaps in the time series using a histogram with bin widths of 10%. The leftmost column shows that 88 patients have time series with fewer than 5% gaps, where a gap is 10 weeks or longer. The remaining 5 patients from the set of 93 are shown in the second group of columns and have no more than 15% of their time series with gaps of this length.

Figure A.15 shows the distribution of all response intervals for participants in Set A. The mode is at 7 days, with counts dropping off rapidly with distance from the mode. The distribution is skew, with many more intervals greater than a week compared with less than a week. There are peaks at 14 and 21 days, these reflecting missed responses in a weekly schedule.

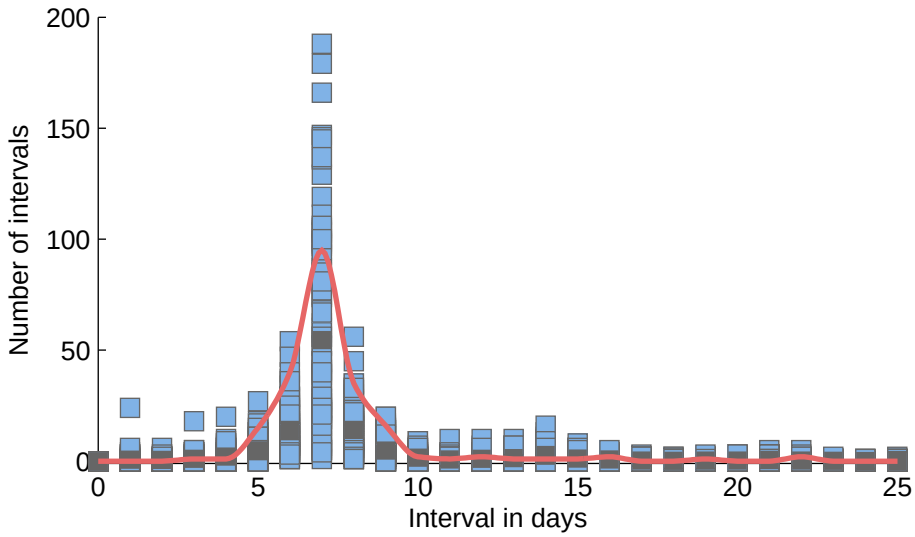


Figure A.15: Distribution of response intervals for participants in Set A. Each square marker represents an interval count for an individual patient. Mean values for counts at given intervals are shown as grey square markers. Peaks in the distribution occur at intervals of 7, 14 and 21 days. The line is a sample distribution from a single patient, smoothed for clarity and included for illustration.

Figure A.15 shows the distribution of all response intervals for participants in Set A. The mode is at 7 days, with counts dropping off rapidly with distance from the mode. The distribution is skew, with many more intervals greater than a week compared with less than a week. There are peaks at 14 and 21 days, these reflecting missed responses in a weekly schedule.

References

[1] Altman, E. G., Hedeker, D., Peterson, J. L., and Davis, J. M. The Altman self-rating mania scale. *Biological Psychiatry 42*, 10 (Nov 1997), 948–955.

[2] American Psychiatric Publishing, Inc. *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. Washington, DC, 2000.

[3] American Psychiatric Publishing, Inc. *DSM-IV-TR Guidebook*. Washington, DC, 2004.

[4] Anand, M. Nonlinear dynamic system model of bipolar mood disorder. *Proc. IEEE, SoutheastCon, 2007.* (2007), 279 – 282.

[5] Arora, S., Little, M. A., and McSharry, P. E. Nonlinear and nonparametric modeling approaches for probabilistic forecasting of the US gross national product. *Studies in Nonlinear Dynamics and Econometrics* (2011), 1–26.

[6] Arroyo, J., and Maté, C. Forecasting histogram time series with k-nearest neighbours methods. *International Journal of Forecasting 25*, 1 (2009), 192–207.

[7] Barkoulas, J., Baum, C., and Chakraborty, A. Nearest-neighbor forecasts of U.S. interest rates. *Boston College Working Papers in Economics*, 313 (2003).

[8] Bellman, R. A Markovian decision process. *Journal of Mathematics and Mechanics 6* (1957).

[9] Berridge, M., Downes, C., and Hanley, M. Neural and developmental actions of lithium: A unifying hypothesis. *Cell 59* (1989), 411–419.

[10] Bishop, C. M., et al. *Pattern recognition and machine learning*. Springer New York, 2006.

[11] Bonsall, M., Wallace-Hadrill, S., Geddes, J., Goodwin, G., and Holmes, E. Nonlinear time-series approaches in characterizing mood stability and mood instability in bipolar disorder. *Proceedings of the Royal Society B: Biological Sciences 279*, 1730 (2012), 916–924.

[12] Bopp, J. M., Miklowitz, D. J., Goodwin, G. M., Stevens, W., Rendell, J. M., and Geddes, J. R. The longitudinal course of bipolar disorder as revealed through weekly text messaging: a feasibility study, text message mood charting. *Bipolar Disorders 12*, 3 (May 2010), 327–334.

[13] Box, G., and Jenkins, G. *Time series analysis: Forecasting and control.*, first ed. San Francisco: Holden-Day, 1970.

[14] Brockwell, P., and Davis, R. *Introduction to time series and forecasting*. Springer Verlag, 2002.

[15] Broersen, P., de Waele, S., and Bos, R. The accuracy of time series analysis for laser Doppler velocimetry. In *Proc. 10th Int. Symp. Laser Techn. Fluid Mech* (2000), Springer-Verlag.

[16] Brown, R. *Statistical forecasting for inventory control*. New York: McGraw-Hill, 1959.

[17] Cade, J. F. Lithium salts in the treatment of psychotic excitement. *Medical Journal of Australia* (1949).

[18] Casdagli, M., Eubank, S., Farmer, J. D., and Gibson, J. State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena 51*, 1 (1991), 52–98.

[19] Chatfield, C. *Time-series forecasting*. Chapman and Hall/CRC, 2002.

[20] Chatfield, C. *The analysis of time series: an introduction*, 6th ed. Chapman and Hall/CRC, 2003.

[21] Chen, Z., Ivanov, P. C. H., Hu, K., and Stanley, H. E. Effect of nonstationarities on detrended fluctuation analysis. *Physical review. E, Statistical, Nonlinear, and Soft Matter Physics 65*, 4 Pt 1 (Apr. 2002).

[22] Cheng, L., Lumb, M., Polgar, L., and Mudge, A. W. How can the mood stabilizer VPA limit both mania and depression? *Mol Cell Neurosci 29* (2005), 155–161.

[23] Christodoulou, G., and Lykouras, E. Abrupt lithium discontinuation in manic-depressive patients. *Acta Psychiatrica Scandinavica 65*, 5 (1982), 310–314.

[24] Cleary, J. G., and Witten, I. Data compression using adaptive coding and partial string matching. *Communications, IEEE Transactions on 32*, 4 (1984), 396–402.

[25] Corlett, P., Taylor, J., Wang, X.-J., Fletcher, P., and Krystal, J. Toward a neurobiology of delusions. *Progress in neurobiology 92*, 3 (2010), 345–369.

[26] Craddock, N., Antebi, D., Attenburrow, M.-J., Bailey, A., Carson, A., Cowen, P., Craddock, B., Eagles, J., Ebmeier, K., Farmer, A., et al. Wake-up call for British psychiatry. *The British Journal of Psychiatry 193*, 1 (2008), 6–9.

[27] Cressie, N. Statistics for spatial data. *Terra Nova 4*, 5 (1992), 613–617.

[28] Dash, B., and Jounious, A. M. M. *Handbook of Ayurveda*. Concept Publishing Company, 1997.

[29] Daugherty, D., Roque-Urrea, T., Urrea-Roque, J., Troyer, J., Wirkus, S., and Porter, M. A. Mathematical models of bipolar disorder. *Communications in Nonlinear Science and Numerical Simulation 14*, 7 (Jul 2009), 2897–2908.

[30] Dawid, A. Properties of diagnostic data distributions. *Biometrics* (1976), 647–658.

[31] Dayan, P., and Huys, Q. J. Serotonin, inhibition, and negative mood. *PLoS Computational Biology 4* (2008).

[32] De Gooijer, J. G., and Hyndman, R. J. 25 years of time series forecasting. *Int. J. Forecast. 22*, 3 (2006), 443–473.

[33] De Gooijer, J. G., and Kumar, K. Some recent developments in non-linear time series modelling, testing, and forecasting. *International Journal of Forecasting 8*, 2 (1992), 135–156.

[34] Diebold, F. X., and Mariano, R. S. Comparing predictive accuracy. *Journal of Business & Economic Statistics 20*, 1 (2002).

[35] Diggle, P. *Time series: a biostatistical introduction*. Oxford University Press, 1990.

[36] Eckner, A. A framework for the analysis of unevenly-spaced time series data. Tech. rep., Working Paper, eckner.com, 2011.

[37] Edelson, R., and Krolik, J. The discrete correlation function–a new method for analyzing unevenly sampled variability data. *The Astrophysical Journal 333* (1988), 646–659.

[38] Erdogan, E., Ma, S., Beygelzimer, A., and Rish, I. Statistical models for unequally spaced time series. *Proceedings of SIAM Data Mining* (2005).

[39] Farmer, J. D., and Sidorowich, J. J. Predicting chaotic time series. *Physical review letters 59*, 8 (1987), 845.

[40] Faul, S., Gregorcic, G., Boylan, G., Marnane, W., Lightbody, G., and Connolly, S. Gaussian process modeling of EEG for the detection of neonatal seizures. *IEEE Transactions on Biomedical Engineering 54*, 12 (Dec 2007), 2151–2162.

[41] Fenn, H. H., Robinson, D., Luby, V., Dangel, C., Buxton, E., Beattie, M., Kraemer, H., and Yesavage, J. A. Trends in pharmacotherapy of schizoaffective and bipolar affective disorders: a 5-year naturalistic study. *American Journal of Psychiatry 153*, 5 (1996), 711–713.

[42] Fernández-Rodríguez, F., and Sosvilla-Rivero, S. Testing nonlinear forecastability in time series: theory and evidence from the EMS. *Economics Letters 59*, 1 (1998), 49–63.

[43] Fernández-Rodríguez, F., Sosvilla-Rivero, S., and Andrada-Félix, J. Exchange-rate forecasts with simultaneous nearest-neighbour methods: Evidence from the EMS. *International Journal of Forecasting 15*, 4 (1999), 383–392.

[44] Fletcher, P. C., and Frith, C. D. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience 10*, 1 (2008), 48–58.

[45] Friston, K. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences 360*, 1456 (2005), 815–836.

[46] Friston, K. The history of the future of the Bayesian brain. *NeuroImage 62*, 2 (2012), 1230–1233.

[47] Friston, K. J. Hallucinations and perceptual inference. *Behavioral and Brain Sciences 28*, 06 (2005), 764–766.

[48] Gardner, E. S. J. Exponential smoothing: The state of the art, part II. *International Journal of Forecasting 22* (2006), 637–666.

[49] Gelder, M., Mayou, R., and Geddes, J. Oxford core texts. *Psychiatry, 2nd edition, Oxford* (1999).

[50] George, D., and Hawkins, J. A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on* (2005), vol. 3, IEEE, pp. 1812–1817.

[51] Gigerenzer, G. Mindless statistics. *The Journal of Socio-Economics 33*, 5 (2004), 587–606.

[52] Glenn, T., Whybrow, P. C., Rasgon, N., Grof, P., Alda, M., Baethge, C., and Bauer, M. Approximate entropy of self-reported mood prior to episodes in bipolar disorder. *Bipolar Disorders 8*, 5p1 (Oct 2006), 424–429.

[53] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation 101*, 23 (2000).

[54] Goodwin, G. M., and Geddes, J. R. What is the heartland of psychiatry? *The British Journal of Psychiatry 191*, 3 (2007), 189–191.

[55] Gottschalk, A., Bauer, M. S., and Whybrow, P. C. Evidence of chaotic mood variation in bipolar disorder. *Archives of General Psychiatry 52*, 11 (Nov 1995), 947–959.

[56] Gottschalk, A., Bauer, M. S., and Whybrow, P. C. Low-dimensional chaos in bipolar disorder? – reply. *Archives of General Psychiatry 55*, 3 (Mar 1998), 275–276.

[57] Grassberger, P., and Procaccia, I. Characterization of strange attractors. *Physical review letters 50*, 5 (1983), 346–349.

[58] Haslett, J. On the sample variogram and the sample autocovariance for non-stationary time series. *Journal of the Royal Statistical Society: Series D (The Statistician) 46*, 4 (1997), 475–484.

[59] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. *The elements of statistical learning*, vol. 2. Springer, 2009.

[60] Healy, D. *Mania: A Short History of Bipolar Disorder*. Johns Hopkins University Press, 2006.

[61] Hegger, R., Kantz, H., and Schreiber, T. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos: An Interdisciplinary Journal of Nonlinear Science 9*, 2 (1999), 413–435.

[62] Hohwy, J. *The predictive mind*. Oxford University Press, 2013.

[63] Holmes, E., Deeprose, C., Fairburn, C., Wallace-Hadrill, S., Bonsall, M., Geddes, J., and Goodwin, G. Mood stability versus mood instability in bipolar disorder: a possible role for emotional mental imagery. *Behaviour research and therapy 49*, 10 (2011), 707–713.

[64] Holt, C. C. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting 20*, 1 (Jan 2004), 5–10.

[65] Hyndman, R., Koehler, A., Ord, K., and Snyder, R. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer, 2008.

[66] Insel, T. R., Cuthbert, B. N., Garvey, M. A., Heinssen, R. K., Pine, D. S., Quinn, K. J., Sanislow, C. A., and Wang, P. S. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *American Journal of Psychiatry 167*, 7 (2010), 748–751.

[67] Jakes, I. *Theoretical approaches to obsessive-compulsive disorder*, vol. 14. Cambridge University Press, 2006.

[68] Jamison, K. R. *Touched with fire*. Simon and Schuster, 1996.

[69] Jones, R. Time series analysis with unequally spaced data. *Handbook of statistics 5* (1985), 157–177.

[70] Judd, L., Akiskal, H., Schettler, P., Coryell, W., Endicott, J., Maser, J., Solomon, D., Leon, A., and Keller, M. A prospective investigation of the natural history of the long-term weekly symptomatic status of bipolar II disorder. *Arch Gen Psychiat 60(3)* (2003), 261 to 269.

[71] Judd, L. L. The long-term natural history of the weekly symptomatic status of bipolar I disorder. *Archives of General Psychiatry 59*, 6 (Jun 2002), 530–537.

[72] Kantz, H., and Schreiber, T. Human ECG: nonlinear deterministic versus stochastic aspects. In *Science, Measurement and Technology, IEE Proceedings-* (1998), vol. 145, IET, pp. 279–284.

[73] Kantz, H., and Schreiber, T. *Nonlinear Time Series Analysis*, first ed. Cambridge University Press, 2003.

[74] Karlsson, M., and Yakowitz, S. Nearest-neighbor methods for nonparametric rainfall-runoff forecasting. *Water Resources Research 23*, 7 (1987), 1300–1308.

[75] Katschnig, H. Are psychiatrists an endangered species? observations on internal and external challenges to the profession. *World Psychiatry 9*, 1 (2010), 21–28.

[76] Kingdon, D., and Young, A. H. Research into putative biological mechanisms of mental disorders has been of no value to clinical psychiatry. *The British Journal of Psychiatry 191*, 4 (2007), 285–290.

[77] Knill, D. C., and Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences 27*, 12 (2004), 712–719.

[78] Kolmogorov, A. N. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell Istituto Italiano degli Attuari 4*, 1 (1933), 83–91.

[79] Krystal, A. D. Low-dimensional chaos in bipolar disorder? *Archives of General Psychiatry 55*, 3 (Mar 1998), 275–276.

[80] Kugiumtzis, D. Surrogate data test for nonlinearity including nonmonotonic transforms. *Physical Review E 62*, 1 (2000), R25–R28.

[81] Lambdin, C. Significance tests as sorcery: Science is empirical–significance tests are not. *Theory & Psychology 22*, 1 (2012), 67–90.

[82] Lee Rodgers, J., and Nicewander, W. A. Thirteen ways to look at the correlation coefficient. *The American Statistician 42*, 1 (1988), 59–66.

[83] Little, M., McSharry, P., Moroz, I., and Roberts, S. Nonlinear, biophysically-informed speech pathology detection. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006 Proceedings.: Toulouse, France* (2006).

[84] Little, M., McSharry, P. E., Moroz, I., and Roberts, S. Testing the assumptions of linear prediction analysis in normal vowels. *Journal of the Acoustical Society of America 119*, 1 (2006), 549–558.

[85] Lorenz, E. N. Deterministic nonperiodic flow. *Journal of the atmospheric sciences 20*, 2 (1963), 130–141.

[86] Lorenz, E. N. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric sciences 26*, 4 (1969), 636–646.

[87] Manji, H., and Lenox, R. The nature of bipolar disorder. *J Clin Psychiatry 61* (2000), 42–57.

[88] McGuire-Snieckus, R., McCabe, R., and Priebe, S. Patient, client or service user? A survey of patient preferences of dress and address of six mental health professions. *Psychiatric Bulletin 27*, 8 (2003), 305–308.

[89] McSharry, P. E. The danger of wishing for chaos. *Nonlinear Dynamics, Psychology, and Life Sciences 9*, 4 (2005), 375–397.

[90] Meng, Q., and Peng, Y. A new local linear prediction model for chaotic time series. *Physics Letters A 370*, 5 (2007), 465–470.

[91] Mezzich, J. International surveys on the use of ICD-10 and related diagnostic systems. *Psychopathology. 35*, 2-3 (2002), 72–75.

[92] Miller, G. Why is mental illness so hard to treat? *Science 338*, 6103 (2012), 32–33.

[93] Mitchell, P. B. On the 50th anniversary of John Cade's discovery of the anti-manic effect of lithium. *Australian and New Zealand Journal of Psychiatry 33*, 5 (1999), 623–628.

[94] Mobaji, H., and Karahalios, K. HCI applications for aiding children with mental disorder. *Crossroads, ACM Student Magazine 12(2)* (2005).

[95] Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. Computational psychiatry. *Trends in cognitive sciences 16*, 1 (2012), 72–80.

[96] Moore, P. J., Little, M. A., and McSharry, P. E. Forecasting depression in bipolar disorder. In *Proceedings, 31st International Symposium on Forecasting* (2011), International Institute of Forecasters.

[97] Moore, P. J., Little, M. A., McSharry, P. E., Geddes, J., and Goodwin, G. Forecasting depression in bipolar disorder. *IEEE Transactions on Biomedical Engineering 59*, 10 (2012).

[98] Moore, P. J., Little, M. A., McSharry, P. E., Goodwin, G. M., and Geddes, J. R. Correlates of depression in bipolar disorder. *Proceedings of the Royal Society B: Biological Sciences 281*, 1776 (2014).

[99] Moran, P. The statistical analysis of the canadian lynx cycle. I: structure and prediction. *Aust. J. Zool 1* (1953), 163 to 173.

[100] Mumford, D. On the computational architecture of the neocortex. *Biological cybernetics 66*, 3 (1992), 241–251.

[101] Muth, J. F. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association 55*, 290 (Jun 1960), 299.

[102] National Institute for Health and Care Excellence. *CG38: The management of bipolar disorder in adults, children and adolescents, in primary and secondary care.* London, 2006.

[103] O'Hagan, A. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society 40*, 1 (1978), 1–42.

[104] Osborne, A. R., and Provenzale, A. Finite correlation dimension for stochastic systems with power–law spectra. *Physica D: Nonlinear Phenomena 35*, 3 (1989), 357–381.

[105] Pandi-Perumal, S. R., and Kramer, M. *Sleep and mental illness*. Cambridge University Press, 2010.

[106] Paparella, F., Provenzale, A., Smith, L. A., Taricco, C., and Vio, R. Local random analogue prediction of nonlinear processes. *Physics Letters A 235*, 3 (1997), 233–240.

[107] Pearson, K. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London 58*, 347-352 (1895), 240–242.

[108] Peng, C. K., Havlin, S., Stanley, H., and Goldberger, A. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time-series. *Chaos 5*, 1 (1995), 82–87.

[109] Pincus, S. Approximate entropy as a measure of system complexity. *Proc Natl Acad Sci USA 88* (1991), 2297 to 2301.

[110] Pincus, S. M. Quantitative assessment strategies and issues for mood and other psychiatric serial study data. *Bipolar Disorders 5*, 4 (Aug 2003), 287–294.

[111] Rasmussen, C. *Evaluation of Gaussian Processes and other Methods for Non-linear Regression.* PhD thesis, University of Toronto, 1996.

[112] Rasmussen, C., and Williams, C. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[113] Rasmussen, C. E. Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning*. Springer, 2004, pp. 63–71.

[114] Rehfeld, K., Marwan, N., Heitzig, J., and Kurths, J. Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics 18*, 3 (2011), 389–404.

[115] Rush, A. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological psychiatry 54*, 5 (2003), 573.

[116] Rush, A., Gullion, C. M., Basco, M. R., Jarrett, R. B., Trivedi, M. H., et al. The inventory of depressive symptomatology (IDS): psychometric properties. *Psychological medicine 26*, 3 (1996), 477–486.

[117] Ryabko, B. Y., and Monarev, V. A. Experimental investigation of forecasting methods based on data compression algorithms. *Problems of Information Transmission 41*, 1 (2005), 65–69.

[118] Salimi-Khorshidi, G., Nichols, T. E., Smith, S. M., and Woolrich, M. W. Using Gaussian-process regression for meta-analytic neuroimaging inference based on sparse observations. *IEEE Transactions on Medical Imaging 30*, 7 (Jul 2011), 1401–1416.

[119] Scargle, J. Studies in astronomical time series analysis. II-statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal 263* (1982), 835–853.

[120] Schneider, T., and Neumaier, A. Algorithm 808: ARfit–a Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS) 27*, 1 (2001), 58–65.

[121] Schreiber, T., and Schmitz, A. Improved surrogate data for nonlinearity tests. *Physical Review Letters 77*, 4 (1996), 635–638.

[122] Small, M. *Applied nonlinear time series analysis: applications in physics, physiology and finance*, vol. 52. World Scientific, 2005.

[123] Small, M. *Dynamics of Biological Systems*. Chapman & Hall/CRC Mathematical & Computational Biology. Taylor & Francis, 2011.

[124] Small, M., and Tse, C. K. Optimal embedding parameters: a modelling paradigm. *Physica D: Nonlinear Phenomena 194*, 3 (2004), 283–296.

[125] Small, M., Yu, D., and Harrison, R. G. Surrogate test for pseudoperiodic time series data. *Physical Review Letters 87*, 18 (2001), 188101.

[126] Stegle, O., Fallert, S., MacKay, D., and Brage, S. Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering 55*, 9 (Sep 2008), 2143–2151.

[127] Sundararajan, S., and Keerthi, S. S. Predictive approaches for choosing hyperparameters in Gaussian processes. *Neural Computation 13*, 5 (2001), 1103–1118.

[128] Sutton, R. S., and Barto, A. G. *Reinforcement Learning*. MIT Press, Cambridge, MA, 1998.

[129] Takens, F., et al. Dynamical systems and turbulence. *Lecture notes in mathematics 898*, 9 (1981), 366.

[130] Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., and Doyne Farmer, J. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena 58*, 1 (1992), 77–94.

[131] Tong, H., and Lim, K. S. Threshold autoregression, limit cycles and cyclical data. *Journal of the Statistical Society 42*, 3 (1980), 245.

[132] Toth, E., Brath, A., and Montanari, A. Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology 239*, 1 (2000), 132–147.

[133] Voss, R. F. 1/f noise: diffusive systems and music. Tech. rep., California Univ., Berkeley (USA). Lawrence Berkeley Lab., 1975.

[134] Wehr, T. A., and Goodwin, F. Rapid cycling in manic-depressives induced by tricyclic antidepressants. *Arch Gen Psychiat 36* (1979), 555 to 559.

[135] White, P., Rickards, H., and Zeman, A. Time to end the distinction between mental and neurological illnesses. *BMJ: British Medical Journal 344* (2012).

[136] Williams, C. K., and Barber, D. Bayesian classification with Gaussian processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 20*, 12 (1998), 1342–1351.

[137] Williams, R., Cheng, L., Mudge, A., and Harwood, A. A common mechanism of action for three mood-stabilizing drugs. *Nature 417* (2002), 292–295.

[138] Winters, P. R. Forecasting sales by exponentially weighted moving averages. *Management Science 6*, 3 (Apr 1960), 324–342.

[139] World Health Organization. *Manual of the international statistical classification of diseases, injuries, and causes of death, Sixth Revision.* Geneva, 1949.

[140] Wu, C., and Chau, K. Data-driven models for monthly streamflow time series prediction. *Engineering Applications of Artificial Intelligence 23*, 8 (2010), 1350–1367.

[141] Yakowitz, S. Nearest neighbour methods for time series analysis. *Journal of time series analysis 8*, 2 (1987), 235–247.

[142] Yeragani, V. K., Pohl, R., Mallavarapu, M., and Balon, R. Approximate entropy of symptoms of mood: an effective technique to quantify regularity of mood. *Bipolar Disorders 5*, 4 (Aug 2003), 279–286.

[143] Yu, K., and Jones, M. Local linear quantile regression. *Journal of the American statistical Association 93*, 441 (1998), 228–237.

[144] Yule, G. U. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society A Mathematical Physical and Engineering Sciences 226*, 636-646 (Jan 1927), 267–298.