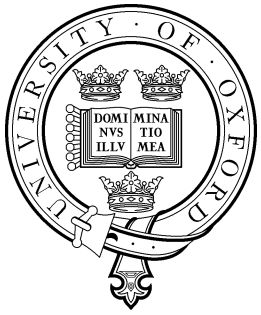
Personalised Modelling of Geographic Movements in Depression



Niclas Palmius
Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2018

Personalised Modelling of Geographic Movements in Depression

Niclas Palmius
Wolfson College

Doctor of Philosophy
Trinity 2018

It has been estimated that as many as 41 % of individuals will experience depression at some point in their lives, a serious and debilitating mental illness. One way to improve the care that patients receive, and reduce the burden on the healthcare system, is to monitor the health of patients over time. One emerging area of longitudinal monitoring is using behavioural data, especially from smartphones, to assess the health of individuals in the community. This thesis explores the relationship between depression and the geographic movements of individuals.

Geolocation data were analysed from a longitudinal study of over 130 participants, including patients with bipolar disorder and borderline personality disorder, one of the largest datasets from these patient groups in the community. After suitable preprocessing, several features hypothesised to be related to mental health were extracted from the raw geographic coordinates.

A new method for clustering stationary places is described, which is shown to provide more consistent and reliable results than alternative methods. The features used were extended from previous work to provide better characterisation of individuals. To reduce noise in the extracted features and improve predictive performance, a Kalman filter-based technique is presented. This is demonstrated to provide improved performance over the unfiltered features, demonstrating the importance of considering noise in the feature space as well as in the data space.

In patients diagnosed with bipolar disorder, statistical differences are shown between weeks where participants self-reported clinical-level depression and weeks where they did not. Classification of depression from the geolocation-derived features is demonstrated in these patients, achieving an accuracy of over 75 %.

Regression of the level of depression, however, proved challenging, and it is demonstrated that one of the reasons for this is that individuals naturally have different baseline values of features, but also fundamentally different behavioural traits under different levels of depression.

A new technique is therefore presented that finds groups of individuals that have similar regression models linking their geolocation-derived features with their self-

reported depression levels. This is applied to 59 participants who provided more than 5 weeks of data for analysis, and finds that there are indeed several groups of individuals that appeared to behave similarly. Regression models trained on these grouped individuals significantly improve on the regression results from models trained over all participants and models trained on limited calibration data.

Overall, this thesis demonstrates the utility of geolocation as a marker of depression, but also highlights the importance of appropriate handling of data to maximise performance. Personalisation of models is also demonstrated to be critical for maximising performance.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

Ac . . . . . Classification accuracy

AMoSS . . . . The Automated Monitoring of Symptom Severity study

APA . . . . . American Psychiatric Association

ASRM . . . . Altman self-rating mania scale (Altman et al., 1997)

AUC . . . . . Area under (the ROC) curve

BD . . . . . Bipolar disorder

BMI . . . . . Body mass index

BPD . . . . . Borderline personality disorder

BPS . . . . . Bipolar spectrum disorder

CDF . . . . . Cumulative distribution function

CQC . . . . . The Care Quality Commission

CRP . . . . . Chinese restaurant process

CV . . . . . Cross-validation

DALY . . . . . Disability-adjusted life-year

DP . . . . . Dirichlet process

DSM . . . . . Diagnostic and Statistical Manual of Mental Disorders

DSM 5 . . . . 5^{th} edition of the DSM (APA, 2013)

DSM-IV . . . 4^{th} edition of the DSM (APA, 2000)

ECG . . . . . Electrocardiogram

EMA . . . . . Ecological momentary assessment (questionnaire)

EQ-5D . . . . EuroQol Group 6 question quality of life questionnaire (EuroQol Group, 1990)

ESA . . . . . Emotional self-awareness (questionnaire)

F_1 . . . . . F_1 score

FPR . . . . . False positive rate

GAD . . . . . Generalised anxiety disorder

GAD-7 . . . . The generalised anxiety disorder 7-item questionnaire (Spitzer et al., 2006)

GLM . . . . . Generalised linear model

GP . . . . . General practitioner

GPS . . . . . Global positioning system

HC . . . . . Healthy control

i.i.d. . . . . Independent and identically distributed

LDA . . . . . Linear discriminant analysis

MAE . . . . . Mean absolute error

MAP . . . . . Maximum *a posteriori*

MC . . . . . Monte Carlo (method)

MCMC . . . . Markov chain Monte Carlo (method)

MDD . . . . . Major depressive disorder

MDE . . . . . Major depressive episode

MLE . . . . . Maximum likelihood estimate

NHS . . . . . UK National Health Service

NICE . . . . . The National Institute for Health and Care Excellence

NMAE . . . . Normalised mean absolute error

NRMSD . . . Normalised root mean square deviation

PCA . . . . . Principal component analysis

PDF . . . . . Probability density function

PDF . . . . . Probability density function (of continuous distribution)

PMF . . . . . Probability mass function

PPG . . . . . Photoplethysmography

QDA . . . . . Quadratic discriminant analysis

QIDS . . . . . Used interchangeably with QIDS-SR_{16}

QIDS-SR_{16} . . 16-item quick inventory of depressive symptomatology questionnaire (Rush et al., 2003)

RMSD . . . . Root mean square deviation

ROC . . . . . Receiver operating characteristic curve – a plot of the FPR against the TPR

Se . . . . . Classification sensitivity (synonymous with TPR)

Sp . . . . . Classification specificity

SVD . . . . . Singular value decomposition

TPR . . . . . True positive rate (synonymous with sensitivity)

UK . . . . . United Kingdom

WHO . . . . . World Health Organization

Guide to Notation

This section describes the notation used in this thesis.

Constants and variables

- Constants are represented as italic upper-case Latin characters:

$$A = 10$$

- Scalar values are represented as italic Greek or lower-case Latin characters:

$$\alpha = a = 3$$

- Vectors are represented as bold Greek or upright lower-case Latin characters:

$$\boldsymbol{\alpha} = \mathbf{a} = [1, 2, 3, 4]^\top$$

- Vectors are columns unless otherwise specified.
- Matrices are represented as bold upper-case Greek or Latin characters:

$$\boldsymbol{\Delta} = \mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

- Feature matrices are $N \times p$ with p features on the columns and N data samples on the rows, i.e. $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top$ where $\mathbf{z}_i = [z_{i1}, \ldots, z_{ip}]^\top$.

General notation

$\lvert \cdot \rvert$	Absolute value.
$\lvert\lvert \cdot \rvert\rvert$	Euclidean norm.
$\mathbf{1}_n$	Column vector of n ones.
$\boldsymbol{a}^\top$ / $\mathbf{A}^\top$	Vector / matrix transpose.

$\mathbf{I}_n$	Diagonal matrix $\mathbf{I}_n = (d_{i,j})$ of size $n \times n$ where $d_{i,j} = 1$ if $i = j \; \forall \; i, j \in [1, n]$, $d_{i,j} = 0$ otherwise.
$\mathbb{I}(\cdot)$	Indicator function where $\mathbb{I}(\cdot) = 1$ if the expression in the brackets is true, or $\mathbb{I}(\cdot) = 0$ otherwise.
$\ell(\cdot)$	Loss function.
θ	General set of model parameters or variables.

Probability notation

$\text{cov}(\cdot)$	Covariance matrix of data points.
$\mathbb{E}(\cdot)$	Expected value, or expectation.
$F_X(\cdot)$	Cumulative distribution function of random variable X.
$\mathcal{L}(\cdot)$	Likelihood function.
$p(\cdot)$	Probability distribution / density / mass function (see notes).
$p(\cdot \mid \cdot)$	Conditional probability (conditional on random variable).
$p(\cdot \,;\, \cdot)$	Conditional probability (parametrised by constant value).
$\Phi(\cdot)$	Normal (Gaussian) cumulative distribution function.
$\pi(\cdot)$	Undefined/unspecified prior distribution.

Probability distributions

$\text{Bern}(\cdot)$	Bernoulli distribution
$\text{Beta}(\cdot)$	Beta distribution
$\text{Cat}(\cdot)$	Categorical distribution
$\text{Dir}(\cdot)$	Dirichlet distribution
$\text{Exp}(\cdot)$	Exponential distribution
$\text{Exp}_n(\cdot)$	Exponential distribution returning n independent values
$\text{Inv-Gamma}(\cdot)$	Inverse gamma distribution
$\mathcal{N}(\cdot)$	Normal (Gaussian) distribution
$\mathcal{N}_n(\cdot)$	Multivariate normal (Gaussian) distribution with n dimensions
$\text{Unif}(\cdot)$	Uniform distribution

Notes

The term *probability mass function* (PMF) is used to describe the probability distribution for a discrete random variable, and the term *probability density function* (PDF) is used to describe the probability distribution for a continuous random variable. Probability distributions are denoted $p(\cdot)$ for both discrete and continuous random variables, and for clarity distribution parameters may be included as constant model parameters, i.e. $p(\cdot\,;\,\cdot)$, or as conditioned variables, i.e. $p(\cdot\,|\,\cdot)$. The words 'distribution' and 'mass' are used interchangeably for discrete random variables, and the words 'distribution' and 'density' are used interchangeably for continuous random variables. The word 'distribution' is thus used for both continuous and discrete random variables. Where a random variable is distributed according to one of the standard distributions above, the standard distribution notation may be used in place of $p(\cdot)$ for the distribution by parametrising or conditioning on the distribution parameters. For example, if $x \sim \mathcal{N}(\mu, \sigma)$ then

$$p(x) = \begin{cases} p(x\,;\,\mu,\sigma) = \mathcal{N}(x\,;\,\mu,\sigma) & \text{if } \mu \text{ and } \sigma \text{ are constant model parameters;} \\ p(x\,|\,\mu,\sigma) = \mathcal{N}(x\,|\,\mu,\sigma) & \text{if } \mu \text{ and } \sigma \text{ are random variables in a} \\ & \text{hierarchical model.} \end{cases}$$

The *cumulative distribution function* (CDF) of a numeric random variable X is denoted with a capital letter F_X, and is defined as

$$F_X(x) = p(X \;\leq\; x) = \begin{cases} \displaystyle\int_{-\infty}^{x} p(z)\,dz & \text{if } X \text{ is continuous; or} \\ \displaystyle\sum_{z \leq x} p(z) & \text{if } X \text{ is discrete.} \end{cases}$$

As with the probability distribution function, distribution parameters may be included as parametrised or conditioned variables, i.e. $F_X(x\,;\,\theta)$ or $F_X(x\,|\,\theta)$. Where the context is clear, the random variable subscript may be omitted by simply writing $F(x\,|\,\theta)$. In the case where $x \sim \mathcal{N}(\mu, \sigma)$,

$$F(x\,|\,\mu,\sigma) = \Phi(x\,|\,\mu,\sigma) = \int_{-\infty}^{x} \mathcal{N}(z\,|\,\mu,\sigma)\,dz.$$

It is health that is real wealth and not pieces of gold and silver.

— Mahatma Gandhi

Chapter 1

Introduction

It has often been said that there is *no health without mental health* (M. Prince et al., 2007; HMG/DH, 2011; Becker & Kleinman, 2013). This simple statement – first endorsed by the European regional office of the World Health Organization (WHO) in 2005 (WHO Europe, 2005) – touches on the central role of cognition and mental well-being to human existence. Otherwise fit and healthy individuals can be debilitated by mental illnesses – "mental pain is as real as physical pain, and it is often more severe", state Layard et al. (2012). Haig (2015) provides a first-hand insight into the pain of living with what he calls "the double-whammy of anxiety and depression":

> *This wasn't having a crazy thought. This wasn't being a bit wacky. This was pain. I had been okay and now, suddenly, I wasn't. I wasn't well. So I was ill.*

The WHO reports that neuropsychiatric conditions account for around one third of years lost to disability among adults aged over 15 years (Mathers et al., 2008). In the United Kingdom it is estimated that mental health issues account for almost half of all health issues for people under the age of 65 (Layard et al., 2012). The percentage of the UK population suffering from mental illness is estimated at 17 %, with most cases being depression (8 %) or anxiety disorders (8 %) (Layard et al., 2012). Moffitt et al. (2010) suggests that as many as 41 % of individuals will experience depression in their lifetime. Similar statistics are seen in absenteeism from work (NICE, 2009; WHO, 2003), and in proportions of incapacity benefit claimants, where over 70 % are

for mental or behavioural disorders (Cattrell et al., 2011). Madsen et al. (2017) have also shown that job strain is associated with an increased risk of clinical depression.

The UK National Health Service (NHS) spends over £12 bn annually on mental health services, but this still represents only 13 % of the total NHS budget (Layard et al., 2012). Recent austerity programmes have lead to reductions in income for up to 40 % of mental health trusts (Gilburt, 2016) meaning that this relatively small amount is being increasingly stretched. Additionally, a further £22 bn funding shortage throughout the health service needs to be closed by 2020/21 through efficiency savings (NHS, 2014; Dunn et al., 2016). It is highly likely that mental health services will be affected by the required frugality.

At the same time, a combination of factors including reduced stigma surrounding mental health; an ageing population; and improved handling of mental illness in the criminal justice system, means that demand for mental health services is expected to rise (CQC, 2017).

In this setting, it is crucial to be able to maximise the efficiency of the available resources while ensuring that patients receive the care that they require. One way that the NHS is aiming to do this is by developing the care available in the community to improve efficiency as well as patient experience (NHS England, 2016).

While these changes to the funding and delivery of mental health services will be challenging to healthcare providers, it presents an opportunity for radical innovation and reform within the mental health system (McDaid & Knapp, 2010). Application of new technology is likely to take a central role, and yet technological innovation in mental health has been scant compared to other areas of healthcare, something that will be explored in greater detail in section 1.2 and chapter 2.

One area that has only recently started to become feasible is the use of long-term behavioural data to assess the mental health of individuals. A source of such behavioural data are modern smartphones, which are highly pervasive and often include sensors that can measure levels of physical activity, geographic location and exposure to light levels – all features that are thought to relate to mental illness. The increased comfort and reduced power requirements of modern physiological sensors,

sometimes embedded in other consumer devices such as smartwatches, means that longitudinal physiological data can also be collected and analysed. Related work using these types of pervasive sensors will be discussed in detail in sections 2.2.2 to 2.2.4.

1.1 Innovation in healthcare

Healthcare provision has changed dramatically over the past century. Innovation can be found across the field, from improved diagnostics, treatments, pharmaceuticals and disease management to new methods of service delivery and funding.

On a national (and indeed global) level, innovation in healthcare aims to optimise what Berwick et al. (2008) call the "triple aims" of healthcare:

a) improving the experience of care for individuals;

b) improving the overall health of populations; and

c) reducing the costs of care for populations (per capita).

This is a careful balancing act where pursuing one aim can have negative results on another. Further complicating realisation of progress towards these aims is that although they are very relevant to individuals and society, often they are not all relevant to individual businesses or service providers.

Much of the innovation in healthcare is intrinsically linked to technology, ranging from advanced imaging through surgical robotics and automated sample analysis. While advances in medicine are largely driven by increases in our understanding of physiology and anatomy, technology is often the enabler of change.

One emerging area of healthcare innovation where technology is playing a central role is in the collection and processing of medical data. Data collection has long been an integral part of healthcare. From the earliest days, vital signs about patients have been diligently collected and recorded. Whilst data were traditionally (and in many cases still are) collected by hand and recorded on paper notes (Wilson et al., 2013), modern technology enables data to be automatically and continuously collected. The types of data that can be collected are also being transformed, with many sensors now able to provide continuous longitudinal data. At the same time, advances in data

storage technology mean that it is now possible to store this continuously collected data long term. This enables databases of patient data to be explored in search of new markers of illness that may not have been previously available (Wong et al., 2017).

The combination of technology with the ability to collect and store vast amounts of data from patients gives us the potential to use this to improve delivery of care. This may be through improved selection of treatment, earlier detection of ineffective treatment or detection of changes in condition.

Although it is now possible to continually collect orders of magnitude more data than a century ago, one thing that has changed little is our ability as individuals to analyse it and extract useful information. One of the current themes in healthcare research, therefore, is the development of new analytic techniques for healthcare data. This is where machine learning techniques, often used together with disparate yet parsimonious sensors to collect data, can provide a revolutionary change to the way that healthcare is delivered.

1.2 Innovation in mental healthcare

While it is clear that technology has transformed most areas of healthcare, mental health has seen comparatively little technological innovation. The diagnostic front still largely relies on traditional and subjective methods of diagnosis performed using check-list type criteria (Tuso, 2014) often based on the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) published by the American Psychiatric Association (APA), now in it's fifth edition (APA, 2013). Mental health professionals are acutely aware of the limitations of their current assessment protocols (Owen, 2014). To quote Singh & Rose (2009), "psychiatry has long been a second-class citizen in science and medicine".

One of the core challenges with psychiatry is the disconnect between the etiological and/or pathogenic processes that underlie mental illness, and the syndromic categories used for diagnosis (Svrakic et al., 2002; G. Miller, 2010; Hickie et al.,

2013). Diagnostic manuals, such as the DSM, attempt to split the range of syndromic features into convenient distinct illnesses (Hyman, 2010; Owen, 2014). This is an intuitive concept, appealing to humanity's desire to split our knowledge into neat compartments (Rosch, 1999), a notion that stretches right back to the philosophy of Aristotle (Edel, 1995). It is also relatively reliable, with field trials of a selection of DSM-5 diagnoses by Regier et al. (2013) showing mixed, but overall good, agreement between clinicians.

Concerns have been raised, however, that this reliability of diagnosis, may come at the cost of *validity* of diagnosis (Conway et al., 2017), i.e. the reliability may hide different underlying pathways of illness. Mental health researchers suspect that single illnesses defined in the DSM may actually have several different underlying processes (G. Miller, 2010; Nasrallah, 2013; van Loo et al., 2012). Similarly, patients often present with comorbid symptoms (Jacobi et al., 2004; Kessler et al., 2007; Moffitt et al., 2007) that can make diagnosis into the DSM categories challenging (Wu & Fang, 2014), and may suggest that multiple conditions should be combined (Moffitt et al., 2007). There is some evidence that different presentations of comorbidity are found in different demographic groups, indicating that they may represent different etiologic pathways (Lamers et al., 2011). These factors may also go some way to explaining differences in the efficacy of medication or treatment on individuals that otherwise present with similar symptoms (Fava et al., 2008; Hickie et al., 2013). Large databases of patient data, often from multiple sources, may provide opportunities to better understand mental illnesses (Monteith et al., 2016). For example, there have been attempts to use data-driven approaches to better understand features of various mental illnesses from electronic medical records (Lyalina et al., 2013; V. M. Castro et al., 2015), although few practical results have emerged.

Fundamentally, the interest in the root causes of mental illness stem from the desire to improve patient care. While the questions surrounding the value and validity of psychiatric diagnoses (Heckers, 2015) will be debated for years to come, millions of

patients are suffering right now. There is therefore growing interest in using technology to assist in mental health assessment, treatment, and monitoring. Current work in this area will be explored in greater detail in section 2.2.

→

1.3 Thesis outline

This thesis explores in depth the specific relationship between mental health and the geographic movements of individuals as measured from their mobile phones. Specifically, it explores the following topics:

a) the differentiation of mental health disorders using geographic movements;
b) the detection of severity of depression using geographic movements; and
c) improvements in the detection of severity of depression through the development of a new personalised non-parametric regression model.

1.3.1 Contributions

This thesis makes contributions in a number of areas relevant to the development of geolocation-derived markers of mental health and the analysis of these types of markers.

The patient populations studied (described in chapter 4) and the length of data available from these patients are unique. While patients with bipolar disorder are relatively well studied, the community-based nature of the patients is unusual. Patients with borderline personality disorder are much less studied, and this is the first study to explore the relationship between geolocation-derived behavioural markers and self-reported depression levels in this patient group.

New methods for imputing missing geolocation data are presented in section 5.5. New methods of extracting and clustering stationary places visited in the time-series are also presented in sections 5.8 and 5.9 respectively. These methods combined are crucial to ensuring that the features calculated from the geolocation data accurately describe the behaviour of individuals in the presence of missing data in the raw data stream. The new location clustering method is demonstrated to provide improved consistency over alternative methods when applied to geolocation data.

The features extracted from the geolocation data are based on previous work by Saeb et al. (2015a) and others, but have been extended in a number of ways. Firstly new ways to quantify the diurnal regularity of individuals were developed

(described in sections 5.10.9 and 5.10.10). These are demonstrated to provide similar or improved performance when individually classifying depression in section 6.1.1 and improved performance when quantifying depression levels in section 6.4.2. A new adaptive method of filtering the extracted features based on a Kalman filter is described in section 6.1.2.2, which enables online filtering of the feature data to remove noise, emulating offline window-based smoothing. This method adapts to handle missing data and provides superior performance over an equivalent window filter. The improvement in performance is demonstrated in sections 6.1.4 and 6.1.5. Previous studies have not considered the effects of noise in the extracted features. Finally, section 6.1.3 describes filtering of the weeks of data included in the analysis to exclude weeks based on a criteria of missing data and time spent away from home. This is also demonstrated to improve results and has not been shown in previous work.

Classification of depression in patients with bipolar disorder using the geolocation-derived features is demonstrated in section 6.3 to achieve an accuracy of over 75 %.

Chapter 7 describes a new Bayesian non-parametric method for finding groups of individuals who have similar regression models linking their objective features and mental state. The application of this model to depression-estimation in the AMoSS cohort demonstrates for the first time the utility of finding sub-population models of mental state. This model also has applications to other fields where sub-groups may have different regression models. While previous work, e.g. by Servia-Rodríguez et al. (2017), demonstrated that demographics correlate with mood (see section 2.2.4.9 for more details) the model presented here can find groups without needing explicit demographic information.

The results in chapter 7 are also the first to compare the performance of personalised models cross-validated over all data available from an individual against using data just from the start of the recording for that individual. This is important because many studies on objective markers of mental illness use the former method which is demonstrated to give significantly improved results over the latter method, but does not generalise to how the model would perform in practice due to the limited

training data available. The personalised model is demonstrated to improve over fully personalised models trained only on calibration data from the test individual.

1.3.2 Implications

Some of the key challenges in the provision of mental healthcare services, and the potential impact of reliable objective measures of mental illness can be summarised in the following (closely related) areas.

Cost of delivery The scale in terms of patient numbers, and the associated financial costs, involved in mental healthcare have been outlined in the introduction to this chapter. Exacerbating the situation, a combination of factors including reduced stigma surrounding mental health; an ageing population; and improved handling of mental illness in the criminal justice system, mean that demand for mental health services is likely to rise (CQC, 2017). There is therefore clearly a pressure to reduce costs within the system while maintaining and improving the care that patents receive in the face of increased demand. One way to achieve this is to improve the allocation of resources in the system. Being able to monitor the mental state of patients continuously means that physicians can be alerted when patients are deteriorating and may be able to improve outcome and cost through early intervention.

Parity of care While there are currently significant differences in mental health service provision throughout the United Kingdom, both in terms of quality (CQC, 2017) and spend (Mental Health Taskforce, 2016), these differences pale into insignificance when considering the global disparity between available healthcare services (World Health Organization, 2011). Some of the contributing factors identified by the Mental Health Taskforce (2016) are poor reporting and transparency in performance. The ability to objectively monitor patient groups could enable identification of areas in need of improvement or additional resources.

Community care Throughout healthcare, but especially in mental health, there is an increasing drive to move patient care from institutional settings to the community (Ham et al., 2012). As above, the ability of healthcare providers to be able to monitor patients means that they can lead more regular lives in the community when well, while still receiving appropriate care when needed.

Validity of diagnosis Validity of the current diagnostic frameworks was described above. It is possible that the ability to find groups of similar patients within the population, using method such as the one presented in chapter 7 in the broader population may reveal distinct phenotypes of illness that can in turn inform diagnostic methods.

Choice of treatment options Related to validity of diagnosis, choosing the appropriate treatment for mental health patients can be difficult. It has been suggested that improved decision support tools may aid clinicians in choosing appropriate treatments (Kessler, 2018) and the groups of similar individuals identified using the model presented in chapter 7 may be useful to identify patients suited to specific treatments.

Objective clinical trials Running clinical trials with mental health patients is challenging because of the difficulty of quantifying the result of the intervention. Improved methods to quantify mental state using objective markers could transform the way that clinical trials are run.

Long-term management Finally, mental health patients with long-term conditions, especially bipolar disorder, commonly self-monitor their health using a variety of metrics (Murnane et al., 2016). The ability to provide accurate indicators of health, rather than the general metrics currently employed, could help patients better understand their illness, and potentially avoid serious episodes through early identification.

1.3.3 Structure

The remainder of this thesis is structured as follows.

The next two chapters give background and introduce related work in the fields of objective markers of mental health and machine learning respectively. Chapter 4 describes the data analysed in this thesis, and chapter 5 describes the extraction of relevant features from the recorded geolocation data. Chapter 6 describes the identification of mental health state and classification of mental health disorders using global models trained on the extracted geolocation features. Chapter 7 is concerned with the development of personalised modelling techniques and applies these models to mental health quantification from objective features. Finally chapter 8 concludes the thesis and gives an overview of future work.

→

1.4 Publication list

The following core publications have arisen out of the work in this thesis:

- Palmius, N., Osipov, M., Bilderbeck, A. C., Goodwin, G. M., Saunders, K., Tsanas, A., & Clifford, G. D. (2014). A multi-sensor monitoring system for objective mental health management in resource constrained environments. In *Appropriate Healthcare Technologies for Low Resource Settings (AHT 2014)*. doi: 10.1049/cp.2014.0764

  This conference paper presents an outline of the AMoSS study described in greater detail in chapter 4.

- Palmius, N., Tsanas, A., Saunders, K. E. A., Bilderbeck, A. C., Geddes, J. R., Goodwin, G. M., & De Vos, M. (2017). Detecting Bipolar Depression From Geographic Location Data. *IEEE Transactions on Biomedical Engineering*, *64*(8), 1761–1771. doi: 10.1109/TBME.2016.2611862

  This journal paper presents the geolocation derived features described in chapter 5 and the population-level depression classification model in chapter 6.

- Palmius, N., & De Vos, M. (2016). Non-Parametric Subset Selection for Personalised Time-Series Regression. In *Practical Bayesian Nonparametrics Workshop at the 30th Conference on Neural Information Processing Systems (NIPS 2016)*.

  This conference workshop paper presents a preliminary version of the personalised regression model in chapter 7.

- Palmius, N., Saunders, K. E. A., Carr, O., Geddes, J. R., Goodwin, G. M., & De Vos, M. (in press). Group-personalized regression models for prediction of mental health scores from objective mobile phone data streams. *Journal of Medical Internet Research*. doi: 10.2196/10194

  This journal paper presents full results of the personalised regression model in chapter 7 applied to the AMoSS participants.

Chapter 2

Mental Health Background and Related Work

The previous chapter outlined the growing interest in using technology to assist in mental health assessment, treatment, and monitoring. Related work in this area will be explored in greater detail in section 2.2, after an introduction to two important mental health conditions studied in this thesis.

2.1 Relevant mental health conditions

The techniques described in this thesis may be relevant to a wide range of mental health disorders, but the focus is on two specific conditions that will be briefly reviewed here.

2.1.1 Bipolar disorder

Bipolar disorder (BD), sometimes known as bipolar spectrum disorder (BPS) (Merikangas et al., 2011) and formerly known as manic-depressive illness (Healy, 2010), is mental disorder characterised by drastic mood shifts (Valenza et al., 2013) between states of depression and mania (defined below). BD is thought to be responsible for the loss of more disability-adjusted life-years (DALYs) than all forms of cancer or major neurological conditions such as epilepsy or Alzheimer's disease. This is because of the severity of illness in both extreme mood states combined with the typically early

onset and course that can persist through the whole life of a sufferer (Merikangas et al., 2011).

Bipolar disorder is one of the most studied forms of mental illnesses. This is partially due to the severity and persistence of the disorder (Kessler et al., 2007), but also because understanding the swings between mood states may provide opportunities to prevent relapse.

There are different types of bipolar disorder depending on the level of functional impairment in manic episodes (see below), but population surveys have revealed that between 4 % and 6 % of adults may be sufferers of any type (Merikangas et al., 2011). A significant percentage of cases that do not meet the criteria for BD are diagnosed instead as unipolar major depression (Merikangas et al., 2011).

2.1.1.1 Depressive episodes

A major depressive episode (MDE) is defined in the DSM 5 as a continuous period of at least 2 weeks with the following nine symptoms, of which five must be met for diagnosis (APA, 2013):

1. depressed mood most of the day;
2. diminished interest or pleasure in all or most activities;
3. significant unintentional weight loss or gain;
4. insomnia or sleeping too much;
5. agitation or psychomotor retardation noticed by others;
6. fatigue or loss of energy;
7. feelings of worthlessness or excessive guilt;
8. diminished ability to think or concentrate, or indecisiveness; and
9. recurrent thoughts of death.

Depressive episodes can occur in individuals either with or without manic episodes, but a history of at least one manic episode is critical for a diagnosis of BD. Individuals who have not also had a manic episodes, but have experienced at least two MDEs are instead diagnosed with major depressive disorder (MDD). While MDD is more

prevalent, BD has been shown to have higher persistence, severity, and impairment than MDD (Kessler et al., 2007).

2.1.1.2 Manic episodes

A manic episode is broadly defined in the DSM 5 as a continuous period of at least 1 week of abnormally and persistently elevated, expansive, or irritable mood with the following seven symptoms, of which three must be present to a significant degree for diagnosis (APA, 2013):

1. increased self-esteem or grandiosity;
2. decreased need for sleep (e.g. feels rested after only 3 hours of sleep);
3. more talkative than usual or pressure to keep talking;
4. flight of ideas or subjective experience that thoughts are racing.
5. distractibility (i.e. attention too easily drawn to unimportant or irrelevant external stimuli);
6. increase in goal-directed activity (either socially, at work or school, or sexually) or psychomotor agitation; and
7. excessive involvement in pleasurable activities that have a high potential for painful consequences.[1]

Mania is the defining feature of BD, and a single episode of mania is sufficient to diagnose an individual – a history of depressive episodes is not required.

2.1.1.3 Euthymia

Euthymia is the time when a BD patient is not in either a manic or depressive episode (Suppes & Dennehy, 2010). While in theory patients are healthy in these periods, in reality functional impairment resulting from sub-syndromal features of episodes and mood instability often persists (Strejilevich et al., 2013). Judd et al. (2003) found that BD patients are completely free from manic or depressive symptoms only around

[1] For example engaging in unrestrained buying sprees, sexual indiscretions, or foolish business investments.

half of the time on average. There is also strong evidence for cognitive impairment during euthymia (Robinson et al., 2006).

2.1.2 Borderline personality disorder

Borderline[2] personality disorder (BPD) is a common and severe disorder associated with high rates of self-harm,[3] self-mutilation and suicide.[4] BPD is characterised by a pervasive pattern of impulsivity and instability in emotions, interpersonal relationships, and identity, often associated with aggression and anger (Leichsenring et al., 2011). The DSM 5 has the following nine symptoms of BPD, of which five must be met for diagnosis (APA, 2013):

1. frantic efforts to avoid real or imagined abandonment;
2. a pattern of unstable and intense interpersonal relationship characterized by alternating between extremes of idealization and devaluation.
3. identity disturbance (e.g. persistently unstable self-image or sense of self);
4. impulsivity in at least two areas that are potentially self-damaging (e.g. spending, sex, substance abuse, reckless driving, binge eating);
5. recurrent suicidal behaviour, or threats, or self-mutilating behaviour;
6. affective instability due to a marked reactivity of mood (e.g. intense episodic dysphoria, irritability, or anxiety lasting from a few hours to a few days);
7. chronic feelings of emptiness;

[2] The "borderline" adjective in BPD is somewhat misleading and is the result of historical legacy from the original description of the disorder by Stern (1983) (Paris, 2005). Stern believed that patients with the disorder were on a "border" between psychosis and neurosis, based on the theory of psychoanalysis, popular at the time, where all mental disorders lie on a continuum. With the publication of the DSM-III in 1980, the theory of psychoanalysis fell out of favour (Mayes & Horwitz, 2005), but the "borderline" name persisted. While Paris (2005) believes that eventually the name will be changed following increased understanding of the psychopathology of the disorder, for the time being it is considered sufficient.

[3] Zanarini et al. (2006, 2013) reports that around 90 % of inpatients with a diagnosis of BPD have a history of self-harming, of which 60 % started self-harming in adolescence or earlier, initially out of frustration or anger, but over time the predominant reason becomes to relieve anxiety or to control emotional pain (Zanarini et al., 2013).

[4] Suicide rates among patients with BPD have been reported to be as high as 10 % (Stone, 1990, ch. 4; Paris et al., 1987; Paris & Zweig-Frank, 2001), with up to 70 % of patients attempting suicide. As would be expected, BPD comorbid with other serious disorders, such as MDD or substance abuse, increases the risk of suicide (Gunderson, 2008, ch.4; Kolla et al., 2008).

8. inappropriate, intense anger or difficulty controlling anger; and
9. transient, stress-related paranoid ideation or severe dissociative symptoms.

Prevalence of BPD is thought to be between 0.5 % and 5.9 % in the general US population and it forms 10 % of outpatient case-load and 15 % to 25 % of inpatient case-load (Leichsenring et al., 2011).

Both BD and BPD are characterised by drastic swings in mood. Note the similarities between symptom 6 in the list above with the manic episode in BD. However, the scale on which the swings occur is different, with episodes in BD lasting weeks or months, while periods of high or low mood in BPD can pass in days or hours. BPD is also commonly comorbid with other mood or personality disorders, which can present a more complex picture.

Unlike BD, which is classed as a mood disorder, BPD falls in the family of personality disorders. There are ten personality disorders listed in the DSM 5 (APA, 2013), of which BPD is the most prevalent (Leichsenring et al., 2011). Personality disorders are best considered from the perspective of the range of human personality traits that have evolved to handle the many different environmental and social situations faced. Everyone has an overall personality strategy comprised of different levels of these traits. The range of personality strategies has largely evolved through prehistory, but society has developed faster than our personality strategies can evolve. This means that society as a whole places implicit bounds on the personality traits considered to be "normal", and individuals with personality strategies containing traits outside of that range are labelled as having a personality disorder (Moran & Crawford, 2013; Beck, 2015).

It is therefore important to distinguish personality disorders from other mental illnesses in the strict meaning of the expression. While mental illnesses, such as MDD or episodes in BD, are periods when the individual is unwell,[5] a personality disorder is simply a reflection on the ordinary personality of the individual (Millon et al., 2004). This is a stable state for the individual, albeit outside social norms,

[5] Mood disorders can be seen as synonymous to a physiological illness, like a cough or a cold (see Millon et al., 2004, figure 1.3, p. 10).

unlike the transient states of illness in mood disorders. In other words, the personality disorder *is* the personality of the patient, rather than a mood episode that the patient temporarily suffers with.

The underlying cause of BPD is unknown, but there is evidence of genetic factors being involved. Studies have reported heritability of between 35 % and 67 % in various twin studies, and a meta-analysis by Amad et al. (2014) showed that there are several specific genes that are linked to the development of BPD. There also appear to be environmental factors involved in the pathogenesis of BPD in some cases. Symptoms often start to show in adolescence (Biskin, 2015), and in 90 % of inpatient cases studied by Zanarini et al. (1997), patients reported childhood abuse or neglect. Similar percentages are reported by other studies[6] (see Ball & Links, 2009; Gunderson, 2008, ch. 2). However, certain genes may make individuals more susceptible to developing BPD, but only in combination with childhood abuse or neglect, and conversely certain genes may protect individuals (Amad et al., 2014). These factors combined indicate a complex interaction between environmental and genetic factors, known as gene-environment interaction, interplay or correlation. Following from the evolutionary theory of the development of personality presented earlier, this can be viewed through the development of latent personality traits. In most individuals they remain hidden, but in response to specific environmental conditions or exposures they are revealed.[7]

Patients with borderline personality disorder are often seen as difficult to treat (Millon et al., 2004, p. 511), with Silver (1983) describing them as "characterologically difficult" patients (Leszcz, 1989). Personality disorders were traditionally seen as resistant to psychotherapy (Perry et al., 1999; Perry, 2001), which is understandable, given that the treatment is attempting to change the regular personality of the individual. With that said, if patients are aware of how their actions affect themselves and those around them, especially in less severe cases, therapy can be effective

[6] It should be noted that these studies of childhood abuse and neglect are retrospective and therefore may include an element of recall bias (Ball & Links, 2009).

[7] Gene-environment interactions have also been reported in a number of other serious mental illnesses, see Uher (2014) for a review.

(Millon et al., 2004, pp. 511–512). Comorbid mood disorders can also respond to treatment, but the effect may be reduced (Leichsenring et al., 2011).

Borderline personality disorder is widely researched (Moran & Crawford, 2013) and considered to be a promising candidate for early intervention and longitudinal monitoring of progression. Unlike many personality disorders that are relatively stable for the individual, BPD has a lifelong course with symptoms peaking in severity in late adolescence and gradually improving (Biskin, 2015). All symptoms do not improve equally, however, and affective symptoms tend to persist longer then impulsivity features and behavioural aspects (Gunderson et al., 2011; Zanarini et al., 2007). While many patients remiss completely to no longer fill the diagnostic criteria for BPD, Gunderson et al. (2011) found that 12 % of patients relapse. Being able to detect, or even predict, signs of this relapse could be very valuable. Additionally, the ability to monitor progression of symptom severity through therapy would be useful since there are several different therapy options available (Lieb et al., 2004).

2.2 Assessment of mental health

A core challenge of assessing mental health severity is that mental illnesses are, generally speaking, illnesses of the mind,[8] and the ability to make assessments of the mind is an open area of research. Current usual practice in the assessment of mental illness is a face-to-face verbal interview conducted by a psychiatrist or mental health professional (Millon et al., 2004, p. 124), sometimes supplemented by patient-reported questionnaires (D'Avanzato & Zimmerman, 2017).

Face-to-face interviews can take several forms. An unstructured interview, sometimes known as a clinical interview, is used to collect symptom information from the patient to support a diagnosis (Jones, 2010). Although this is the most commonly used assessment technique, the interview does not follow a standard structure, meaning that the interviewer decides how to structure the interview. The quality of the results from an unstructured interview are therefore highly dependent on the ability

[8] Some authors dispute this definition, e.g. Szasz (1960) who argues that mental illnesses are in fact not illnesses at all.

of the interviewer to determine the appropriate information needed to arrive at a diagnosis and/or assess the severity of illness in the patient. Structured interviews instead follow a prescribed interview format with a standardised sequence of questions (Segal & Williams, 2014). Structured interviews aim to reduce the dependence on the skill of the interviewer, and are widely considered to lead to more consistent results between interviewers (Barrio-Minton, 2017, p. 106). Semi-structured interviews are a compromise between the two previous methods, with a general systematic structure prescribed, but allowing the interviewer to deviate from the structure as deemed appropriate (Lucas, 2003).

Although generally considered to be the "gold standard" assessment method (Asmundson et al., 2014), the downsides of structured interviews are that they can take a long time to complete (Barrio-Minton, 2017, p. 106) and may reduce the rapport between the interviewer and interviewee (Segal & Williams, 2014).

All face-to-face interviews, even structured interviews, are subjective by their nature. This is leading to an increasing interest in finding objective measures of mental health that reduce the subjectivity of the assessment. These objective measures are referred to as "laboratory findings" in the DSM-IV (APA, 2000).

Another important aspect of mental health management is monitoring changes in the severity of illness in patients. Monitoring severity of illness can be important to check the efficacy of a treatment plan (Moran & Crawford, 2013), but is also useful in episodic disorders such as BD where early intervention at the start of an episode may reduce the severity of the episode (Morriss et al., 2007). Interviews, however, require dedicated time for each patient, which places an upper limit on the regularity at which patients can be assessed, and as discussed earlier are subjective in nature. There is therefore increasing interest in passive objective measures that can longitudinally monitor the health of the patient without (or with minimal) patent or clinician involvement.

When looking for new methods to assist the assessment of mental health, there are several areas being explored, as described in the following subsections. Patent-

reported metrics (section 2.2.1) can be easily collected using modern technology,[9] but are inherently subjective and somewhat cumbersome for the patient, especially when collected for extended periods. Biological markers (section 2.2.2) describe factors directly linked to the biology of the patient, such as genetic factors in mental illness, which may show useful risk factors for diagnosis. Finally physiological markers (section 2.2.3) are autonomous processes of the body, and behavioural markers (section 2.2.4) indicate changes to behaviour.

2.2.1 Patient-reported measures

Arguably the simplest way to gather long term information about the health (mental or otherwise) of a patient is to simply ask them to tell you. This is usually in the form of a questionnaire completed by the patient, which could be completed either on paper or electronically. There are two main types of patient-reported questionnaires commonly used.

The first type of questionnaire are clinically-validated psychiatric questionnaires (several examples of which are described in greater detail in section 4.4), which are structured questionnaires that are completed by the patient either on paper or digitally. These questionnaires are generally focussed on specific conditions, and ask the patient to rate different diagnostic features of the condition. The responses for the individual questions can usually be combined to give a single score indicating the severity of illness of the individual. These questionnaires can provide valuable assistance to healthcare professionals, and can be used by patients for self-monitoring their condition. While many of these questionnaires have been shown to correlate well with clinical diagnosis, they are subject to recall bias because they often relate to the past one or two weeks.

A different type of questionnaire, sometimes known as ecological momentary assessment (EMA) questionnaires (Saeb et al., 2015b; Wang et al., 2014), attempt to

[9] Prior to the widespread adoption of Internet-enabled smartphones, studies were performed that required users to send reports via text messages (Bopp et al., 2010) or that used a simple applet to collect responses to send via text message (Reid et al., 2009, 2011).

capture simpler data about the individual, often at a higher frequency. Commonly these are focussed on mood, but may ask simple questions about what the individual is currently doing. These questions can take several forms, such as asking patients to rate their mood on an appropriate scale, to choosing descriptive words for their mood or current activity. Saeb et al. (2015b) found a weak relationship between EMA scores and PHQ-9 depression questionnaire scores (similar to the QIDS questionnaire described in section 4.4.1.1) but found that some questions were more predictive than others. Unsurprisingly, EMA questionnaires completed nearest to the depression questionnaire were found to be most strongly correlated.

Both types of questionnaire have a number of disadvantages, however, including the subjectivity of the answers and the different baselines that different individuals have. They may also be a burden on the patient to perform, especially when they are required for extended periods, which may lead to missing data or complete withdrawal from providing the data, especially when the incentive for the patient is unclear (Drake et al., 2013). As part of the StudentLife study (described below), Wang et al. (2014) found a relatively steady drop-off in the number of EMA responses over the 10 weeks of the study, although the number of daily prompts was higher than would be reasonable in practice (up to 13 each day). Wang et al. also found that that participants preferred shorter, quicker, questionnaire styles.

There is colloquial evidence that self-monitoring can improve symptoms of depression (Groot, 2010; Wikberg et al., 2016) but no significant effect was found in a randomised control trial by Wikberg et al. (2017). Wikberg et al. (2017) did however find a significant improvement in adherence to antidepressants in patients who were self-monitoring their levels of depression.

2.2.2 Biological markers of mental health

Moving from subjective patient-reported indicators, the most fundamental form of objective assessment is to consider core biological features, for example genetic and cellular markers.

Genetic factors are known to be involved in many mental conditions (Kendler, 2013), and epigenetic influences on gene expression are also thought to be important (Nestler et al., 2016). While genetics may provide useful risk factors for diagnosis and personalisation of treatment options, unless the biological mechanism by which the genes influence the mind is understood (Kendler, 2013), these genetic markers themselves are not useful for long term monitoring.

It is possible to make structural and functional assessments of the brain using a variety of techniques,[10] but while there has been progress in identifying features that correlate with mental illness (Lewis & Higgins, 1996), these methods are often too impractical to use for longitudinal monitoring.

The methods described above provide a number of tools that may be useful, either in isolation or in tandem, to assist screening or diagnosis of various mental health disorders. This could have important implications, especially where it might be possible to detect signs of an illness before symptoms present.

2.2.3 Physiological markers of mental health

While biological aspects of the body tend to change slowly in response to environmental stimuli, physiology covers the structure and function of organs, which need to respond more quickly. Many physical parts of the body are affected by mental health.

2.2.3.1 Photoplethysmography

Photoplethysmography (PPG) non-invasively measures the blood volume changes in the microvascular bed of tissue using a light source and light detector (Allen, 2007).

Hu et al. (2011) and Pham et al. (2013) applied a variety of nonlinear complexity measures to PPG data recorded from 195 patients diagnosed with a variety of mental illnesses and 113 students. Pham et al. achieved sensitivity and specificity of 99 % differentiating the patient group from the control group using a k-NN classifier.

[10] Brain structure and/or function can be measured using techniques such as computed tomography (CT); magnetic resonance imaging (MRI); magnetic resonance spectroscopy (MRS); single photon emission computed tomography (SPECT); positron emission tomography (PET); functional MRI (fMRI); magnetoencephalography (MEG); or electroencephalogram (EEG) (Turkington, 2002).

2.2.3.2 Voice

Voice, speech and language characteristics have long been considered as important descriptive features for episodes in bipolar disorder (Kraepelin, 1921), and other mental illnesses (see Oyebode, 2015, ch. 10). Speech is also an aspect of human behaviour that is used daily by most people, making it a highly accessible physiological signal for analysis.[11] Mobile phones provide a unique way to analyse voice in a natural setting, when users are engaged in normal telephone conversations. While the amount of voice data available from phone conversations is likely to be considerably less (for most individuals) than from face-to-face conversations, the user speaks directly to the phone so identification of the patient is simpler, making analysis easier.

Several objective experiments on small samples of patients started in the late 1970s analysing recordings of patients reading identical passages of text. Darby & Hollien (1977) tested several properties of recorded speech from 6 severely depressed patients before and after electroconvulsant therapy (which appeared to reduce or eliminate symptoms) but did not find any meaningful correlations. A similar study by Askenfelt & Sjölin-Nilsonne (1980) comparing 4 depressed patients and 4 healthy controls also found limited differences between the two groups. In a more detailed analysis of 28 patients with acute major depression and 13 healthy controls, Nilsonne et al. (1988) found significant differences in the variability of the fundamental frequency of speech between the two groups. Ellgring & Scherer (1996) reported changes in simple voice features in a longitudinal study of patients during depression and in remission. Speech rate was found to be be the most strongly correlated with well-being, but only significantly in female participants (although this could be in part due to the small sample size of male participants).

Nilsonne et al. (1988) points out that in early work, fewer variables could be considered in each experiment given that specialised equipment may be required. In modern work, thousands of speech-related variables can be calculated in real-time

[11] While speech itself is more behavioural than physiological, some features of voice such as unconscious variations in tone can be considered physiological.

on a mobile device during a phone call. Such a technique was used by Faurholt-Jepsen et al. (2016a) to analyse the mental state of 28 BD participants as part of the MONARCA study (described below). 6552 features were calculated during phone calls using the openSMILE library (Eyben et al., 2010), which were used to classify episode type based on fortnightly clinician ratings. Classifying depressed episodes from euthymic state had an accuracy of 0.68 and classifying manic or mixed episodes from euthymic state had an accuracy of 0.74.

Muaremi et al. (2014) and Grünerbl et al. (2015) used a similar but reduced set of features generated by the openSMILE library on a separate cohort of 6 BD patients from the MONARCA project, but trained personalised models on individual patients. Classification was performed on a daily basis, and each model was trained on 66 % of the data from the participant and tested on the remaining 33 % of data. Muaremi et al. demonstrated an average F_1 score across all participants of 0.80 for detecting the clinically-scored state of the participant using a Random Forest classifier. Muaremi et al. also included features of conversation characteristics, adapted from Feese et al. (2011), such as the number of changes between speakers and how long each speaker spoke for, as well as statistics of all phone calls made and received. Including these features in the model improved the average F_1 score across all participants to 0.83, but the individualised nature of these results make them impractical for real-world use. The reasons for this will be discussed in detail in chapter 7.

2.2.4 Behavioural indicators of mental health

Physiological markers show promise as objective markers for mental health, but can be cumbersome to measure, especially continuously, and sometimes require specialised (although increasingly portable) equipment.[12] A further form of assessment is to consider the overall behaviour of the individual and how this changes with changes in

[12] Botella et al. (2016) tested the acceptance of a computerized cognitive behavioural therapy software program with and without physiological sensors, and found that the sensors did not significantly affect user satisfaction, however the results may have been affected by the additional reimbursement provided to the group with the sensors. Botella et al. (2016) also did not demonstrate that the use of sensors improved the outcome of the program, but it is unclear how they were used.

severity of illness. Many mental illnesses are characterised by changes in behaviour, for example the reduced activity typically seen in depressive episodes to the high and erratic behaviour reported in manic episodes in BD. One of the major advantages with using behavioural markers is that many can be measured using smartphones and smartwatches, making them potentially very unobtrusive methods of continuous assessment.

The development of objective behavioural indicators of mental health is not new, but has traditionally been restricted by the difficulties of collecting and analysing longitudinal data from patients over long periods of time. As discussed briefly in chapter 1, modern smartphones are powerful and highly pervasive devices that often include a variety of sensors that can measure features that are thought to relate to mental illness. This has enabled large amounts to data to be collected but analysis requires specialised techniques due to the amount and type of data.

The remaining sections of this chapter review some of the current work in objective behavioural indicators of mental health categorised by the modality analysed. Results on individual modalities will be given where available, and results from studies that only present combined results using multiple modalities will be presented in section 2.2.4.9.

2.2.4.1 Major projects

Many of the results on behavioural markers for mental health have come out of a few major projects that will be briefly introduced here and referred to later.

MONARCA The MONARCA[13] project (Riva et al., 2011; Puiatti et al., 2011) developed a mobile and web-based platform to monitor objective markers of mental health in patients with BD, either in the community or in a hospitalised setting. The MONARCA project is now being commercially developed as the Monsenso mHealth solution.[14]

[13] MONitoring, treAtment and pRediCtion of bipolAr episodes (Faurholt-Jepsen et al., 2013).

[14] See https://www.monsenso.com/ for details of the commercial solution.

StudentLife The StudentLife study (Wang et al., 2014) recorded wide-ranging smartphone-based data from a cohort of 48 students enrolled in a smartphone programming course at Dartmouth College (30 undergraduates and 18 graduate students) for a full 10-week term. Students completed several questionnaires assessing their mental health at the start and end of the study, including the 9-item Patient Health Questionnaire-9 (PHQ-9) depression questionnaire by Kroenke et al. (2001). EMA questionnaires relating to mood, stress and sleep were also collected at several times during the day. Mood was collected using a photographic affect meter (PAM) questionnaire that asks the user to select an image that most closely resembles their current mood from a set of pre-defined images (Pollak et al., 2011). Much of the data from the StudentLife study has been publicly released.[15]

2.2.4.2 Physical activity levels

One of the most studied behavioural markers of mental health is the physical activity level of individuals, collected either from a smartphone accelerometer or a standalone actigraphy device. This is not surprising given the diagnostic symptoms of depression given in section 2.1.1.1, many of which could result in reduced overall activity levels; or the diagnostic symptoms of mania given section 2.1.1.2, many of which may result in increased or erratic overall activity levels.

Early work on the relationship between depression and physical activity levels relied on self-reported activity levels (e.g. Goodwin, 2003; Galper et al., 2006; De Mello et al., 2013). As part of the MONARCA study, Faurholt-Jepsen et al. (2015) found a weak but significant correlation between daily self-reported activity levels and clinician-rated levels of both depression and mania in 30 BD patients. While studies using self-reported measures are relatively simple to run on large cohorts and demonstrated the hypothesised trends, self-reported activity has been shown to have poor validity as a proxy for direct measures (S. A. Prince et al., 2008).

[15] The StudentLife dataset is available from http://studentlife.cs.dartmouth.edu/dataset/, but some potentially personally-identifiable data were redacted.

Early studies on objective measures of activity, such as the OPTIMI project (Botella et al., 2011), used custom hardware to measure activity levels,[16] but later work uses off-the-shelf wearable devices or smartphones. The Help4Mood project (J. C. Castro et al., 2011, 2012; de Cerio et al., 2012, 2013), for example, used a network of activity sensors, including a smartphone; wristwatch; belt sensor; keyring sensor; and under-mattress sensor, where the user chose the devices that they were comfortable with (Fuster-Garcia et al., 2015). However, no results on patients have been presented.

In a comparison of unipolar patients; bipolar patients; and healthy controls, Faurholt-Jepsen et al. (2012) found that the unipolar patients had significantly lower levels of overall fitness than the bipolar patients or healthy controls, indicating that general inactivity may be a factor in unipolar depression. Comparing unipolar and bipolar patients, significant differences were only found when patients were not reporting depressive symptomatology, where it was found that bipolar patients had up to 63 % lower activity levels depending on the time of day. Wielopolski et al. (2015) found significantly lower levels of physical activity in acute unipolar depressed patients compared to healthy controls, and found that physical activity correlated with improvement in health in the 19 bipolar patients.

As part of the StudentLife study, Wang et al. (2014) found that levels of activity were weakly significantly negatively correlated with self-rated loneliness scores, but not with depression. Farhan et al. (2016a) also did not find significant correlations between simple activity/inactivity features and self-reported PHQ-9 depression scores from 79 students recorded using either Android and iPhone devices.

2.2.4.3 Geographic movements and locations

While physical activity levels can be considered as a local measure of activity, geographic movements may provide a similar indicator over a wider scale.

One of the first studies of the relationship between objective measures of geolocation and mental health was by Gruenerbl et al. (2014), and later Grünerbl et al.

[16] The OPTIMI project has not published results on objective activity data.

(2015), as part of the MONARCA project. Gruenerbl et al. studied 10 patients with bipolar disorder who were hospitalised at the start of the study period. Patients were initially resident at the hospital, but may have been discharged during the data collection period. Even while resident, patients were able to move around the hospital and nearby town. Patients were clinically assessed every three weeks through the 12 week duration of data collection, with additional intermediate telephone assessments. Simple daily features were extracted from the geolocation data focussing on the amount of time spent outdoors (although it is not clear how this was determined) but also including the number of locations visited and the total distance covered. Separate Naïve Bayes models were trained per individual and tested using three-fold cross-validation repeated 500 times. This resulted in a mean 81 % of days correctly classified across all the patients, but given the limited number of states exhibited by individual patients (usually only 2) and likely high intra-patient correlations of features, these results may not generalise well. It does however demonstrate the potential for geolocation as a feature of bipolar episodes.

In a community cohort of 18 individuals recruited over the internet, Saeb et al. (2015a) studied the relationship between geolocation-derived features and depression as measured using the self-reported PHQ-9 questionnaire. Saeb et al. found significant correlations of certain geolocation-derived features onto PHQ-9 scores and reported a classification accuracy of 86.5 % for detecting mild depression. The most significantly correlated geolocation-derived features included the normalised entropy of time spent in different locations (see section 5.10.2); the level of diurnal regularity (see section 5.10.8); the variance in location coordinates; and the percentage of time spent at home. Using the same data, Saeb et al. (2015b) separately reported that significant correlations between several geolocation-derived features and PHQ-9 scores were only found when the features were calculated over two weeks of geolocation data, and calculating the features on individual days did not result in significant correlations. A similar methodology applied to 48 university students by Saeb et al. (2016) found comparable results. The participants in both of these studies were not diagnosed with any mental disorder and as such the generalisability of the findings

to clinical populations is unclear. Many of the features used by Saeb et al. (2015a, 2015b, 2016) are described in detail in section 5.10.

Farhan et al. (2016a) calculated similar geolocation-derived features from 79 students using both Android and iPhone devices. Similar to Saeb et al. (2015a), Farhan et al. found significant correlations between the normalised entropy of time spent in different locations and the percentage of time spent at home with self-reported PHQ-9 scores, but only on data recorded from Android devices. Notably the strength of the correlations reported by Farhan et al. was lower than those reported by Saeb et al. (2015a) and more closely matching the results reported in section B.1 of appendix B. On data recorded from iPhones, significant correlations were only found for the number of locations visited, which were not found to be significant from Android devices. Combining the geolocation-derived features and simple activity/inactivity features, Farhan et al. reported classification of depression in their cohort ($n = 19$ of 79) with sensitivity of 0.73/0.83 and specificity of 0.97/0.92 on iOS/Android devices respectively. It should be noted, however, that diagnosis of depression was made at initiation of the study and all depressed participants underwent treatment during the course of data collection. Some PHQ-9 scores from the depressed individuals were therefore below the threshold for depression, but they were still treated as depressed for the purposes of classification.

Yue et al. (2017) extended the results by Farhan et al. (2016a) by improving the data quality of geolocation data collected from the Android devices in the study. Yue et al. found that there were major limitations to the quality of geolocation data recorded on the Android devices (a problem described in greater detail in section 5.5) and developed a fusion algorithm to fill in missing data using wireless network access point associations and activity levels. Before data fusion, only 30 % of two-week periods analysed had 80 % data coverage, while after data fusion this increased to 54 %. This improved the reliability of the extracted features, which resulted in increased significance of the correlations between PHQ-9 scores and the number of locations visited; the variance in location coordinates recorded; the entropy and normalised entropy of time spent in different locations; and the percentage of time spent at home.

It also resulted in improved specificity of classification of depression to 0.96, although the same limitations as above still apply, and sensitivity decreased to 0.73.

Canzian & Musolesi (2015) reported correlations between geolocation-derived features and daily self-reported PHQ-8 depression scores[17] in 28 participants who provided at least 20 usable data points. Correlations were calculated for each participant individually, and significant correlations were shown between depression scores and the maximum distance between any two locations over the last 14 days for 18 of the 28 participants.

Mehrotra & Musolesi (2017) considered geolocation-based features correlated with self-reported emotional states in 22 participants who provided at least 21 days of labelled data. In addition to standard features, Mehrotra & Musolesi considered the similarity of the sequences of places visited on consecutive days. Instead of considering the variance of places visited as used by Saeb et al. (2015a), Mehrotra & Musolesi calculated the area of the smallest convex polygon that contains all the places visited. In total, 8 features were calculated, and correlations were strongest between differences of consecutive sequences of places and happiness ratings, although the features were optimised to maximise correlations. The area of places visited was strongly correlated with happiness ratings on weekdays. Significance levels were not reported.

Another way to consider geographic locations is to broadly identify the places where an individual spends time, for example, home; work; shop; restaurant; etc. Saeb et al. (2017) found that certain semantic locations were weakly significant predictors of self-reported depression and anxiety levels in a community sample of 206 adults without a history of mental illness.

Proxies for geographic movements can be obtained by monitoring the number of different cell phone base stations that the user connects to (Al Agha et al., 2016) or the number of wireless networks available to the user during the day. While both of these methods are imprecise, they are simple methods that indicate how much the user is moving between different locations and have lower power requirements than

[17] The PHQ-8 questionnaire is similar to the PHQ-9 questionnaire by Kroenke et al. (2001) but omits a question on suicide ideation (see Kroenke et al., 2009, for details).

recording GPS coordinates. Faurholt-Jepsen et al. (2014) found weakly significant correlations between the number of different cell phone base stations connected and clinician-rated levels of depression (but not mania) in a community sample of 17 BD patients as part of the MONARCA project. In a follow-up study on 17 different BD patients, Faurholt-Jepsen et al. (2016b) found weakly significant correlations between the number of cell phone base stations connected and clinician-rated levels of both depression and mania.

The second diagnostic symptom of depression describes *diminished interest or pleasure in all or most activities* which can manifest itself in patients spending a lot of time at home (see symptoms of MDE in section 2.1.1.1). Staying at home has also been shown by Adams et al. (2004) to be a risk factor for depression in elderly populations. One feature commonly investigated in relation to depression is therefore the amount of time spent at home. Chow et al. (2017) studied GPS traces of college students and found a significant positive correlation between the amount of time spent at home and negative affect, but not with the self-reported depression questionnaire used. Sabatelli et al. (2014) used connected wireless access points to determine the time spent at home, but did not find consistent individualised correlations with patient-reported depression on four BD participants. Petersen et al. (2014) also found that time spent at home correlates significantly with loneliness in elderly adults, which is in turn linked to depression (Aylaz et al., 2012) and correlates with poor outcome (van Beljouw et al., 2010). As part of the StudentLife study, Wang et al. (2014) calculated a simple metric of total distance travelled during each day, and found that this was weakly significantly negatively correlated with self-rated loneliness, but not depression.

2.2.4.4 Sleep patterns

Sleep is widely considered to be an important factor in the progression of episodes in unipolar depression and bipolar disorder (Franzen & Buysse, 2008).

As part of the StudentLife study, sleep was classified by taking into consideration ambient light and sound levels, activity levels, and phone usage. Wang et al.

(2014) found that sleep duration was significantly negatively correlated with self-rated depression scores and stress levels.

2.2.4.5 Circadian / diurnal regularity

Regularity in daily routine has long been indicated as a factor in depression (Ehlers et al., 1988) and bipolar disorder (Malkoff-Schwartz et al., 1998, 2000) where circadian rhythm disruption is considered by some to be a "core feature" (Gonzalez, 2014). Keeping to regular daily routines has been suggested as both prevention (Frank et al., 2006) and therapy (Frank et al., 1997, 2000). This link has been shown from adolescents (Wolfson & Carskadon, 1998) to the elderly (Prigerson et al., 1995). Increased lifestyle regularity has also been shown to improve sleep quality (Monk et al., 2003; Carney et al., 2006) which in turn impacts levels of depression (Franzen & Buysse, 2008). Carney et al. (2006) also indicated that social interaction, another indicator of positive mental health, may improve regularity and hence sleep. It is clear that the factors of social interaction, lifestyle regularity and sleep are all deeply interrelated and correlated with mental health. The causality here is not clear, but Malkoff-Schwartz et al. (1998, 2000) have shown that disrupted social routine may precipitate mania, and Soreca et al. (2009) argue that rhythm irregularity might even be a marker of vulnerability to mood disorders.

Saeb et al. (2015a) developed a feature describing the level of diurnal regularity calculated from geolocation traces, which is described in more detail in section 5.10.8. This was found to be significantly negatively correlated with self-reported levels of depression.

2.2.4.6 Use of technology

Given the pervasive nature of technology in modern society, especially mobile phones, the amount that someone interacts with that technology may be a marker of mental health.

Saeb et al. (2015a) studied the frequency of phone usage and the total time spent interacting with the phone in a community cohort of 21 individuals recruited over the

internet. In this study, Saeb et al. found that both phone usage features were weakly significantly correlated with levels of depression reported using the PHQ-9 questionnaire, with more depressed individuals interacting more with their phones. Saeb et al. also reported classification accuracy of 74.2 % for detecting mild depression from the amount of time spent using the phone. Faurholt-Jepsen et al. (2014), however, did not find significant correlations between the amount of time that the screen was on and clinician-reported levels of mania or depression in 17 BD patients. The same result was found in a follow-up study of 29 BD patients by Faurholt-Jepsen et al. (2016b).

The finding by Saeb et al. (2015a) is contradicted by Matthews et al. (2017) in a qualitative study of BD patients, who found that although participants reported increased use of technology during and prior to manic episodes, in depressive episodes they reduced their use use of technology significantly. However, Matthews et al. also report that perceived negative comparisons to others on social media may precipitate or exacerbate depressive symptoms.

Alvarez-Lozano et al. (2014) considered correlations between the type of smartphone apps used and self-reported mood scores in 18 BD patients as part of the MONARCA study. Analysis was performed per participant, and while significant correlations were found between app categories and self-reported metrics, no overall trends were presented. Turning this around, Stachl et al. (2017) reported that app usage may be predicted by personality trait, but their sample size of 137 participants was small given the number of variables being investigated.

Mehrotra et al. (2016) calculated several features of app and phone usage, but also considered how the user responded to notifications from apps, such as the number of notifications responded to and the time taken to respond. Mehrotra et al. found strong correlations between all of the features calculated over 14 days and PHQ-8 scores on 25 individuals, but the levels of depression found were not disclosed.

2.2.4.7 Social activity

Mobile phones give a unique source of objective data on social activity from the levels of calls and text messages.

In an initial study of 17 BD patients as part of the MONARCA project, Faurholt-Jepsen et al. (2014) found weakly significant correlations between the number of outgoing calls and clinician-rated levels of depression (but not mania); but did not find significant correlations for the number of outgoing text messages. However, in a follow-up study of 61 BD patients, Faurholt-Jepsen et al. (2015) found significant positive correlations of the number of outgoing calls with the clinician-rated levels of mania (but not with depression); and positive correlations of the duration of calls with clinician-rated levels of depression (but not with mania). Weakly significant correlations were also found with the duration of incoming calls (correlated with both depression and mania) and the duration of outgoing calls (correlated with depression only). Outgoing or incoming text messages were not significantly correlated with either depression or mania. In a further follow-up study on subsets of BD patients, Faurholt-Jepsen et al. (2016b) found the following results: weakly positive correlations between the number of outgoing text messages and mania (but not depression); no significant correlations with the duration of calls, the number of outgoing calls, or the number of characters in outgoing text messages; significant negative correlations between the number of characters in incoming text messages and mania; and weakly significant positive correlations between depression and the number of incoming calls and the number of missed calls. The final correlation (depression and the number of incoming calls and the number of missed calls) can possibly be explained by friends and family knowing that the patient is unwell and calling to check on them. Petersen et al. (2016) found conflicting results when studying loneliness in the elderly, with loneliness levels significantly negatively correlated with the number of incoming phone calls (but not outgoing calls). Taken together, the results from these studies are clearly conflicting in many aspects.

In the StudentLife study, social interaction was determined automatically by monitoring the microphone on the phone, and classifying speech and conversation. Wang et al. (2014) found that conversation frequency and duration were significantly negatively correlated with self-rated depression scores and stress levels.

Closely related to use of technology in the previous subsection, use of social media may also be predictive of mental health state. In a qualitative study, Matthews et al. (2017) found that BD patients reported excessive use of social media during manic episodes and reduced use during depressive episodes. Patients also reported differences in the way that they used social media during mania, such as sending uncharacteristically explicit messages or photos, or revealing sensitive information.

Tsakalidis et al. (2016) extracted semantic data from text messages and social network posts (from Facebook and Twitter) of students. Tsakalidis et al. found that ngrams, word embeddings, and topics (features used in sentiment analysis) were the most predictive, and achieved reasonable prediction of positive affect and well-being, with correlations between predicted and self-reported values of 0.84 and 0.87 respectively.

2.2.4.8 Other behavioural and lifestyle manifestations

Stress may be another factor in depression (Iacovides et al., 2003), and Wahle et al. (2016) used the number of calender entries as a proxy for stress but only presented combined results with other features (see below for details).

Excessive risk-taking is a key diagnostic feature of manic episodes in particular (see item 7 in the list of diagnostic features in section 2.1.1.2). This may take the form of increased use of online shopping or gambling (Matthews et al., 2017).

2.2.4.9 Combined results

Many studies, including some mentioned above, combine several modalities in a single model of mental health.

Wahle et al. (2016) used a combination of features derived from the metrics discussed previously for 36 participants trialling a new app for managing depression.

Because the features were also fed back to the participants with recommendations to take action, simple and understandable features were used such as general activity; walking time; location variance; and the number of phone calls made and received. Classification was also attempted using these features to estimate PHQ-9 scores, with an accuracy of 61.5 % achieved using a Random Forest classifier.

In a large study of affect and sensor data collected from 18 000 online participants, Servia-Rodríguez et al. (2017) reports the relationship between objective data and mood broken down into two dimensions of (a) positivity/negativity; and (b) alertness. A range of objective data were collected, such as magnitude of acceleration as a measure of physical activity; calls made and received; text messages sent and received; and ambient noise detected on the microphone. The number of locations visited was also recorded as well as the amount of time spent at each location, with an attempt to label places through user reports. Using a deep neural network to classify mood from the objective data, Servia-Rodríguez et al. found that classifying positive and negative mood achieved the highest accuracy of ≈68 % at weekends, and ≈62.5 % on weekdays. Due to the quantity of data collected, Servia-Rodríguez et al. were also able to analyse the correlations between participant demographics and the objective data, finding that almost all the demographic data about the participants were correlated with their objective data.

2.2.4.10 Limitations

Many of the studies described above use smartphones as the principal data collection method. This has many advantages as discussed, but a limitation is that data collection can only start once the relevant app is installed by the user. Analysis of the use of certain social networks can provide historical data with the consent of the user. Matic & Oliver (2016) propose using logs from mobile network providers of access locations; calls; and data usage, but this has major privacy considerations. The utility of historical data without labelling of mental state may also be limited, but in the case of patients with a history of clinical mental illness it may be possible to incorporate clinical records.

2.2.5 Discussion

This chapter has first introduced two relatively prevalent mental health conditions that are thought to be promising candidates for the application of new technology. A background to current methods for assessment of mental health was given, followed by survey of related work in methods of longitudinal and objective assessment of mental health. This ranges from simple questionnaires to biological, physiological and behavioural markers of mental health.

It is clear that there is significant interest in this area, and there are many potential areas to explore. Using smartphones or wearable devices is becoming increasingly popular due to their ubiquity. The next chapter provides a summary of machine learning techniques used in the rest of this thesis, then chapter 4 describes the collection of geographic location data for analysis.

Chapter 3

Machine Learning Background and Related Work

The field of machine learning has many tools available that can be leveraged to analyse data. The term *machine learning* was first described by Arthur Samuel (1959) as

> *the programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning.*

Samuel studied machine learning on the problem of training machines to play the game of checkers, but remarked, quite rightly, that "the principles of machine learning verified by these experiments are, of course, applicable to many other situations". Since then, machine learning has grown into a thriving academic field and commercial industry (Faggella, 2016) with many different branches applicable to a wide range of applications. It is interesting to note that the idea of Samuel in using games as a test-bed for new techniques is still very relevant today (Silver et al., 2016). This chapter introduces some of the techniques used in the rest of this thesis.

→

3.1 Standard models

3.1.1 Linear regression

The most basic method for estimating a continuous value $\hat{y}$ from feature data is *linear regression* using the model

$$\hat{y} = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_p z_p \tag{3.1}$$

$$= \beta_0 + \sum_{j=1}^{p} \beta_j z_j \tag{3.2}$$

$$= \beta_0 + \mathbf{z}^\top \boldsymbol{\beta} \tag{3.3}$$

where $\mathbf{z} = [z_1, \ldots, z_p]^\top$ is a vector of the p feature values; $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top$ is a vector of feature weights; and β_0 is the intercept (Hastie et al., 2009). It is often computationally helpful to include β_0 in the $\boldsymbol{\beta}$ vector by denoting

$$\boldsymbol{\beta}^* = [\beta_0, \beta_1, \ldots, \beta_p]^\top; \text{ and} \tag{3.4a}$$

$$\mathbf{z}^* = [1, z_1, \ldots, z_p]^\top \tag{3.4b}$$

in which case equation (3.3) can be written as

$$\hat{y} = \mathbf{z}^{*\top} \boldsymbol{\beta}^*. \tag{3.5}$$

The values of β_0 and $\boldsymbol{\beta}$ can be optimised by minimising the residual sum of squares

$$\text{RSS}(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{3.6}$$

$$= \sum_{i=1}^{N} \left(y_i - \beta_0 - \mathbf{z}_i^\top \boldsymbol{\beta} \right)^2 \tag{3.7}$$

$$\text{RSS}(\boldsymbol{\beta}^*) = \sum_{i=1}^{N} \left(y_i - \mathbf{z}_i^{*\top} \boldsymbol{\beta}^* \right)^2 \tag{3.8}$$

over N training samples. In the case where $N > p$ and the combined feature matrix $\mathbf{Z}^* = [\mathbf{z}_1^*, \ldots, \mathbf{z}_N^*]^\top$ is full rank (i.e. all features are linearly independent) equation (3.8) can be written as

$$\text{RSS}(\boldsymbol{\beta}^*) = \left(\mathbf{y} - \mathbf{Z}^* \boldsymbol{\beta}^* \right)^\top \left(\mathbf{y} - \mathbf{Z}^* \boldsymbol{\beta}^* \right) \tag{3.9}$$

where $\mathbf{y} = [y_1, \ldots, y_N]^\top$. By differentiating, equation (3.9) can be minimised algebraically as

$$\boldsymbol{\beta}^* = \left(\mathbf{Z}^{*\top}\mathbf{Z}^*\right)^{-1}\mathbf{Z}^{*\top}\mathbf{y} \tag{3.10}$$

and the values of β_0 and $\boldsymbol{\beta}$ can be extracted from $\boldsymbol{\beta}^*$ according to equation (3.4a).

3.1.1.1 Higher order components

Higher order components fit naturally into the framework above by considering equations (3.1) and (3.2) as

$$\hat{y} = \beta_0 + \beta_1 z_1 + \ldots + \beta_p z_p + \beta_1' z_1^2 + \ldots + \beta_p' z_p^2 + \beta_1'' z_1^3 + \ldots + \beta_p'' z_p^3 + \ldots \tag{3.11}$$

$$= \beta_0 + \sum_{j=1}^{p} \beta_j z_j + \sum_{j=1}^{p} \beta_j' z_j^2 + \sum_{j=1}^{p} \beta_j'' z_j^3 + \ldots \tag{3.12}$$

which, if $\boldsymbol{\beta}^*$ and $\mathbf{z}^*$ in equations (3.4a) and (3.4b) are re-defined as

$$\boldsymbol{\beta}^* = [\beta_0, \beta_1, \ldots, \beta_p, \beta_1', \ldots, \beta_p', \beta_1'', \ldots, \beta_p'', \ldots]^\top; \text{ and} \tag{3.13a}$$

$$\mathbf{z}^* = [1, z_1, \ldots, z_p, z_1^2, \ldots, z_p^2, z_1^3, \ldots, z_p^3, \ldots]^\top \tag{3.13b}$$

can still be solved with equation (3.10) if $N \geq 1 + dp$ where d is the highest power included in the model. The higher model complexity may however lead to over-fitting of the training data.

⟶

3.1.2 Generalised linear models

The ordinary linear regression of equation (3.3) can also be expressed as

$$y \sim \mathcal{N}(\beta_0 + \mathbf{z}^\top \boldsymbol{\beta}, \sigma^2) \tag{3.14}$$

or, equivalently

$$y = \beta_0 + \mathbf{z}^\top \boldsymbol{\beta} + \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{3.15}$$

where the output is now y directly, instead of $\hat{y}$. This form gives a stronger intuition into the interpretation of linear regression. It is assumed that the mean of the response variable y can be estimated as an additive linear weighted combination of the predictors $\mathbf{z}$ and offset β_0, with Gaussian distributed noise with variance σ^2. The optimal values of $\boldsymbol{\beta}$ and β_0 (usually obtained by minimising the residual sum of squares) are also the maximum likelihood estimation of equation (3.14) given the data $\mathbf{z}$ (see also section 3.2.2.2). Even if higher order terms are incorporated as discussed in section 3.1.1.1 this may not be a reasonable assumption.

Given knowledge of, or an intuition of, the distribution of y, a transform can be applied to the output of the linear regression to change the output distribution. These models belong to a class of non-linear extensions of the ordinary linear model known as generalised linear models (GLMs) (McCullagh, 1984; McCullagh & Nelder, 1989; Nelder & Wedderburn, 1972). More formally, the model is defined as an output random variable Y having an expectation

$$\mathbb{E}(Y) = \mu; \tag{3.16}$$

a linear regression

$$\eta = \beta_0 + \mathbf{z}^\top \boldsymbol{\beta}; \tag{3.17}$$

and a *link function*

$$\mu = g(\eta) \tag{3.18}$$

that links η to Y through μ. The output variable Y can be defined with any distribution in the exponential family parametrised by the expectation $\mathbb{E}(Y) = \mu$ as long as an appropriate link function can be defined.

The values of β_0 and $\boldsymbol{\beta}$ can still be found by minimising equation (3.6), but due to the non-linear link function these models cannot be fitted using simple matrix operations such as equation (3.10) and instead are solved numerically using an iterative approximation algorithm such as the Newton-Raphson method (Hastie et al., 2009). Alternatively, because the output Y is a random variable, the model can be trained through maximum likelihood estimation as described in section 3.2.2.2.

One commonly applied GLM uses a logistic link function of the form

$$\mu = g(\eta) = \frac{1}{1 + e^{-\eta}} = \frac{1}{1 + e^{-\left(\beta_0 + \mathbf{z}^\top \boldsymbol{\beta}\right)}} \tag{3.19}$$

where μ is now in the range $[0, 1]$, which makes it useful in regression tasks where the dependent variable is bounded. For regression purposes, equation (3.19) can be generalised as

$$\mu = \frac{b - a}{1 + e^{-\left(\beta_0 + \mathbf{z}^\top \boldsymbol{\beta}\right)}} + a, \tag{3.20}$$

which limits the range of μ smoothly between a and b. This is equivalent to scaling the dependent variable to be in the range $[0, 1]$ and using equation (3.19).

$\longrightarrow$

3.1.3 Logistic regression

If μ in equation (3.19) is interpreted as the mean of a Bernoulli distributed random variable, then the random variable Y has support in $\{0, 1\}$ where μ is the probability that $y = 1$. To formalise this, the probability that $y = 1$ given the data $\mathbf{z}$ can be expressed as

$$p(y = 1 \,|\, \mathbf{z}, \beta_0, \boldsymbol{\beta}) = \frac{1}{1 + e^{-\left(\beta_0 + \mathbf{z}^\top \boldsymbol{\beta}\right)}}. \tag{3.21}$$

In this form, the GLM is known as *logistic regression*.

In order to use the probabilistic output of the logistic regression model for classification a threshold is chosen where samples with a probability greater than the threshold are considered as belonging to the class. Typically a threshold of 0.5 is chosen, which will give the greatest accuracy, but this can be varied in the case of unequal prior probabilities or to tune the model for sensitivity or specificity rather than accuracy (see section 3.4.2 for further details of possible model evaluation metrics).

Logistic regression is one of the more traditional classifiers but it is still useful in many situations given the simplicity of the implementation. One disadvantage of logistic regression for classification is that, with a fixed threshold for classification, it imposes a strictly linear decision boundary over the feature space. This can be seen by setting $p(y = 1 \,|\, \cdot) = 0.5$ in equation (3.21), rearranging and taking logs to give the expression $\beta_0 = -\mathbf{z}^\top \boldsymbol{\beta}$ where $\mathbf{z}^\top$ is the hyper-plane decision boundary.

3.1.4 Naïve Bayes model

Logistic regression fits a logistic curve to training data to predict a binary response. While this provides an output "probability" of the classification of the data, this value is not calculated from the probability densities of the data and is simply a non-linear regression. Therefore it is not a true probability based on the training data. An alternative way to consider the problem of classifying data $\mathbf{z}$ into classes in $\mathcal{Y}$ is to consider the distribution of the training samples of $\mathbf{z}$ within each of the classes in $\mathcal{Y}$.

The naïve Bayes model considers each of the p features to be independently distributed random variables within each class where

$$Z_{i\,|\,y=c} \;=\; \pi(\theta_{i\,|\,y=c}) \;\; \forall\; i \in \{1,\ldots,p\}, c \in \mathcal{Y} \tag{3.22}$$

is the distribution of feature i in class c. Typically $\pi(\cdot)$ is Gaussian, but this is not required if the feature better fits another distribution. The distribution parameters $\theta_{i\,|\,y=c}$ are typically determined from the training data using maximum likelihood estimation as described in section 3.2.2.2.

The probability for data $\mathbf{z}$ to be in class c can then be calculated by direct application of Bayes' rule (see section 3.2.2.1) as

$$p(y = c\,|\,\mathbf{z}) = \frac{p(\mathbf{z}\,|\,y = c)\cdot p(y = c)}{p(\mathbf{z})} \tag{3.23}$$

$$\propto p(\mathbf{z}\,|\,y = c)\cdot p(y = c) \tag{3.24}$$

where

$$p(\mathbf{z}\,|\,y = c) = \prod_{i}^{p} p(z_i|y = c, \theta_{i\,|\,y=c}). \tag{3.25}$$

If all variables are Gaussian then this can be implemented as a multivariate Gaussian distribution with a diagonal covariance matrix. A classification can then be made by choosing the class that maximises $p(y = c\,|\,\mathbf{z})$, i.e.

$$y = \underset{c\,\in\,\mathcal{Y}}{\operatorname{argmax}}\;\; p(y = c\,|\,\mathbf{z}). \tag{3.26}$$

3.1.5 Discriminant analysis models

Discriminant analysis models are similar to the naïve Bayes model except that

a) the features are required to be multivariate Gaussian distributed, i.e.

$$\mathbf{z} \,|\, c \;\sim\; \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c); \text{ and} \tag{3.27}$$

b) the requirement of feature independence is relaxed, i.e. off-diagonal entries in the covariance matrix $\boldsymbol{\Sigma}_c$ are allowed.

3.1.5.1 Linear discriminant analysis

Linear discriminant analysis (LDA) models place an additional restriction that the covariance matrix for all classes must be the same, i.e. $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma} \;\forall\; c \in \mathcal{Y}$. This restriction can be interpreted as having identically shaped multivariate Gaussian clusters placed at different mean locations in $\mathcal{Z}$. The decision boundary for an LDA classifier is a linear function, hence the name LDA.

3.1.5.2 Quadratic discriminant analysis

If the LDA restriction of having the same covariance matrix for all classes is relaxed then the decision boundary becomes a quadratic function. This leads to the alternative quadratic discriminant analysis (QDA) model.

3.2 Probabilistic methods

Probabilistic methods of machine learning rely on constructing a *generative model* of the underlying process. A generative model describes a system as a hierarchical model of random variables where each level is conditioned on one or more of the higher levels. Generative models can be intuitively extended to describe highly complex models with multiple levels of dependent variables.

3.2.1 Bayesian models

Bayesian models consider the measured data in the context of a prior belief about the underlying population. Where a model contains parameter values, θ, a frequentist model estimates the parameter values from the available training data. The Bayesian way to model this data would instead be to apply some reasonable belief to the parameter values in the form of a prior probability distribution in the generative model, i.e. $\theta \sim \pi(\cdot)$.

3.2.2 Inference

The general aim of machine learning methods is to calculate and draw conclusions about latent quantities from values that are observed (Gelman et al., 2013). This process is generally known as *statistical inference* and can take several forms as discussed in the following sections.

3.2.2.1 Definitions

As a generalised notation, the latent variables in the model are referred to as θ and the observed values as x. The generative definition of the model can be used to construct the *joint probability* of x and θ occurring using the chain rule (Murphy, 2012) so that

$$p(x, \theta) = p(x \wedge \theta) = p(x \,|\, \theta) \cdot p(\theta) \tag{3.28}$$

where $p(\theta)$ is the prior probability of θ. Once the joint has been constructed, Bayes' rule (Murphy, 2012) can be used to calculate the probability of θ given the observed data x as

$$p(\theta \,|\, x) = \frac{p(x \,|\, \theta) \cdot p(\theta)}{p(x)} \tag{3.29}$$

which is known as the *posterior probability* of the model. The observed data x is known as the *evidence* of the model with parameters θ.

The normalising constant $p(x)$ in equation (3.29) cannot be expressed directly due to the inherent dependencies on the other model variables. However, it can be

obtained by marginalising the joint distribution over θ

$$p(x) = \int_{-\infty}^{\infty} p(x, \theta)\, d\theta \tag{3.30}$$

where the joint distribution $p(x, \theta)$ can be substituted by equation (3.28). In practice there is often no analytic solution to this integral, but because it is a function of x only, it is constant in the posterior in equation (3.29). This means that it can often be omitted in the posterior distribution by expressing it as

$$p(\theta \,|\, x) \propto p(x \,|\, \theta) \cdot p(\theta) \tag{3.31}$$

which is often known as the *unnormalised posterior density* (Gelman et al., 2013).

3.2.2.2 Maximum likelihood / maximum *a posteriori* estimation

One common form of inference is to determine the value of the θ that makes x most likely to have occurred. In other words, inferring the optimal value of θ that maximises the likelihood of the data, i.e.

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ \mathcal{L}(x \,|\, \theta) \tag{3.32}$$

$$= \underset{\theta}{\operatorname{argmax}}\ p(\theta \,|\, x). \tag{3.33}$$

Where no prior belief is applied on the value of θ this is known as the *maximum likelihood* estimate (MLE). To apply a prior belief $p(\theta)$ on the value of θ, $p(\theta \,|\, x)$ in equation (3.33) is substituted with the posterior definition from equation (3.29) to give

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ \frac{p(x \,|\, \theta) \cdot p(\theta)}{p(x)}. \tag{3.34}$$

In this case θ^* represents the value of θ at the mode of the posterior distribution, hence this method is known as *maximum a posteriori* (MAP) estimation. Because $p(x)$ is constant over θ, equation (3.34) can be simplified to give

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ p(x \,|\, \theta) \cdot p(\theta) \tag{3.35}$$

$$= \underset{\theta}{\operatorname{argmax}}\ p(x, \theta). \tag{3.36}$$

3.2.3 Posterior variance

The MAP estimate of the posterior distribution gives a point estimate that is useful for many applications. However, it can be also useful to obtain a metric for how reliable the MAP estimate is. Because the posterior distribution is just a probability distribution, the variance (and higher order moments) of the distribution can be calculated using expectations.

3.2.4 Sampling

The difficulty of calculating the MAP estimate of θ^* in equations (3.35) and (3.36) varies considerably depending on the model. There is often no closed-form solution, but gradient descent (or other standard numerical optimisation methods) can be used. However, these general methods are prone to finding local optima in complex multimodal posterior densities. In some applications it is also useful to know about the different modes in the posterior, some of which may not be found using standard optimisation methods. Finally, while such methods will generally converge to a (local) MAP estimate, higher order moments will not be available.

One approach is to perform a grid search over the whole parameter space of θ. However, the support of one or more of the components of θ is typically infinite and the search must therefore be artificially bounded. The prior belief placed on the components of θ can be used to bound the search within the most probable values, but this does not factor in data that may overwhelm the prior. Additionally, even if suitable bounds can be found (typically over a large range), the granularity of the grid must be chosen appropriately to maximise the parameter space explored against performance. In smooth posterior distributions it may be possible to start with a coarse grid and focus in on areas of interest, but this may still miss narrow modes.

A more Bayesian approach is to use the prior belief to choose values of θ from. Consider that each model variable θ_i is distributed by prior distribution π_i. The *base distribution* can then be defined as

$$G_0 = \{\pi_1, \ldots, \pi_p\}, \tag{3.37}$$

and a unique draw j from the base distribution is denoted as $\theta^{(j)} \sim G_0$. By repeatedly drawing values of $\theta^{(j)}$ from G_0 and evaluating the posterior using equation (3.31), the posterior can be searched without artificially bounding the search space, or having to choose the grid resolution.

This is the most basic of the so-called *Monte Carlo* (MC) methods, named after the European city known for casino gambling due to their stochastic nature (Metropolis & Ulam, 1949; Eckhardt, 1987; Metropolis, 1987). When using MC methods, the process of drawing values is known as *sampling*.

A problem with the previously described method of repeatedly sampling values of $\theta^{(j)}$ from G_0 is that samples of θ with expectation $\mathbb{E}(G_0)$ are used to infer the properties of $p(\theta \mid x)$. The posterior expectation $\mathbb{E}(\theta \mid x)$ is likely significantly different from $\mathbb{E}(G_0)$ which means that if the prior distributions π_i are not well defined then it can take many samples to find the posterior mode(s), and even more to have sufficient surroundings samples to determine the properties of the mode(s).

Instead, what if there was a way to directly sample from the posterior distribution? More formally, a way to sample data points $\mathbf{\Theta} = \{\theta_1, \theta_2, \ldots\}$ from the whole posterior distribution with probability $p(\theta_i \mid x)$. Data points near the mode(s) of $p(\theta \mid x)$ will be more common in the sampled data set $\mathbf{\Theta}$. Using these samples, the different modes of the posterior can be found, as well as higher order moments or other statistics of interest. As more samples are generated, information about the posterior can be determined with ever increasing accuracy.

As discussed above, $p(\theta \mid x)$ is usually intractable and cannot be sampled directly. However, for a given θ, it is usually possible to calculate the value of $p(\theta \mid x)$. Using this property, several numeric MC methods have been developed to allow sampling directly from $p(\theta \mid x)$.

3.2.4.1 Metropolis-Hastings sampling

While working on many-body problems in statistical mechanics, Metropolis et al. (1953) modified the Monte Carlo method in what has been called one of the top 10 algorithms of the 20$^{\text{th}}$ century by the editors of *SIAM News* (Cipra, 2000). The key

contribution of Metropolis et al. was to sample a proposed value of θ^* from a *proposal distribution* $\pi(\cdot)$ conditioned on the current sample θ^i, rather than randomly sampling each sample from the whole space of θ.

Specifically, Metropolis et al. used a symmetric distribution for θ^* centred on θ^i, for example, a Gaussian distribution

$$\theta^* \,|\, \theta^i \;\sim\; \mathcal{N}(\theta^i, \sigma^2) \tag{3.38}$$

where σ^2 is the standard deviation. The proposed value θ^* is then accepted as the next sample θ^{i+1} with probability

$$\begin{aligned} p^{\text{accept}} &= \begin{cases} \frac{p(\theta^* \,|\, x)}{p(\theta^i \,|\, x)} & \text{if } \frac{p(\theta^* \,|\, x)}{p(\theta^i \,|\, x)} < 1 \\ 1 & \text{otherwise} \end{cases} \\ &= \min\left(1, \frac{p(\theta^* \,|\, x)}{p(\theta^i \,|\, x)}\right). \end{aligned} \tag{3.39}$$

If the proposed value is not accepted, the current value is retained for the next sample θ^{i+1}. This is known as the *Metropolis algorithm*.

The full chain of samples $\boldsymbol{\Theta} = \theta^0, \theta^1, \ldots$ can be interpreted as a random walk through the space of θ with samples concentrated around the higher probability region(s) of the posterior distribution $p(\theta \,|\, x)$. The sampled values $\boldsymbol{\Theta}$ can hence be seen as draws from $p(\theta \,|\, x)$ and can be used to calculate properties of the posterior distribution. The ability of moving to parts of the posterior with lower probability means that it can sample effectively from multi-modal posterior distributions.

If the proposal distribution is asymmetric, i.e. $q(\theta^* \,|\, \theta^i) \neq q(\theta^i \,|\, \theta^*)$, then the acceptance probability in equation (3.39) will not result in the correct samples from $p(\theta \,|\, x)$. In this case, Hastings (1970) provided an adapted acceptance probability of

$$p^{\text{accept}} = \min\left(1, \frac{p(\theta^* \,|\, x) q(\theta^i \,|\, \theta^*)}{p(\theta^i \,|\, x) q(\theta^* \,|\, \theta^i)}\right) \tag{3.40}$$

$$= \min\left(1, \frac{p(\theta^* \,|\, x)/q(\theta^* \,|\, \theta^i)}{p(\theta^i \,|\, x)/q(\theta^i \,|\, \theta^*)}\right) \tag{3.41}$$

giving rise to the more general *Metropolis-Hastings algorithm*. This modification becomes particularly relevant when dealing with non-infinite distributions of θ, for example if θ is constrained to be non-negative.

3.2.4.2 Conjugate priors

As discussed above, the posterior distribution $p(\theta \mid x)$ is often intractable. However, there is a class of models where it is possible to sample directly from $p(\theta \mid x)$.

Considering the generative model

$$x \mid \mu \sim \mathcal{N}(\mu, 1) \tag{3.42a}$$

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2), \tag{3.42b}$$

the task is to sample μ from

$$p(\mu \mid x) = \frac{p(x \mid \mu)p(\mu)}{p(x)} \propto p(x \mid \mu)p(\mu) \propto p(x, \mu). \tag{3.43}$$

In this case, both $p(x \mid \mu)$ and $p(\mu)$ are normally distributed, and $p(\mu \mid x)$ can be expressed analytically as

$$p(\mu \mid x) \propto \overbrace{\frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp\left\{-\frac{(x-\mu)^2}{2 \cdot 1}\right\}}^{p(x \mid \mu)} \cdot \overbrace{\frac{1}{\sqrt{2 \cdot \sigma_0^2 \cdot \pi}} \exp\left\{-\frac{(\mu-\mu_0)^2}{2 \cdot \sigma_0^2}\right\}}^{p(\mu)} \tag{3.44}$$

or, with n data samples $\mathbf{x} = [x_1, \ldots, x_n]^\top$,

$$p(\mu \mid \mathbf{x}) \propto \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\mu)^2}{2}\right\}\right] \cdot \frac{1}{\sqrt{2\sigma_0^2\pi}} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}. \tag{3.45}$$

This can be rearranged so that

$$p(\mu \mid x) \propto \exp\left\{-\frac{\mu^2 - 2\mu'\mu}{2\sigma^{2\prime}}\right\} \propto \exp\left\{-\frac{(\mu-\mu')^2}{2\sigma^{2\prime}}\right\}, \tag{3.46}$$

where

$$\mu' = \frac{\mu_0 + \sigma_0^2 \mathbf{x}^\top \mathbf{1}_n}{1 + n\sigma_0^2} = \frac{\frac{1}{n}\mu_0 + \frac{1}{n}\sigma_0^2 \mathbf{x}^\top \mathbf{1}_n}{\frac{1}{n} + \sigma_0^2} = \frac{\frac{1}{n}\mu_0 + \sigma_0^2 \bar{\mathbf{x}}}{\frac{1}{n} + \sigma_0^2} \quad ; \quad \text{and} \tag{3.47a}$$

$$\sigma^{2\prime} = \frac{\sigma_0^2}{1 + n\sigma_0^2} \tag{3.47b}$$

from which it is clear that μ can be sampled from a Gaussian distribution with mean μ' and standard deviation $\sigma^{2\prime}$.

This ability to form an analytical expression of the posterior that can be directly sampled is only possible with certain combinations of likelihood function and prior. A prior that enables such an expression to be formed together with a given likelihood function is called a *conjugate prior* for the likelihood.

3.2.4.3 Gibbs sampling

In practice it is rarely possible to form an analytic expression for the full posterior distribution that allows easy sampling. However, even in complex models it is often possible to express subsets of the model in analytically tractable ways.

This is the underlying concept of the *Gibbs sampler*, first described by Geman & Geman (1984) for restoration of noisy images using a Bayesian generative model. The Gibbs sampler was further formalised and popularised outside of the field of image-processing by Gelfand & Smith (1990).

The Gibbs sampler considers any model with model variables $\theta_1, \theta_2, \ldots, \theta_p$. The algorithm proceeds by looping through all variables, θ_i, in turn at each time step, t, and samples a new value for $\theta_i^{(t)}$ conditioned on the sampled values of $\theta_{j\neq i}$. If $\theta_{j\neq i}$ has already been sampled at time t then the new sample of $\theta_i^{(t)}$ is conditioned on $\theta_{j\neq i}^{(t)}$ otherwise it is conditioned on $\theta_{j\neq i}^{(t-1)}$. Algorithm 3.1 describes the Gibbs sampler in full.

The act of sampling each variable θ_i conditioned on all the others has a significant simplifying effect on the joint probability. All the multiplicative terms in the joint not involving θ_i now become constant and can be dropped while still retaining proportionality. Many of the variables $\theta_{j\neq i}$ will also be dropped off, and the only ones guaranteed to exist in $p(\theta_i \mid \theta_{j\neq i})$ are the variables with direct conditions in the generative model, i.e. variables that are conditioned on θ_i, or variables that θ_i is explicitly conditioned on. If the conditioned variables are conjugate priors for θ_i, then it is said to be *conditionally conjugate* and the variable can be directly sampled as detailed in the previous section.

A Gibbs sampler implemented in this way can be very flexible. In the case of a supervised problem where the $\mathbf{y}$ values are known from the training data, the sampling of $y_1, \ldots, y_n$ can simply be omitted and the rest of the variables sampled. In the other extreme, where no labels are available, the sampler will attempt to identify the groups in an unsupervised way directly from the data. If only a subset of data points are

Algorithm 3.1 Generic Gibbs sampler.

data: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ the observed data.
result: $\boldsymbol{\Theta}$ samples of model variables $\theta_1, \ldots, \theta_p$.

begin
  $t \leftarrow 0$
  $\theta_1^{(t)} \sim \pi(\theta_1)$ ▷ Sample all variables from their prior (conditional) distributions, or set to some reasonable value.
  $\vdots$
  $\theta_p^{(t)} \sim \pi(\theta_p)$
  for $t = 1, 2, \ldots$ **do** ▷ Repeat until convergence.
    $\theta_1^{(t)} \sim p(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, \theta_4^{(t-1)}, \ldots, \theta_p^{(t-1)}, \mathbf{X})$ ▷ Sample θ_i conditioned on current sampled values of $\theta_{j \neq i}$. If $\theta_{j \neq i}$ has been sampled at time t condition on $\theta_{j \neq i}^{(t)}$ otherwise condition on $\theta_{j \neq i}^{(t-1)}$.
    $\theta_2^{(t)} \sim p(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, \theta_4^{(t-1)}, \ldots, \theta_p^{(t-1)}, \mathbf{X})$
    $\theta_3^{(t)} \sim p(\theta_3 \mid \theta_1^{(t)}, \theta_2^{(t)}, \theta_4^{(t-1)}, \ldots, \theta_p^{(t-1)}, \mathbf{X})$
    $\theta_4^{(t)} \sim p(\theta_4 \mid \theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_p^{(t-1)}, \mathbf{X})$
    $\vdots$
    $\theta_p^{(t)} \sim p(\theta_4 \mid \theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_{p-1}^{(t)}, \mathbf{X})$
  end for
end

labelled, then the sampler can be run in a semi-supervised way, using the labels where available and inferring the others.

While having conditional conjugacy is highly convenient for sampling, it is not required. Where there are some conditional densities that are not conjugate, Metropolis-Hastings samples can be used instead as described in section 3.2.4.1.

3.3 Non-parametric methods

The models described above all belong to the class of *parametric* models, where there are a fixed number, m, unknown parameters, $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$, specified in the model definition. The aim when performing inference is therefore to determine the optimal values of the m parameters in $\boldsymbol{\theta}$, and perhaps other statistical properties as described above.

An alternative way to specify certain types of model is to say that the number of parameters, m, is unknown, and then try to determine this as an additional model parameter. This is particularly suitable to mixture models, for example, where instead of specifying the number of classes, k, up front, the value of k is allowed to adjust to match the data. This model could be defined as having k classes; each with one mean parameter, μ_i; and a single parameter, q, specifying the distribution between the classes. With this definition, there are a total of $m = k + 2$ unknown parameters, $\boldsymbol{\theta} = \{k, \mu_0, \ldots, \mu_{k-1}, q\}$. Because the total number of parameters is a function of k, which is also itself one of the unknown parameters, the model can be said to have an unknown, and potentially infinite (but countable), number of unknown parameters. This is known as a *non-parametric model*.

Before considering how to perform inference over a model with a potentially infinite number of unknown parameters, consider how to increase the number of states in the mixture model discussed above. If the number of classes were fixed to $k = 2$, the q parameter can be used as the probability of a data point, i, being in class 1, with the class, denoted y_i, drawn from a Bernoulli distribution. An alternative (and equivalent) way to define the model is to define $\mathbf{q}^* = [q - 1, q]^\top$ and draw the class from a categorical distribution, i.e.

$$y_i \sim \text{Cat}(\mathbf{q}^*). \tag{3.48}$$

It is now easy to see how this simple model can be extended to cover k classes instead of just two. The $\mathbf{q}^* = [q_1^*, \ldots, q_k^*]^\top$ parameter, now a vector of length k, defines the probability of drawing each of the k classes, and has the following properties: a) $q_i^* \geq 0 \; \forall \; i$; and b) $\sum_{i=1}^k q_i^* = 1$. A distribution that provides vectors that meet these requirements is the *Dirichlet distribution*, so the model can be defined with

$$\mathbf{q}^* \sim \text{Dir}(\boldsymbol{\alpha}). \tag{3.49}$$

The $\boldsymbol{\alpha}$ parameter of the Dirichlet distribution is also a k-length vector, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k]^\top$, that defines the concentration of mass in $\mathbf{q}^*$.

Commonly $\boldsymbol{\alpha} = \alpha \mathbf{1}_k$, in which case the following properties apply:

- **if** $0 < \alpha < 1$**:** mass is concentrated in small subsets within $\mathbf{q}^*$, i.e. as $\alpha \to 0$, $q_i^* \to 1$ for any one $i \in [1, k]$ while $q_{j \neq i}^* \to 0$. However, it does not specify *which* i mass is concentrated on in each draw.
- **if** $\alpha = 1$**:** mass is randomly allocated throughout $\mathbf{q}^*$, i.e. all possible instances of $\mathbf{q}^*$ are equally likely.
- **if** $\alpha > 1$**:** mass is distributed evenly between all $q_i^* \ \forall \ i$, i.e. as $\alpha \to \infty$, $q_i^* \to 1/k \ \forall \ i$.

If different values are given to different elements of $\boldsymbol{\alpha}$ then mass distribution within $\mathbf{q}^*$ can be more controlled. Elements of $\boldsymbol{\alpha}$ with higher values are more likely to have higher values in $\mathbf{q}^*$ and vice versa. The Dirichlet distribution is the conjugate prior of the categorical distribution and so this is a natural extension of the above model and permits easy sampling.

Considering the case when $0 < \alpha < 1$, this can be seen to concentrate probability mass in $\mathbf{q}^*$ to a subset of elements. This also has the effect of regularising the total number of classes included in the model, by setting multiple $\mathbf{q}_i^* \approx 0$. If the number of classes in the model is unknown, then one possibility is to set $k \gg$ the anticipated number of classes, and use a low value of α to concentrate the mass. While this will concentrate mass on a subset of $\mathbf{q}^*$, it will not concentrate the mass on specific classes. To concentrate mass on the first classes in $\mathbf{q}^*$, α_i can be set to exponentially decreasing values.

This model can be extended so that $k = \infty$ with exponentially decreasing values of α_i to effectively exclude having over a certain number of classes. This is the basis of the *Dirichlet process* (DP) model. The DP is defined as a non-parametric model parametrised by a base distribution G_0 and a concentration parameter α, denoted as

$$\theta^{(g)} \,|\, G \ \sim \ G \tag{3.50a}$$

$$G \,|\, G_0 \ \sim \ \mathrm{DP}(\alpha, G_0) \tag{3.50b}$$

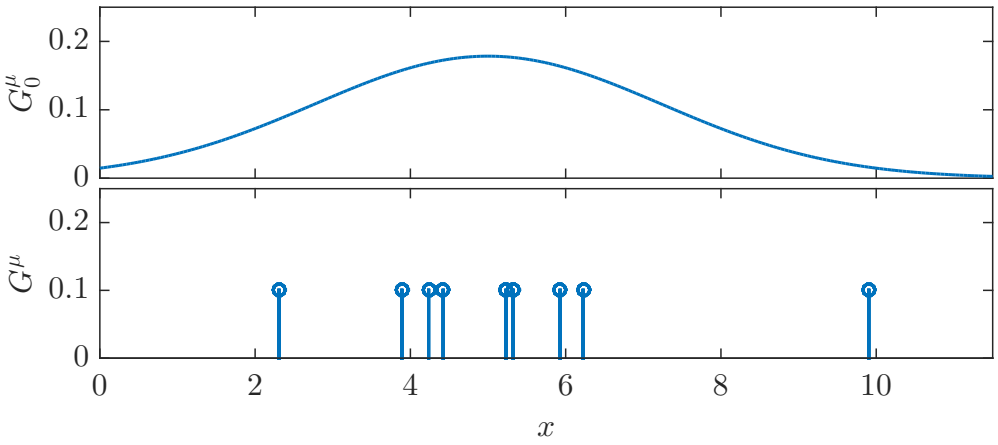


Figure 3.1: Example of a DP prior for the model given in equation (3.51). The top graph shows the base distribution G_0^μ, and the lower graph shows discrete draws from G_0^μ.

where $\theta^{(g)}$ is a model variable and the $^{(g)}$ superscript indicates that θ is a grouped variable. The base distribution G_0 defines the prior distribution of all grouped variables, which may be independent or conditional; continuous or discrete; or some combination thereof. The base distribution often has infinitely many realisations. However, a draw G from the DP contains a finite number of realisations of G_0. This may contain realisations that are already used, or new draws from the prior. A draw from G returns one of the realisations.

To exemplify this, the previously described mixture model can be represented as

$$x_i \mid \mu_i^{(g)} \sim \mathcal{N}(\mu_i, 1) \tag{3.51a}$$

$$\mu_i^{(g)} \mid G \sim G \tag{3.51b}$$

$$G \mid G_0 \sim \mathrm{DP}(\alpha, G_0) \tag{3.51c}$$

where G_0 is a distribution over μ such that

$$G_0^\mu = \mathcal{N}(\mu_\mu, \mu_{\sigma^2}). \tag{3.51d}$$

A Gaussian prior can be specified as the base distribution with $\mu_\mu = 5$; and $\mu_{\sigma^2} = 5$. This model is shown in figure 3.1 above, where the top graph shows the base distribution G_0^μ with infinitely many realisations. Drawing μ_i directly from G_0^μ would result in a negligible probability of any two points having the same μ_i. Because

explicit group indices are no longer defined, the aim is to sample from G_0^μ in such a way that $\mu_i = \mu_{j \neq i}$ for all $j \neq i$ that are in the same group. The value of μ_i therefore becomes an implicit group identifier.

G therefore provides a set of discrete realisations from G_0. In this case G^μ indicates realisations of the μ variable, but any number of variables can be represented in G. The bottom graph in figure 3.1 shows 10 discrete draws from G_0^μ.

The α parameter controls the probability of drawing a new realisation from G_0 rather than one of the existing realisations in G. Specifically, if n samples, $\theta_1, \ldots, \theta_n$, have been drawn from the DP, then the next sample, θ_{n+1}, will be set according to

$$\theta_{n+1} \,|\, \theta_1, \ldots, \theta_n \;=\; \begin{cases} \theta_1 & \\ \;\vdots & \text{each with probability } \frac{1}{n-1+\alpha} \\ \theta_n & \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha}. \end{cases} \tag{3.52}$$

It can be seen that with larger values of α it is more likely that a new value will be drawn from G_0.

Note that in equation (3.52) the values of $\theta_1, \ldots, \theta_n$ are not unique, they are simply previous draws from G using equation (3.52), which means that there are likely to exist repeated values. This implies another important property of the DP. If there are m instances where $\theta_i = \theta_{j \neq i}$ then there is effectively a probability $\frac{m}{n-1+\alpha}$ of setting θ_{n+1} to θ_i. This property is sometimes referred to as the *rich gets richer* property of the DP, which helps clusters to form (Wallach et al., 2010).

In practice, working directly with $\theta_1, \ldots, \theta_n$ is often not useful. If $\phi_1, \ldots, \phi_k$ instead denote the k unique values of $\theta_1, \ldots, \theta_n$, then equation (3.52) can be written as

$$\theta_{n+1} \,|\, \phi_1, \ldots, \phi_k \;=\; \begin{cases} \phi_1 & \\ \;\vdots & \text{each with probability } \frac{n_c}{n-1+\alpha} \\ \phi_k & \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases} \tag{3.53}$$

where n_c is the number points in group c, i.e. $n_c = \sum_{i=1}^{n} \mathbb{I}(\theta_i = \phi_c)$. Equation (3.53) is commonly referred to as the Chinese restaurant process (CRP) prior (Pitman, 2006)

from a fanciful analogy with customers entering a Chinese restaurant and being more likely to sit at tables where previous customers are already sitting.

The probability of drawing a given sequence of n samples with k unique values of θ_i, each with n_c samples allocated, can be derived from equation (3.53) as

$$p(\theta_1, \ldots, \theta_n \,;\, \alpha) = \alpha^k \prod_{i=1}^{n} \frac{1}{i-1+\alpha} \prod_{c=1}^{k} (n_c - 1)! = \alpha^k \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{c=1}^{k} \Gamma(n_c). \tag{3.54}$$

3.3.1 Inference

As discussed in detail earlier, the general aim when performing inference in probabilistic models is to determine properties about the posterior distribution

$$p(\theta \,|\, x) = \frac{p(x \,|\, \theta) \cdot p(\theta)}{p(x)} \tag{3.55}$$

$$\propto p(x \,|\, \theta) \cdot p(\theta). \tag{3.56}$$

The difference in the non-parametric model is that the number of elements of θ is not known in advance. This poses some unique challenges when performing inference because variables need to be dynamically created and/or destroyed as required.

One well known algorithm was developed by Radford Neal (2000) in a paper discussing various different methods of sampling a DP. Neal's final algorithm was numbered 8 in the paper, and it has since become known simply as *Algorithm 8*.

Neal's Algorithm 8 is an iterative sampling algorithm that introduces auxiliary variables to increase the chance of moving to a new state drawn from G_0. The only requirement to run Algorithm 8 is that it is possible to calculate the joint $p(\mathbf{x}_i, \phi_c)$ for data sample i under model parameter(s) ϕ_c. In each iteration, data sample i can be allocated to either one of the existing groups $c \in \mathcal{C}$ with parameters ϕ_c or to one of m new groups $c \in \mathcal{C}^*$ with parameters $\phi_c \sim G_0$. The probability of allocating data sample i to one of the existing groups or one of the new groups is given by

$$p(c_i = c \,|\, \mathcal{C}, \mathcal{C}^*) \propto \begin{cases} \frac{n_{-i,c}}{n-1+\alpha} \cdot p(\mathbf{x}_i, \phi_c) & \forall\ c \in \mathcal{C} \\ \frac{\alpha}{|\mathcal{C}^*|} \cdot \frac{1}{n-1+\alpha} \cdot p(\mathbf{x}_i, \phi_c) & \forall\ c \in \mathcal{C}^* \end{cases} \tag{3.57}$$

where $n_{-i,c}$ denotes the number of data samples allocated to group c excluding data sample i; and $|\mathcal{C}^*|$ indicates the number of groups in $\mathcal{C}^*$. The algorithm is summarised in algorithm 3.2 on the following page.

Algorithm 3.2 DP model inference using Radford Neal's Algorithm 8 (Neal, 2000).

data: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ the observed data.
result: $\boldsymbol{\Theta}$ samples of model variables $\theta_1, \ldots, \theta_p$.

begin
  for $i = 1, \ldots, n$ **do**
    $c_i \leftarrow 1$ ▷ Assign all data points to group 1.
  end for
  $\phi_1 \sim G_0$ ▷ Assign parameters for group 1 from base distribution.
  for $t = 1, 2, \ldots$ **do** ▷ Repeat until convergence.
    for $i = 1, \ldots, n$ **do** ▷ Loop over all data samples.
      Let $\mathcal{C}$ be the set of all unique groups $c_{j \neq i}$.
      Let $\mathcal{C}^*$ be a set of m new groups with parameters ϕ_c.
      for $c \in \mathcal{C}^*$ **do** ▷ Set initial parameter values for new
        $\phi_c \sim G_0$ groups from base distribution.
      end for
      if i is the only data sample in group c_i **then** add ϕ_{c_i} to $\mathcal{C}^*$.
      Assign c_i from $\mathcal{C}$ or $\mathcal{C}^*$ using equation (3.57).
    end for
    Let $\mathcal{C}$ be the set of all unique groups c_j.
    for $c \in \mathcal{C}$ **do**
      Sample parameter values ϕ_c conditional on data samples in group c.
    end for
  end for
end

3.4 Model evaluation metrics

This section describes some basic model evaluation metrics for regression and classification that are used throughout the rest of this thesis.

3.4.1 Regression models

The performance of regression models is commonly evaluated using the mean absolute error (MAE) or root mean square deviation (RMSD). The MAE is a simple metric averaging the absolute deviation of the predictions from the true value, defined as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \tag{3.58}$$

An alternative metric is the RMSD, which instead averages the the square of the deviation, defined as

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{3.59}$$

The RMSD is useful in many situation because it is most influenced by points with large errors. This makes it useful as a training metric where it is useful to penalise large misclassifications more than smaller ones. However, because this amplifies the effect of larger errors it may give misleading results if these errors are from outliers (Chai & Draxler, 2014). It can also make it more difficult to interpret the result. MAE has therefore been proposed by Willmott & Matsuura (2005) among others as a more suitable model evaluation metric.

Both the MAE and RMSD are affected by the scale of the data, which can make it difficult to compare the results to other models on data with different scales. The normalised mean absolute error (NMAE) and normalised root mean square deviation (NRMSD) are variations of equations (3.58) and (3.59) that normalise the metrics by dividing by the range of the data, i.e.

$$\text{NMAE} = \frac{\text{MAE}}{y_{max} - y_{min}} = \frac{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|}{y_{max} - y_{min}} \tag{3.60}$$

$$\text{NRMSD} = \frac{\text{RMSD}}{y_{max} - y_{min}} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{y_{max} - y_{min}} \tag{3.61}$$

3.4.2 Classification models

Results from classification models are commonly presented as accuracy (Ac), sensitivity (Se) and specificity (Sp). Accuracy defines the overall percentage of points classified correctly, defined as

$$\mathrm{Ac} = \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i)}{N}, \tag{3.62}$$

where y_i denotes the correct class of data point i, and $\hat{y}_i$ denotes the estimated value. Accuracy is useful to provide a relative indication or how well a model performs, but in the case of unbalanced data sets it can provide misleading results. Sensitivity and specificity are commonly used in addition to accuracy to give an indication of how well the model performs on each class. Sensitivity, also known as the true positive rate (TPR), is the proportion of the true positive samples correctly classified, defined as

$$\mathrm{Se} = \mathrm{TPR} = \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = 1) \cdot \mathbb{I}(y_i = 1)}{\sum_{i=1}^{N} \mathbb{I}(y_i = 1)}, \tag{3.63}$$

while specificity is the proportion of true negative samples correctly classified, defined as

$$\mathrm{Sp} = \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = 0) \cdot \mathbb{I}(y_i = 0)}{\sum_{i=1}^{N} \mathbb{I}(y_i = 0)}. \tag{3.64}$$

A further useful metric is the false positive rate (FPR), the proportion of true negative samples incorrectly classified as positive, defined as

$$\mathrm{FPR} = \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = 1) \cdot \mathbb{I}(y_i = 0)}{\sum_{i=1}^{N} \mathbb{I}(y_i = 0)} = 1 - \mathrm{Sp}. \tag{3.65}$$

An optimal classifier has high accuracy, sensitivity, and specificity. However, when used as metrics for model training, optimising for one of the metrics above can often be at the expense of one of the others. In particular, optimising only for specificity or sensitivity will tend to severely negatively affect the other. An alternative metric is the F_1 score (F_1), defined as

$$\mathrm{F}_1 = \frac{2 \cdot \mathrm{Se} \cdot \mathrm{Sp}}{\mathrm{Se} + \mathrm{Sp}}, \tag{3.66}$$

which combines the sensitivity and specificity with equal weighting to provide a single score (van Rijsbergen, 1979). It is also possible to put more emphasis on sensitivity or specificity.

Classification performance can also be presented using the receiver operating characteristic (ROC) curve, which is defined as a plot of the FPR against the TPR of the classifier. The different values of FPR and TPR between 0 and 1 are generated by adjusting the class acceptance threshold for the classifier. The area under this curve (known as the the AUC) is also often presented as an indicator of the performance of the model to different inputs.

3.5 Model validation

A performance evaluation of a model must give as accurate an estimate of performance on unseen data as possible. Care must be taken to ensure that the results are presented fairly – a model trained and tested on the same data will perform unrealistically well.

For this reason, a subset of the data are commonly excluded when training the model and are subsequently used to evaluate the performance of the trained model. This has two downsides. Firstly the excluded data are essentially wasted for training the model. Secondly this assumes that the excluded data are representative of the population. With a large data sets these two downsides reduce in importance, but for smaller data sets they can make a big difference to the results.

In order to make the best use of the available data and provide a representative indicator of performance on unseen data, a method known as *cross-validation* (CV) is commonly used. Cross-validation involves splitting all the available data into distinct subsets for training and evaluating the model. By randomly partitioning the data into training and test subsets, estimates can be made of the performance on new data as well as the sensitivity of the model to specific set of training data.

Two common methods are k-fold cross validation and leave-one-out cross validation. In k-fold cross validation the whole data set is split into k equal sized groups. One of the groups is left out and the model is trained on the other $k-1$ groups. The

model performance is then evaluated on the left out group, and the process iterated k times until all groups have been used for testing. If k is set to the number of data points in the whole data set, then the special case of leave-one-out cross-validation is obtained. Leave-one-out cross-validation has the advantage that it makes the best use out of the available data. The main disadvantage is that it does not give any indication of the variance in the performance of the model. In k-fold and leave-one-out cross-validation, the overall performance metric can be calculated after iterating over all folds, which results in a single metric. With k-fold cross-validation, different splits can be tested to provide an indication of the sensitivity of the model to different training subsets. This is not possible in leave-one-out cross-validation because the folds are fully deterministic.

Independence of samples must also be considered. Using any of the above methods as described assumes that all samples in the data set are independent. In many cases, for example features derived from recorded geolocation data, the features from one individual cannot be assumed to be independently distributed under a global distribution. Leave-one-participant-out cross-validation can instead be used to validate models where groups of samples (from a study participant) are not thought to be independent. In this method all the labelled data for a participant are used for testing, while the model is trained on the labelled data for the remaining participants. This is repeated for all participants to give an overall result. As with leave-one-out cross-validation, this only provides a single result since the folds are deterministic.

3.5.1 Group equalisation

In classification tasks care must be taken when training and evaluating models where the class distributions are unequal, or unbalanced. Consider a data set where 90 % of the samples belong to class G_1, and 10 % belong to class G_2. A classification accuracy of 90 % can be achieved trivially by classifying all samples as class G_1 (sensitivity will be 100 % while specificity will be 0 %, resulting in an F_1 score of 0 %). This demonstrates the importance of presenting appropriate statistics of the classification results, and optimising for an appropriate statistic.

When training a model, one way to handle unbalanced data sets is to incorporate the class proportions into the model (as done in the naïve Bayes model described in section 3.1.4) or to adjust the acceptance probability on the output (for example adjusting the acceptance probability in logistic regression as described in section 3.1.3).

An alternative way is to over-sample the under-represented group. This involves artificially increasing the size of the under-represented group by either duplicating training samples exactly, or by drawing new samples from the same distribution. The former method can be problematic because it can over-fit the model to the duplicated samples, while the latter method requires that there are enough data to accurately determine the distribution from which to draw samples.

A final method, and the one used for classification of depression in chapter 6, is to sub-sample the over-represented group. This method reduces the size of the whole data set to ensure that both groups are equally represented. In each iteration of cross-validation, regardless of the method, the training data set has n_1 samples from group G_1, and n_2 samples from group G_2. In the case where $n_1 \neq n_2$, $n_{min} = \min(n_1, n_2)$ samples are randomly selected from both G_1 and G_2 to form G_T (i.e. all samples from the under-represented group are used, and the same number of samples are randomly chosen from the over-represented group). G_T is used as the training data to train the model, and performance is tested on the left-out data as usual.

This technique can be applied to all cross-validation methods, and the selection of G_T is repeated M times for each fold or left out participant. With leave-one-out or leave-one-participant-out cross-validation this provides a way to generate statistics on the model performance based on the different sub-samples of training data. Where k-fold cross-validation is used, the random generation of the k partitions is repeated $\lceil N/k \rceil$ times where N is the number of participants. This provides the same number of tests for each data point to generate comparable statistics of model performance.

Group equalisation ignores the underlying group distributions in the population as a whole. This is likely to be important to consider in practice, but can be applied independently of the model training phase by applying a prior to the classification.

Using group equalisation is nevertheless convenient as the results are easily interpretable, especially by presenting the range of statistics described above.

3.6 Feature selection

Typically p features are extracted from the raw data with the aim of applying the chosen machine learning model. The full p-dimensional feature space $\mathcal{Z}$ will generally contain features that are either

a) strongly relevant;

b) weakly relevant but non-redundant;

c) redundant; or

d) irrelevant (Estévez et al., 2009; Yu & Liu, 2004).

In some applications there may be hundreds or thousands of features and the task of the feature selection method is to select the smallest k maximally informative features that optimise the performance of the machine learning model. The metric for optimal performance of the machine learning model varies depending on the model type and application, but the optimal feature set will generally aim to maximise the number of relevant features while minimising the number of redundant features (Peng et al., 2005).

There are three main reasons why feature selection is an important part of pattern recognition tasks. Firstly prediction accuracy is heavily influenced by the number of features used for the prediction task. More features result in a data set with greater variance which will improve in-sample classification at the cost of over-fitting the training data, which can reduce out-of-sample prediction accuracy (Hastie et al., 2009). The second reason is hardware and software resource reduction. A machine learning algorithm run on fewer features will consume less computational resources in both the training and testing phases. Related to this in some settings, and particularly relevant to healthcare, is reducing the number of external sensors required which will reduce cost and may increase the comfort of the patient.

The last reason, again particularly relevant for healthcare, is interpretability of results. The ability to present interpretable results using a minimal set of features is likely to be crucial to convincing clinicians of the merit of the type of systems presented in this thesis. This in turn has implications for the development of relevant features, which also need to be understandable to the psychiatric community.

Although there are many feature selection methods available, the following sections will describe two methods that are used in this thesis.

3.6.1 The Lasso

The least absolute shrinkage and selection operator (the Lasso) is a feature selection method developed by Tibshirani (1996) that attempts to optimise the subset of features presented to any model that is based on a standard linear model of the form

$$\hat{y}_i = f(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n) \tag{3.67}$$

$$= f(\beta_0 + \boldsymbol{\beta}^\top \mathbf{z}) \tag{3.68}$$

where $\mathbf{z} = [z_1, \ldots, z_p]^\top$ is a vector of the p extracted features; $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top$ is a vector of feature coefficients; and β_0 is an intercept. The Lasso attempts to minimize $\sum_i^n (y_i - \hat{y}_i)^2$ subject to the constraint that $\|\boldsymbol{\beta}\|_1 \leq r$. The constraint parameter r acts to restricts the values of the components of $\boldsymbol{\beta}$. When r is large it will have no effect and the solution to β is just a standard linear least squares regression of equation (3.67). However as $r \to 0$ the solutions are shrunken versions of the least squares estimates to the point where some (or all) of the components of $\boldsymbol{\beta}$ are 0.

The Lasso can alternatively be expressed in its Lagrangian dual form as

$$\underset{\beta_0, \boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_i^n \left(y_i - f(\beta_0 + \boldsymbol{\beta}^\top \mathbf{z}_i) \right)^2 + \lambda \sum_j^p |\beta_j| \right\} \tag{3.69}$$

where the regularisation is now controlled by the λ parameter (Hastie et al., 2009, 2015; Osborne et al., 2000). The function of the λ parameter is inverted from the r parameter in that small values of λ have no effect on the solution while large values of λ cause the values of $\boldsymbol{\beta}$ to shrink.

The Lasso constrained equation can be solved using various methods including standard quadratic programming techniques, but there are also other methods that can exploit the particular properties of the problem to solve it more efficiently. One such method is the least angle regression (LARS) algorithm described by Efron et al. (2004), and more recently pathwise coordinate descent by Friedman et al. (2007).

The Lasso can be used for general feature selection by increasing r from 0 until all features are included in the model, i.e. $\beta_i \neq 0 \; \forall \; i$. The order in which the elements become non-zero is the order of the selected features. However, the Lasso is known to perform poorly where there are strong correlations in the features (Zou & Hastie, 2005). This is certainly the case here with the nature of the extracted features and the expected correlation between different data subsets. It is also known that variables that appear correlated may still be useful when used together (Guyon & Elisseeff, 2003) and these may be excluded by the Lasso.

3.6.2 Wrapper feature selection methods

Although the Lasso is well suited to linear regression problems (and derivatives thereof, such as logistic regression and other generalised linear models), it may not provide optimal results when used with a non-linear classifier. Feature selection methods that use the final classifier as part of the feature selection process are known collectively as *wrapper* methods (Guyon & Elisseeff, 2003). These methods can make use of the particular properties of the classifier being used, as well as subtle properties of the individual features that may be missed with more general methods such as the Lasso.

3.6.2.1 Greedy forward selection

For feature selection in chapter 6, a simple wrapper method, often known as *greedy forward selection*, was used. Greedy forward selection is described in algorithm 3.3 and selects features based on an appropriate metric of a classification result of each feature individually, then each remaining feature with the first selected feature etc.

Algorithm 3.3 Greedy forward feature selection method.

This performs classification using each available feature, choosing the best one, then again performs classification on each of the remaining features with the previously selected feature(s) and so on.

1: **data:** $\mathcal{F} = \{f_1, \ldots, f_p\}$ set of p raw model features.
2: **result:** $\mathcal{F}^* = \{f_1^*, \ldots, f_p^*\}$ set of model features ordered by performance.
3: **begin**
4: $\quad \mathcal{F}^* \leftarrow \{\}$
5: $\quad$ **while** $|\mathcal{F}| > 0$ **do**
6: $\quad\quad f_{\text{best}} \leftarrow \{\}$
7: $\quad\quad f_{\text{score}} \leftarrow 0$
8: $\quad\quad$ **for** $f \in \mathcal{F}$ **do** ▷ Loop through all unselected features.
9: $\quad\quad\quad \mathcal{F}' \leftarrow \mathcal{F}^* \cup \{f\}$
10: $\quad\quad\quad f'_{\text{score}} \leftarrow \text{classify}(\mathcal{F}')$ ▷ Perform leave-one-participant-out
11: $\quad\quad\quad$ **if** $f'_{\text{score}} > f_{\text{score}}$ **then** classification using previously se-
12: $\quad\quad\quad\quad f_{\text{best}} \leftarrow f$ lected features and current feature,
13: $\quad\quad\quad\quad f_{\text{score}} \leftarrow f'_{\text{score}}$ returning relevant metric.
14: $\quad\quad\quad$ **end if**
15: $\quad\quad$ **end for**
16: $\quad\quad \mathcal{F}^* \leftarrow \mathcal{F}^* \cup \{f_{\text{best}}\}$ ▷ Append f_{best} to $\mathcal{F}^*$.
17: $\quad\quad \mathcal{F} \leftarrow \mathcal{F} \setminus \{f_{\text{best}}\}$ ▷ Remove f_{best} from $\mathcal{F}$.
18: $\quad$ **end while**
19: **end**

3.6.3 Discussion

This chapter has introduced a number of different machine learning techniques, all of which are applied in some way through the following chapters. These techniques ranged from classical models of regression and classification to more advanced non-parametric methods. The classification models are all applied to the identification of depression and differentiation of mental health conditions in section 6.3 with full results given in sections C.1 and C.2 of appendix C. Standard linear regression and generalised linear models are used to estimate depression levels in section 6.4 with full results given in section C.3. The non-parametric Dirichlet process model, combined with a Bayesian version of the Lasso model, is used to develop the grouped personalised model in chapter 7. The model evaluation and validation techniques are used throughout the thesis.

This is not intended to be an exhaustive survey of literature in this broad area. The models applied have been chosen for a number of reasons as identified in the relevant parts of the thesis. While there are many more classification and regression models available (Murphy, 2012; Hastie et al., 2015), both frequentist and Bayesian, and with varying levels of complexity and non-linearity, it is not obvious that more advanced models will provide superior results in this application. This is discussed in further detail in section 6.5.

One notably absent technique from this chapter is that of artificial neural networks, which are gaining increasing popularity as an automated feature extraction and classification method. Neural networks will be discussed in greater detail in section 5.12.

Chapter 4

Data Description

Data for this thesis were collected as part of the Automated Monitoring of Symptom Severity (AMoSS) study, run in collaboration with the Department of Psychiatry at the University of Oxford.[1] The AMoSS study was set up to investigate how symptom severity in mental health relates to objective metrics derived from sources such as physical activity and movement, social interaction or physiology. Data were collected from patients diagnosed with bipolar disorder (BD); patients diagnosed with borderline personality disorder (BPD); and healthy controls (HC) without any pre-existing mental health diagnosis. All participants were screened by a psychiatrist on entry into the study.

The AMoSS study recruited a total of 140 participants between January 2014 and November 2015, of which 54 were diagnosed with BD; 33 were diagnosed with BPD; and 53 were healthy controls. Participants were initially recruited for 12 weeks with the option to remain in the study for up to a year. After the end of the year, participants could further opt to continue providing data.

This chapter first describes the full range of data collected in the AMoSS study. The geolocation data and questionnaires used throughout the rest of this thesis are then described in detail and characterised. Finally, some of the related analyses of the AMoSS data are summarised.

[1] The AMoSS study was approved by the local NHS research ethics committee (REC reference 13/EE/0288) and all participants provided written informed consent.

4.1 Data collection methodology

Throughout their participation in the study, participants used a custom Android app in conjunction with external devices to collect a full range of behavioural; physiological; and lifestyle data hypothesised to be relevant to mental state.

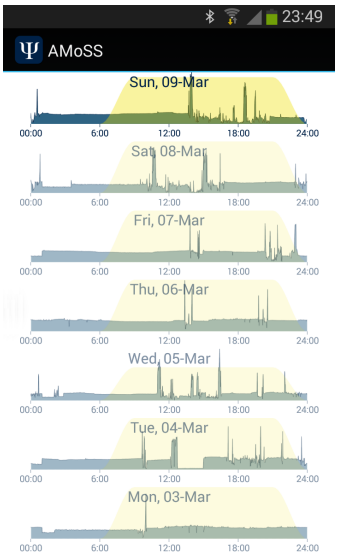
4.1.1 Data collection app

A custom open source[2] Android data-collection app was developed to allow remote real-time monitoring of participant behaviour and physiology. The app was either pre-installed on a Samsung Galaxy SIII or S4 smartphone provided to participants, or could be used on their own phone. The app recorded the following data:

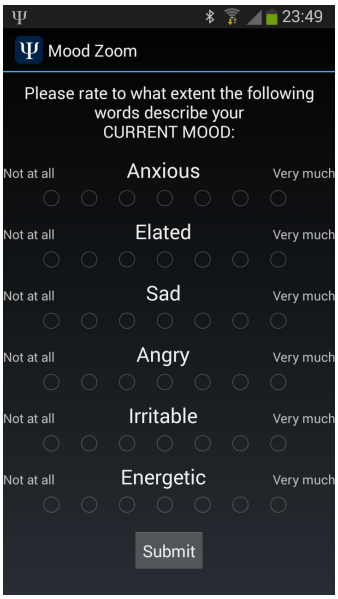
1. Activity level through accelerometry
2. Ambient light level
3. Geographic location
4. Social network activity measured by
    a) phone calls (time and duration)
    b) text messages (time and length)
5. A daily self-reported mood questionnaire (known as *Mood Zoom*, described in section 4.4.2)
6. Participant physiology (blood pressure and temperature can be manually entered or read from Bluetooth connected devices; continuous heart rate and related signals derived from Bluetooth ECG or pulse oximetry devices)

Screenshots of the app are shown in figure 4.1. Participants had full control over the data being collected and were able to disable collection of any modality at any time. Data collected by the app were securely uploaded to a remote server for analysis. An overview of the architecture of the data collection app and server upload mechanism is given by Osipov (2016).

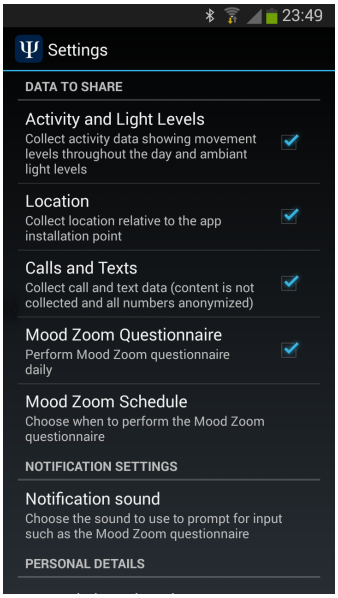
[2] The source code for the data collection app is available from http://amossstudy.bitbucket.io/.



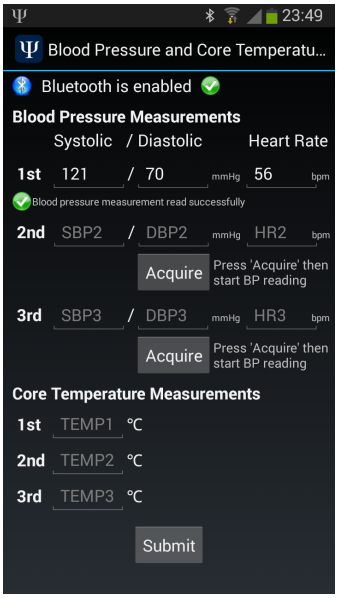
(a) Weekly actigraphy levels.



(b) Daily mood survey.



(c) Data collection settings.



(d) Physiological data entry.

Figure 4.1: AMoSS study data collection app screen-shots. The main screen (a) shows an overview of the recorded actigraphy levels over the last week. The daily mood questionnaire screen (b) was used to collect the simple mood monitoring questionnaire (Mood Zoom) on a daily basis, or 10 times daily during the high intensity monitoring week described in section 4.1.3. The settings screen (c) gave participants full control over the data being collected. The physiological data entry screen (d) was used during the high intensity monitoring week, described in section 4.1.3, to collect physiological data either through connection to external Bluetooth devices or by manual entry.

4.1.2 External devices

In addition to the data recorded from the phone sensors, participants were also provided with a Fitbit One™ wireless activity and sleep tracker[3] and a GENEActiv Original wrist-worn accelerometer.[4] If the phone with the app was not carried on the person, the Fitbit and GENEActiv devices could provide a more complete measure. The Fitbit device collected step-count data as a proxy for activity levels, and the GENEActiv device was configured to provide triaxial accelerometry at 25 Hz which enabled continuous data collection for up to 1 month. The GENEActiv device also provided skin temperature and ambient light level data.

4.1.3 High intensity monitoring week

Additional data were also collected during a one-week "high intensity" monitoring period, usually in the first 12 weeks of participation. During this week, participants provided thrice-daily blood pressure and core body temperature readings through the Android app.

Participants also wore a Shimmer3 portable electrocardiogram (ECG) monitor,[5] and an adhesive Proteus® Wearable Sensing Platform[6] patch. The Shimmer3 device provided full ECG data for around three days. The Proteus patch provided average heart rate and skin temperature every 5 minutes; and accelerometry every minute for the whole week. The Mood Zoom questionnaire (described in section 4.4.2) was also administered 10 times daily instead of once daily during the rest of the study.

[3] See http://www.fitbit.com/uk/one

[4] See http://www.geneactiv.org/actigraphy/geneactiv-original/

[5] see https://www.shimmersensing.com/products/shimmer3-ecg-sensor

[6] see https://www.proteus.com/, although the patch used is no longer available.

4.2 Participant enrolment

Figure 4.2 on the next page shows the profile of the app data collected from study participants from the start of the study until January 2017. Each row represents a participant, and the bars indicate the times when app data were collected (specifically geolocation). The full length of the bar indicates the time that the participant was enrolled in the study and dark grey sections are times when no geolocation data were received from the participant. The end of the bar is determined by either when the participant formally left the study (either at; before; or after one year of enrolment), or the time at which the last data were received from them (if they opted to stay in the study for longer than a year). Missing data within this period may be because participants opted not to provide geolocation data; because they did not use the phone and kept it switched off; or because network issues prevented the data being uploaded. Note that figure 4.2 simply indicates that a data file was uploaded for the date in question, and quality or quantity of the data is not considered here – this will be discussed in detail in section 5.7.

It can be seen that while most participants participated in the study for up to 12 months, several opted to continue for longer. Details of participant demographics will be given in section 5.7 for participants who provided sufficient geolocation data for analysis.

4.3 Geographic location data

This thesis focusses on the geographic location data collected through the android app, which will be discussed in greater detail in this section.

4.3.1 Sources of geographic location data

Location data recorded on Android are a fusion of different sources of location data. Locations read from the global positioning system (GPS) provide the most accurate data, but GPS is often not available indoors or in dense urban areas (Chen et al., 2006) and it has high power requirements (Mayora et al., 2013).

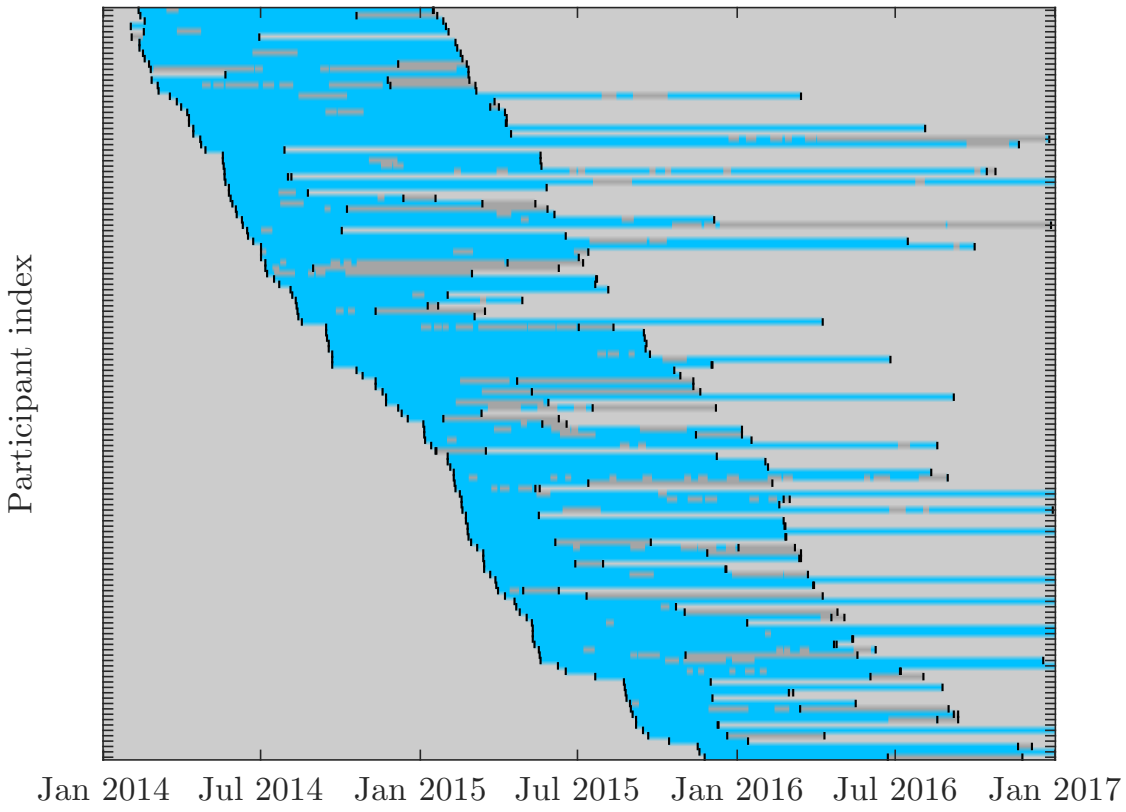


Figure 4.2: Profile of phone app data received for all participants enrolled in the AMoSS study.

Mobile cell tower triangulation (LaMarca et al., 2005; Chen et al., 2006) uses estimates of the distances from the handset to nearby cellphone towers to estimate the current location. This is generally the least accurate source of location data but because the phone is constantly communicating with the cellphone towers anyway it is very power efficient.

The final source of location data uses nearby wireless network access points to determine an approximate location. When requesting the current location from the Android operating system, the location returned can originate from any of the above sources depending on the resource availability on the device at the time. The different sources can also sometimes contradict (Yang et al., 2010).

Additionally, both cellphone tower and wireless network triangulation require a database of the known locations of the cellphone towers or the wireless network access points (Schmidt-Dannert, 2010). This can lead to problems with accuracy (Yang et al., 2010) especially where the database is out of date. This is especially pertinent for wireless access points that may have been moved to new locations while still being listed in the database in their old locations. On Android, location triangulation using cellphone towers or wireless network access points is performed by sending the details of the cellphone tower(s) and/or wireless access point(s) in range of the mobile device to Google over the Internet using the Google Web Services Geolocation API (Fasel, 2011; Feng, 2013; Google, 2018). If the user does not have access to the Internet then it may not be possible for Android to determine the location.

4.3.2 App update

Technical issues with the initial implementation of the app severely reduced the accuracy of the geolocation data collected at the start of the study. An update to the app was released on 1st May 2015 to correct these issues and the data used in this thesis were collected using this updated version. The data collection profile in figure 4.2 can be updated to show when participants upgraded, as shown in figure 4.3 on the following page. The dark blue sections indicate data collected with the updated app and the red crosses indicate the time of upgrade.

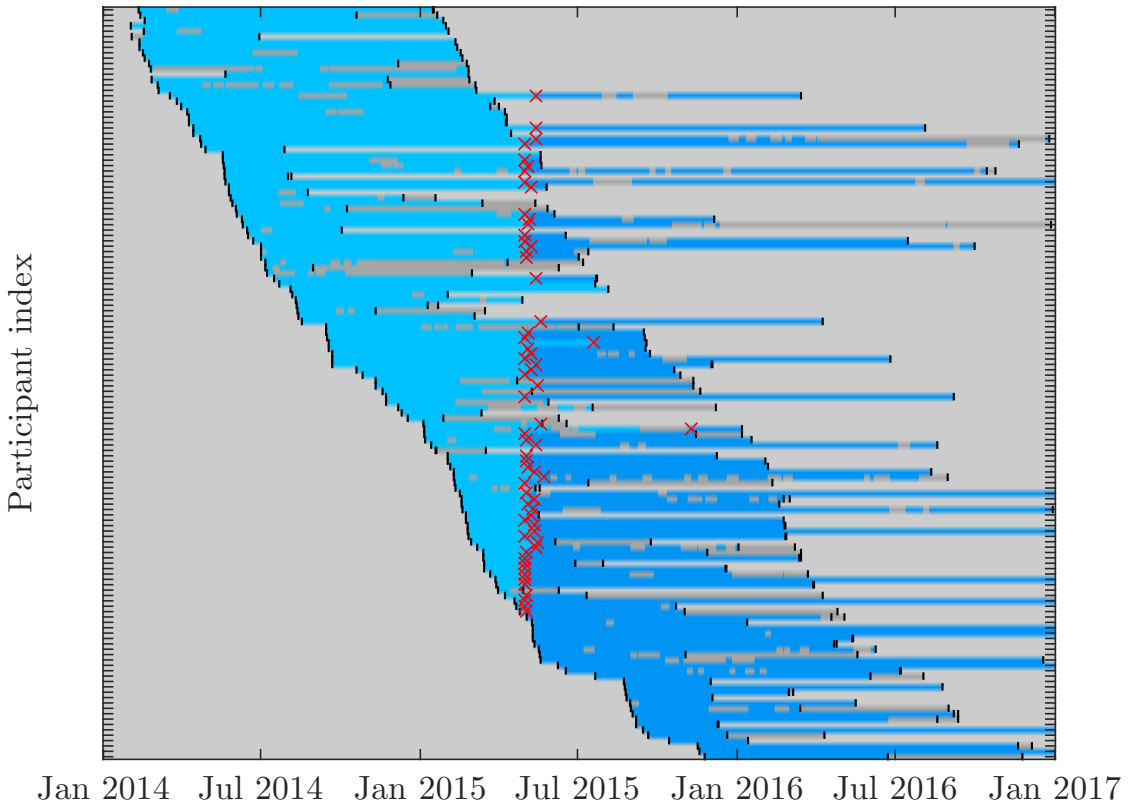


Figure 4.3: Profile of phone app data received for all participants enrolled in the AMoSS study before and after the app update described in section 4.3.2. Light blue bars show the old app and dark blue bars show the updated app. The red cross shows the date that each participant upgraded their app.

4.3.3 Anonymisation of geographic location data

Geographic location data is highly sensitive to collect, and has serious privacy implications (Mayora et al., 2013). For this reason, all recorded data were anonymised on the phone (before being transmitted to our servers) by making the location coordinates relative to a random location in the world, randomly selected when the app is first installed.[7]

4.3.4 Raw geographic location data characteristics

This section describes basic characteristics of the raw recorded geolocation data before any preprocessing is performed. The characteristics of the data after preprocessing are described in section 5.6.

[7] This is a relatively effective method of protecting the privacy of study participants, but it causes a number of potential problems for analysis.

The first problem is due to the variable nature of longitude throughout the world (the length of a unit of latitude in km also varies somewhat depending on the latitude – being approximately 110.5 km at the equator and 111.5 km on the Arctic circle – but the effect is much less significant than the effect on longitude). For example, a unit of longitude at the equator (0°N) approximately 110.6 km; while at 66°N (approximately on the Arctic circle) a unit of longitude is approximately 45.4 km. The process of anonymisation shifts the origin (origin as used in this footnote refers to the intersection between the equator (0°N) and the prime meridian (0°W)) to some unknown location, which means that it is no longer possible to accurately calculate distances in km using the anonymised data. For convenience, the global home location was calculated as described in section 5.2 and all anonymised data points were subtracted from this location to leave the home location at the origin. These transformed coordinates were used for all further calculations. Because the trial was run around Oxford, the home location of all participants was assumed to be in Oxford (51.75°N 1.26°W) where a unit of latitude is 111.3 km and a unit of longitude is 69.1 km. All distances between coordinates were calculated using these dimensions. This means that distances calculated between points close to the home location of the participant are likely to be relatively accurate while distances further from home are not.

Another potential problem presented by the anonymisation process and subsequent transformation relative to the home location is the loss of knowledge of the location of the 180^{th} meridian. This makes calculating relative distances around the 180^{th} meridian difficult, which could in turn affect the stationary location clustering methods described in section 5.8.

Both of the problems discussed above are accepted as limitations of the work presented here and would require more careful consideration before applying these techniques in practice. However, because the main focus of this work is on the behaviours of individuals in their every-day activity – which are likely to be close to home – these aspects are unlikely to have a significant impact.

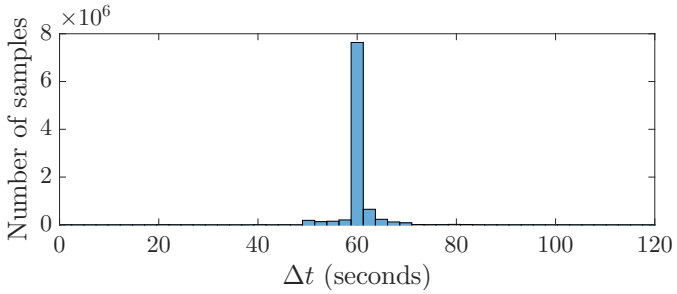


Figure 4.4: Time between consecutive samples of geographic location data (Δt) for participants using the updated app described in section 4.3.2.

4.3.4.1 Sample rate

The raw geolocation data are unevenly sampled as provided to the app by the Android operating system. The actual sample rate of the acquired data is best described by the time between consecutive recorded data points, denoted Δt. This is shown in figure 4.4 above, which shows Δt for all data points from the 88 participants who installed the upgraded app (see table 5.2 for a breakdown of the participants in the study). The mode of the distribution is at 60 s between samples, implying a mean sample rate of 1 sample per minute.

4.3.4.2 Data coverage

The raw data coverage is given as the proportion of each day that has geolocation data recorded. Each recorded data point is defined to be valid since the previous data point, up to a maximum of 5 min, which provides a reasonable estimate of the current location given the expected rate of change in the geographic location of individuals. The distribution of coverage for all participants who installed the updated app and provided data is shown in figure 4.5. For the purposes of this figure, each participant is assumed to be providing data between the first and last day where any data were provided. The annotations show that over 35 % of days from participants have ≥ 95 % data coverage. However, for almost 23 % of days participants provided no data at all.[8]

[8] Note that the estimate for where the data coverage is exactly 0 % is likely to underestimate the true value because days without any data before the first data point or after the last data point are not included. However, for the purposes of this figure, this estimate is deemed acceptable.

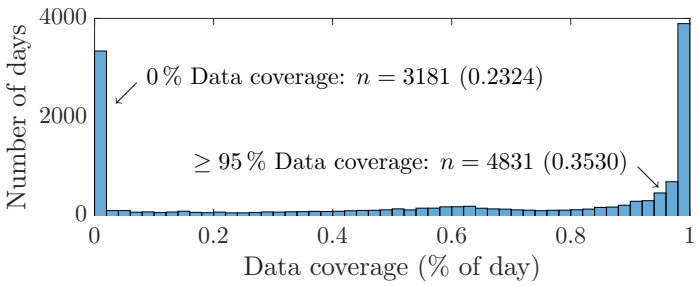


Figure 4.5: Data coverage of the raw geographic location data for participants using the updated app described in section 4.3.2.

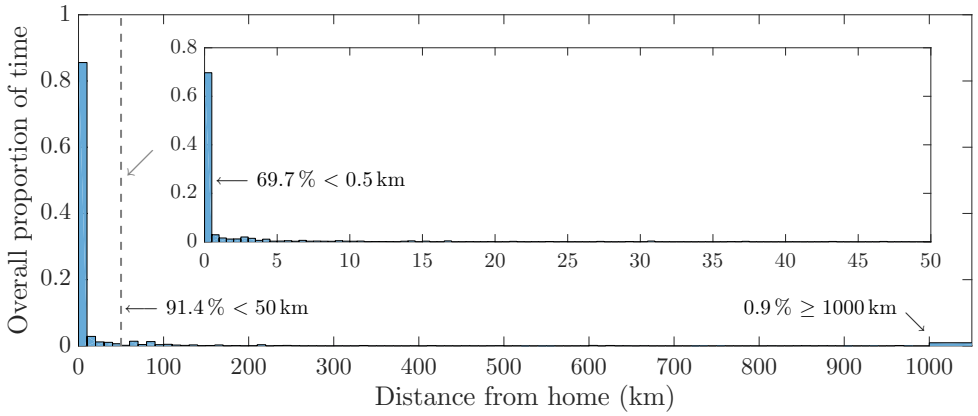


Figure 4.6: Overall distribution of the time participants spent at locations within given distances from home when using the updated app described in section 4.3.2.

4.3.4.3 Data properties

Examples of the raw data recorded are shown later in figure 5.1, but to give an indication of the overall distribution of the raw geolocation data, figure 4.6 above shows the proportion of time that participants spent at locations within given Euclidean distances from their assumed home location. The assumed home location of participants is calculated using the method described in section 5.2.

Figure 4.6 shows that, as expected, participants spend most of their time at or close to home. Overall, participants spent 91 % of their time < 50 km from home, and almost 70 % of their time < 500 m. Note that the distributions presented here are from the raw geolocation data recorded by the app – as demonstrated in figure 4.5, there are high levels of missing data that may affect the results.

4.4 Questionnaires

In order to track psychological state, participants were asked to complete several clinically validated psychiatric questionnaires on a weekly basis as well as a compact mood monitoring questionnaire daily.

4.4.1 Clinically validated psychiatric questionnaires

The clinically validated psychiatric questionnaires were performed through the True Colours monitoring system (Bopp et al., 2010; Malik et al., 2011) which is integrated with the NHS infrastructure in Oxford. True Colours enables completion of pre-defined surveys via either email or text message. The following questionnaires are collected on a weekly basis.

4.4.1.1 QIDS depression questionnaire

The 16-item quick inventory of depressive symptomatology (QIDS-SR_{16} or just QIDS) (Rush et al., 2003) questionnaire is a 16 question diagnostic and monitoring tool for depression covering all nine DSM-IV symptom criterion domains of a) sad mood; b) concentration; c) self-criticism; d) suicidal ideation; e) interest; f) energy/fatigue; g) sleep disturbance (initial, middle, and late insomnia or hyper-somnia); h) decrease/increase in appetite/weight; and i) psychomotor agitation/retardation.

QIDS-SR_{16} returns a total score between 0 (no symptoms) and 27 (severe symptoms), with defined thresholds of a) no depression: 0–5; b) mild depression: 6–10; c) moderate depression: 11–15; d) severe depression: ≥ 16.

There are other questionnaires used to quantify depression such as the Patient Health Questionnaire-9 (PHQ-9) (Kroenke et al., 2001), which has the same range as QIDS-SR_{16} (0–27) but a threshold of 10 for moderate depression. To put these scores in context, a meta-analysis of 14 studies by Gilbody et al. (2007) has shown that PHQ-9 ≥ 10 has pooled sensitivity of 80 % and specificity of 92 % in diagnosing major depressive disorder (MDD). Higher thresholds have also been shown to give improved classification accuracy for MDD (Löwe et al., 2004). Given that the diagnostic criteria

for episodes in MDD are the same as depressive episodes in BD (see section 2.1.1.1), this indicates that these questionnaires are also a suitable indicator of depression in BD. While it is known that there are symptomatic differences between depressive episodes in MDD and BD (Moreno et al., 2012), this is an open area of research outside the scope of this thesis.

4.4.1.2 ASRM mania questionnaire

The Altman self-rating mania scale (Altman et al., 1997) (ASRM) questionnaire is a five question tool designed to measure the presence and severity of manic symptoms in accordance with DSM-IV criteria. The five questions assess increased levels of happy or cheerful mood; increased self-confidence; reduced sleep; increased talking; and increased activity levels over the past week. Each item is scored on a five-point scale from 0 (like usual) to 4 (much more than usual), with total scores ranging from 0 to 20. For detection of episodes of mania, a threshold score of 5.5 is recommended as it has shown an optimal combination of sensitivity and specificity (85 % and 86 %, respectively) (C. J. Miller et al., 2009).

4.4.1.3 GAD-7 anxiety questionnaire

The generalised anxiety disorder (GAD-7) (Spitzer et al., 2006) questionnaire is a brief 7 question tool designed to assess symptoms of generalised anxiety disorder. The survey was developed by constructing a 13 item questionnaire based on the DSM-IV symptom criteria for GAD and then selecting the 7 questions that provided the best diagnostic criterion validity. Spitzer et al. showed that a threshold score of 10 provided optimal performance for detecting generalised anxiety disorder with sensitivity of 89 % and specificity of 82 %.

4.4.1.4 EQ-5D quality of life questionnaire

The EuroQol EQ-5D (EuroQol Group, 1990) is a 6 question quality of life questionnaire providing a quality of life score based public ratings of the possible responses.

4.4.2 Phone-based mood monitoring questionnaire

A custom questionnaire, known as *Mood Zoom*, was developed to assess mood states in participants at a higher frequency than the clinically validated psychiatric questionnaires, and in a simpler and more direct way. Mood Zoom asks participants to rate 6 aspects of their mood on a Likert scale (1–7). The mood descriptors used were how a) anxious; b) elated; c) sad; d) angry; e) irritable; and f) energetic participants felt. Participants were asked to assess how well these descriptors described their mood, from 'not at all' to 'very much', on the custom app. A screenshot of the Mood Zoom questionnaire implemented on the app was shown in figure 4.1(b).

Throughout the study, participants were asked to complete the Mood Zoom questionnaire on a daily basis, with a prompt shown at a user definable time in the evening. This daily questionnaire would ask participants about their mood through the entire day.

4.4.3 Limitations

As mentioned in section 2.2.1, all the questionnaires described above, including the clinically validated questionnaires, are fundamentally only subjective proxies of mental state. Limitations include the subjectivity of the answers and the different baselines that different individuals have. In chapters 6 and 7, models of mental state are built with reference to the patient-reported QIDS questionnaire.[9] It is therefore an accepted limitation that comparing results to such subjective proxies may not represent the true mental state of the individual. The QIDS questionnaire is however a well-validated model of depression in bipolar and unipolar depression.

The value of estimating mental state from objective markers is unlikely to be the absolute accuracy of the estimated value, but rather the relative changes seen over time. While single questionnaire responses for an individual may have limited value, the long-term trends for individuals across different questionnaires have been shown

[9] The approach of developing and comparing models of mental state to subjective patient-reported questionnaires is in common with many previous studies of objective markers of mental health. Some studies use clinician-reported measures, but these are still fundamentally subjective proxies and are usually not available at a high sample rate.

by Johns et al. (2013) to correlate well. This indicates that the personalised models developed in chapter 7 can provide value by finding models that match the general trend of depression for individuals rather than just absolute values. Additionally, as more longitudinal data are available for analysis, the properties of the digital and behavioural phenotypes found may in turn help to inform our understanding of mental illness.

4.4.4 Calendar week data labelling

Although participants are prompted by text message or email to complete the clinically validated questionnaires on a weekly basis, participants sometimes forget or are unable to respond. The results in this thesis are based on analysis of calendar weeks of data, from Monday to Sunday. Within each calendar week, there may be no questionnaire responses provided, a single questionnaire response, or there may be multiple responses. For this reason, the method in algorithm 4.1 was developed to label calendar weeks using either a response within 3.5 d either side of the week, or an average of the linear interpolation between responses on either side of (and during) the week. Some examples of the data labelling are shown in figure 4.7.

→

Algorithm 4.1 Labelling of calendar weeks.

Each calendar week of data is labelled using either a questionnaire response within 3.5 d either side of the week, or the mean of the linear interpolation between responses on either side of (and during) the week.

data: $\mathcal{T} = \{t_1, \ldots, t_n\}$ times of n completed QIDS scores.
$Q(t)$ completed QIDS score at time t.
w_{start} the start date of the week to label.
w_{end} the end date of the week to label.
result: label to use for the week.

begin
  $\mathcal{T}' \leftarrow \{t_i \in \mathcal{T} : (w_{start} - 3.5\,\text{d}) \leq t_i \leq (w_{end} + 3.5\,\text{d})\}$
  $\mathcal{T}^* \leftarrow \{t_i \in \mathcal{T} : (w_{start} - 7\,\text{d}) \leq t_i \leq (w_{end} + 7\,\text{d})\}$
  if $(\mathcal{T}' = \{\})$ **and** $(\mathcal{T}^* = \{\})$ **then**
    return null
  else if $|\mathcal{T}'| = 1$ **then**
    return $Q(\mathcal{T}')$
  else
    Compute linear interpolation of QIDS responses $q(\tau)$ for days $w_{start} \leq \tau \leq w_{end}$ between QIDS scores $Q(t^*)\ \forall\ t^* \in \mathcal{T}^*$.
    return $\sum q(\tau_i)/|\tau|$
  end if
end

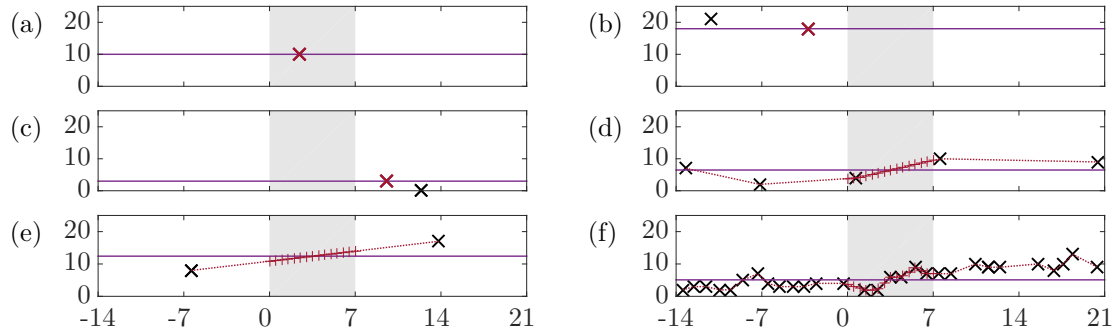


Figure 4.7: Examples of weeks labelled using algorithm 4.1. The grey section shows the week to be labelled; the black crosses show all available questionnaire responses (in this case QIDS scores); and the purple line indicates the label applied. (a-c) demonstrate using the nearest single point during, before and after the week being labelled respectively, and (d-f) show using the mean of the linear interpolation between data points surrounding or within the week being labelled, where the dotted red line shows the linear interpolation and the red crosses indicate the points used when calculating the mean score.

4.5 Broader analysis of AMoSS data

As shown in the earlier sections of this chapter, many source of data were collected from participants in the AMoSS study. This thesis concerns the analysis of the geolocation data collected, and compares it to the depression scores in the QIDS questionnaire, but this is only a subset of the data collected, and several of the other data sources have also been analysed by various individuals.

The first modality to be analysed was the accelerometry data by Osipov (2016). A wide range of features were first extracted from the GENEActiv wrist-worn devices and phone accelerometry data (although some features could not be calculated from the phone accelerometry data). The features extracted from the GENEActiv devices were used to classify first the diagnosis of the participant, and then the mood state (euthymic; depressed; or manic) in the HC and BD participants. The phone accelerometry data were used to build 'fully personalised' models (in the terminology of chapter 7) relating the objective features to the three main clinically validated psychiatric questionnaires collected. These personalised models used cross-validation over all the available data from each participant, which will be discussed in greater detail in section 7.11.

Separate work is currently in progress developing a model of circadian regularity using the phone activity data.

More recently, Carr et al. (2018a) analysed the accelerometry and cardiovascular data from the Proteus patch worn by participants during the high intensity monitoring week described in section 4.1.3. Carr et al. developed an elegant method for quantifying the variability of diurnal regularity in the objective data from individuals and correlated this with variability in their Mood Zoom questionnaire responses. This demonstrated that the different measures of variability were highly correlated, and that BPD participants had particularly high levels of variability in all measures. Carr et al. (2018b) extended this technique to demonstrate that there is also diurnal regularity in the sub-daily sampled Mood Zoom questionnaire during the high intensity week, and that this also correlates with the physiological data in some cases.

The questionnaires, especially the Mood Zoom questionnaire, have had considerable research interest. Bilderbeck et al. (2015) presented a preliminary analysis, quantifying the differences between participant groups visible in figure 5.7 as well as differences in variability in responses to the Mood Zoom and clinically validated questionnaires. Variability in mood has been a theme through much of the analyses of the questionnaires, and Bilderbeck et al. demonstrated that the BPD patients tend to have higher variability in the daily mood scores than the BD patients. This was extended by Saunders et al. (2016; 2017a) who proposed that mood instability measured through daily mood questionnaires may provide an additional tool to distinguish BD and BPD patients when the diagnosis is unclear from the symptoms presented.

Tsanas et al. (2016) also analysed the Mood Zoom questionnaires and found that the 6 moods in the questionnaire could effectively be reduced to a 'positive' mood component; a 'negative' mood component; and a third component most strongly reflecting irritability. Combined, these three latent components explained 85 % of the variability in the Mood Zoom responses. The 'negative' component was then shown to correlate strongly with the QIDS and GAD-7 clinically validated questionnaires. Variability in the clinically validated questionnaires and the latent components of the Mood Zoom questionnaire was then shown to be significantly different between the patient groups in the study and the healthy controls. Tsanas et al. (2017) extended this analysis and applied a sparse version of PCA with similar results, while also demonstrating broad consistency in the latent components between patient groups.

Qualitative studies have also been performed through participant interviews to investigate how the type of monitoring presented here is seen from the perspective of the patient. Saunders et al. (2015a, 2017b) found that monitoring was generally well received and improved patients' understanding of their conditions. It was also colloquially found that monitoring promoted positive behavioural change, but it remains to be seen whether behavioural changes persist long-term. While these effects may impact the data collected, the general principle that just the act of monitoring can positively influence mental state is potentially beneficial.

4.6 AMoSS data limitations

The AMoSS data set is one of the largest longitudinal datasets collected from mental patients in the community. A wide range of data sources were collected, and many participants provided data for over a year. Notwithstanding the limitations of self-reported questionnaires as a proxy of mental health discussed in sections 2.2.1 and 4.4.3, adherence with the questionnaires in the AMoSS study was generally good. This was demonstrated by Tsanas et al. (2016) and can be seen from figure 5.6 in the next chapter where relatively few weeks of valid geolocation data could not be analysed due to lack of questionnaire data.

Nevertheless, the dataset is not without limitations, some of which will be explored here, and in greater detail in the following chapters.

Without additional processing, there are large amounts of missing data in both the activity data (shown by Osipov, 2016) and the geolocation data (shown in section 4.3.4.2). While the data preprocessing techniques described in the next chapter enabled utilisation of much of the data that may otherwise have been unusable, it would naturally have been better not to have had such high levels of missing data in the first place. Additionally, there were high levels of noise in the geolocation data recorded as discussed in section 5.3. When providing location coordinates, Android also provides an estimate of the accuracy of the location (essentially an estimate of the standard deviation). This was not recorded on the custom app, but having access to this might simplify the noise removal process as discussed in greater detail in section 8.3.1.

Technology is also continually improving, and while the phones used by AMoSS participants (mainly Samsung Galaxy SIII or S4 smartphone) were mid to high-end at the time of the study, newer devices naturally have increased processing capacity which should improve the quality of the data recorded if the trial were run again now.

While the use of smartphones as data collection devices is convenient due to their ubiquity and power, the data recorded are inherently representative of the device itself, rather than the user. Some users will always carry their phone with them and

for these individuals the data recorded will be highly accurate. However, some people might leave them in their bag, or on their desk, meaning that important behavioural data are lost. In this regard, geolocation is less likely to be affected than activity since users are more likely to take their phones with them when travelling over larger distances.

However, increased utility of smartwatches for data collection could dramatically change the type and accuracy of data that can be collected. Smartwatches also often have more advanced physiological sensors that can provide behavioural and physiological data in ways that are currently challenging.

The physiological data analysed by Carr et al. (2018a, 2018b) shows the ability of physiological data to quantify mental state. But the devices utilised on the AMoSS study, while powerful, were not suitable for long-term monitoring. Longitudinal physiological data would likely have important predictive power as a personalised model of mental state due to how intricately linked the cardiovascular and cognitive systems are (Thayer & Lane, 2009), but this cannot be demonstrated with the data available.

4.6.1 Discussion

This chapter has outlined the AMoSS study run from 2014, which collected a wide range of data from both healthy controls and patients with mental health diagnoses. The main focus was on the geolocation data that will be analysed in detail in the following chapters, but a summary of the analysis of the other data collected was also provided.

Geolocation was demonstrated previously in section 2.2.4.3 to be a promising candidate as an objective marker of mental state, and this is one of the first studies to collect this level of longitudinal geolocation data from patients in the community. However, the quality of the data was shown to be variable. Significant levels of preprocessing will be required to improve the data coverage and reduce noise to enable extraction of features that accurately reflect the behaviour of the individual. This is described in detail in the next chapter.

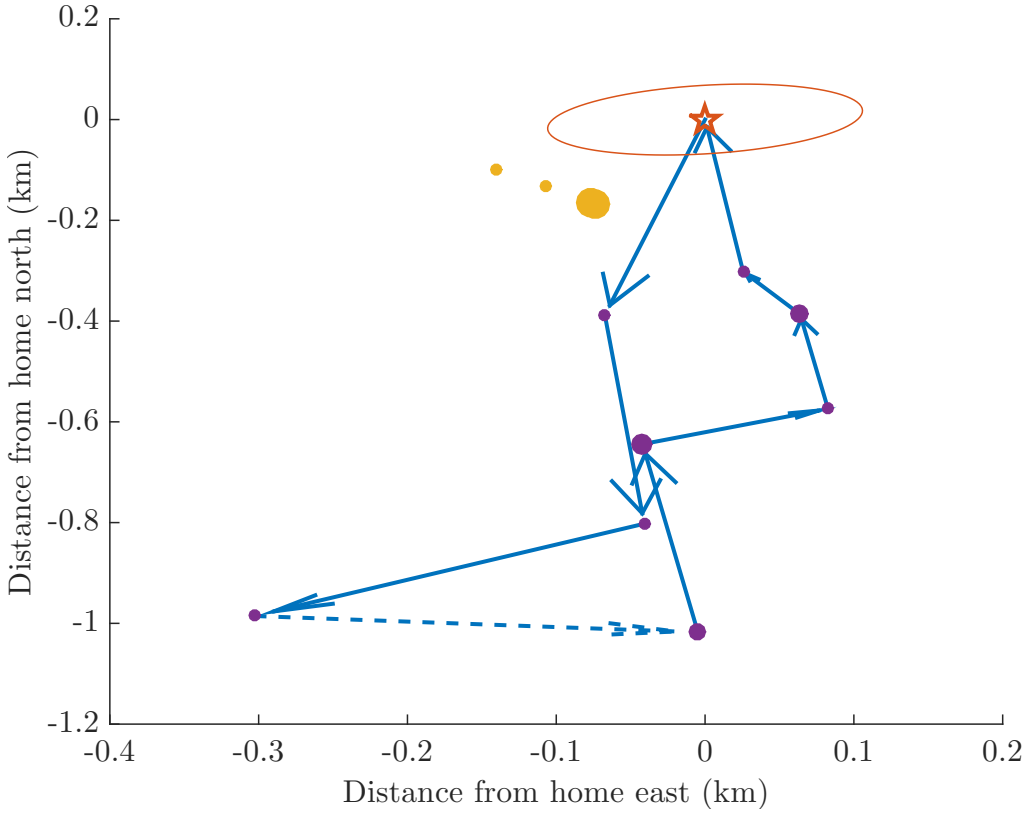
Chapter 5

Using Geolocation Data as an Indicator of Mental Health State

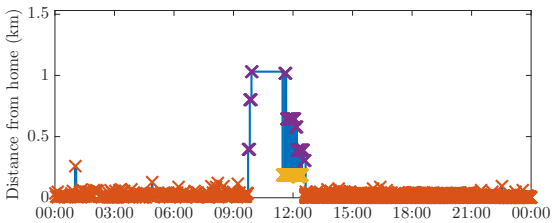
The symptoms of both borderline personality disorder and episodes in bipolar disorder contain characteristics that may be visible in the geographic movements of individuals. This was demonstrated by the diagnostic criteria in section 2.1 and the previous work with this type of data outlined in section 2.2.4.3. Geographic movement behaviour may be captured through the geolocation data recordings from participants in the AMoSS study described in the previous chapter. However, the previous chapter also outlines shortcomings in the data coverage and quality that need to be addressed before accurate features characterising behaviour can be extracted. This chapter describes the processing of the recorded geolocation data and the extraction of features relevant to mental health.

5.1 Android-recorded geolocation data

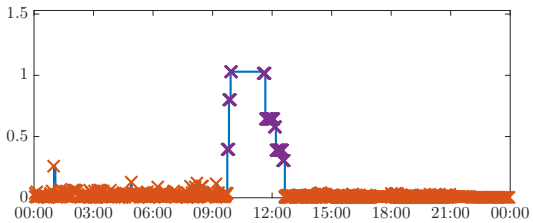
As described in section 4.3.1, location data recorded on Android smartphones can come from a range of different sources. In practice, it was found that the Android operating system would often provide inaccurate location values interspersed between what would appear to be more accurate data points. This can be seen from the data samples in figure 5.1(a) on the next page and figure 5.1(d) on page 93. These graphs show location coordinates recorded from two different participants over a weekend and weekday day, and the route that they most likely took. The dark orange star



(a) Raw data coordinates with inaccurately recorded noise (weekend)



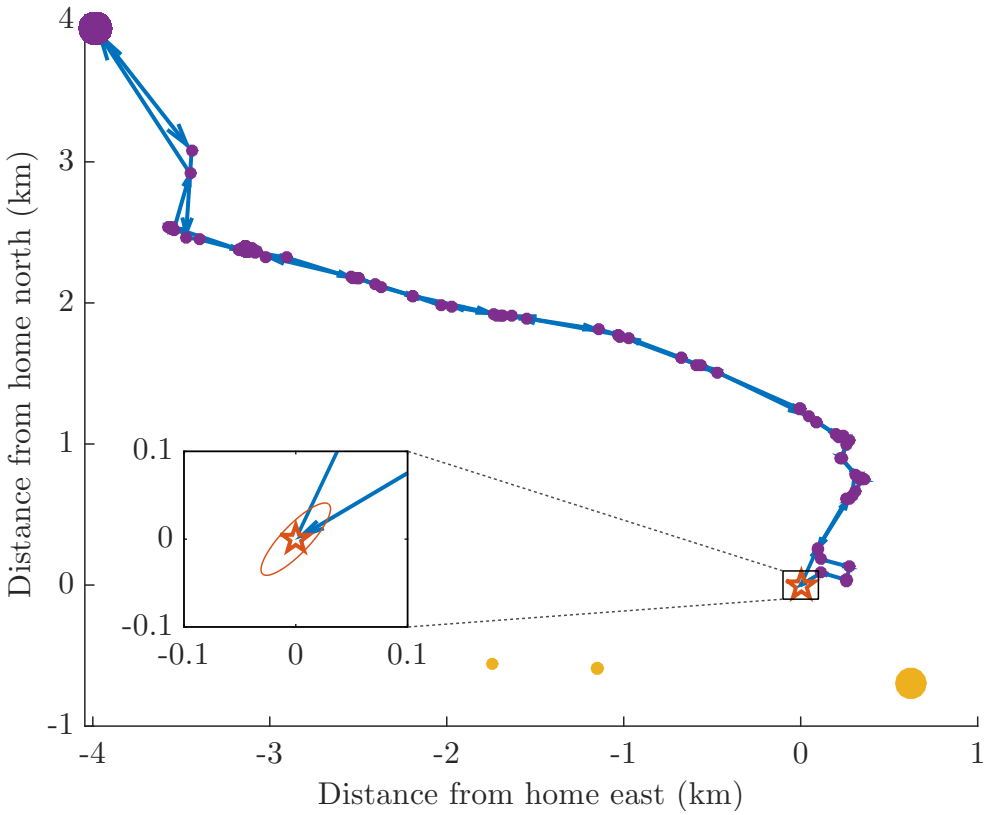
(b) Original data with noise (weekend)



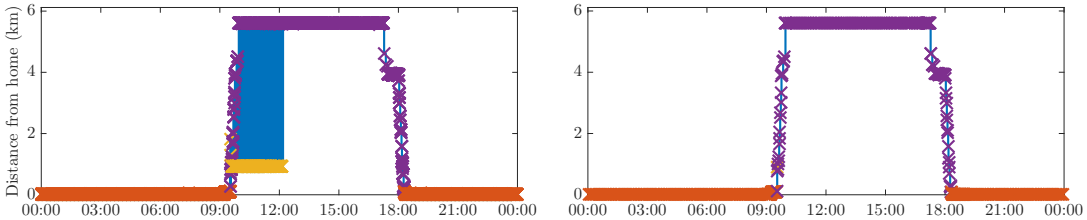
(c) Filtered data (weekend)

Figure 5.1: Examples demonstrating filtering of inaccurate data for a typical weekend (this page, a-c) and weekday (next page, d-f) day from different participants. Plots (a) and (d) on the top row show the coordinates where the participant was recorded relative to their home location. The dark orange star represents (0, 0) which is assumed to be the home location of the participant from which other distances are computed. The surrounding ellipse shows the 95 % confidence interval of the points recorded at the assumed home location modelled as a multivariate normal distribution. The purple points are accurate location readings with the path between them joined by the blue arrows. The broken arrow in (a) indicates that there is a gap of longer than 15 min where no data were recorded between the two points and therefore the location of the participant cannot be accurately determined. The yellow points are inaccurate noise recorded at the same time as the accurate recordings.

(continued on the next page)



(d) Raw data coordinates with inaccurately recorded noise (weekday)



(e) Original data with noise (weekday)

(f) Filtered data (weekday)

Figure 5.1 (cont.): The location coordinates can alternatively be represented as graphs (b-c) and (e-f) on the bottom row that show the Euclidean distance from where the participant was recorded to their assumed home location over the duration of the day. The colours of the markers in graphs (b-c) and (e-f) correspond to the locations with the same colour shown in plots (a) and (d) respectively. The blue shaded areas in (c) and (e) show rapid transitions between the purple (accurate) and yellow (inaccurate) locations (these transitions were excluded from plots (a) and (d) for clarity). It can clearly be seen that the yellow points are inaccurate because they are located far from what appears to be a reasonable path in the plots on the top row and that they occur concurrently with the locations in the reasonable path in the graphs on the left of the bottom row. The size of the point in the plots on the left indicates the length of time spent in each location, scaled between 10 min and 1 h. Plots (c) and (f) show the pre-processed data traces used for further analysis.

indicates the home location of the participant and the surrounding ellipse represents the 95 % confidence interval of the data points recorded at this location, modelled as a multivariate normal distribution. The three points south-west of the ellipse in figure 5.1(a), and to the south of the home location in figure 5.1(d), are samples that appear to be of low accuracy. This can be seen more clearly in figures 5.1(b) and (e) where the vertical axis is the Euclidean distance from home to each location sample and the horizontal axis is the time through the day. This provides a convenient way to visualise the change in location over time. The dark orange points indicate the home data points; the purple points are the true path; and the yellow points are the low accuracy data points. It can be seen that the purple readings can reasonably be assumed to be the true path because they follow a reasonable route, both in terms of the raw coordinates and the distance from home.

5.2 Calculating the 'home' location

Although the recorded geolocation data are anonymised (as described in section 4.3.3), it is convenient to calculate the assumed 'home' location for each participant. This assumed home location is used directly for the calculation of features such as the percentage of time spent at home (see section 5.10.4) and the diurnal movement calculated using the distance from home (see section 5.10.10) as well as providing convenient visualisation.

In order to determine the home location, the mode location between 2 A.M. and 7 A.M. for every day i in the data set was calculated as $\hat{h}_i$. The global home location was then calculated as $\hat{h}_g = \text{mode}(\hat{h}_i)$. All the data for the participant were then subtracted by $\hat{h}_g$ to leave the origin at location coordinates (0, 0) as the assumed home location.

5.3 Data filtering

The inaccurate data points described in section 5.1 would cause problems for the extraction of relevant features because the oscillations between the accurate and in-

Algorithm 5.1 Filtering of recorded location data.

Locations with a high source or destination $\frac{d}{dt}D$ indicate potential noise so if they occur commonly to or from that location then remove all instances of the location. $|\hat{D}|$ indicates the number of occurrences of the mode location.

data: $D(t)$ recorded data points at $t_{min} \leq t \leq t_{max}$.
result: $D'(t')$ filtered data points at $t'_{min} \leq t' \leq t'_{max}$.

begin
    $D' \leftarrow D$
    $t' \leftarrow t$
    repeat
        $t^* \leftarrow t'$ where $\frac{d}{dt'}D' > 100\,\mathrm{km\,h^{-1}}$ ▷ Data points with high $\frac{d}{dt'}D$.
        ▷ Each transition at time $t^\dagger \in t^*$ is between a source coordinate at $D(t^\dagger - 1)$ and a destination coordinate at $D(t^\dagger)$.
        $D_{src} \leftarrow D'(t^* - 1)$ ▷ Get all source coordinates.
        $D_{dest} \leftarrow D'(t^*)$ ▷ Get all destination coordinates.
        $\hat{D}_{src} \leftarrow \mathrm{mode}(D_{src})$ ▷ Get mode source coordinate.
        $\hat{D}_{dest} \leftarrow \mathrm{mode}(D_{dest})$ ▷ Get mode destination coordinate.
        if $|\hat{D}_{src}| \geq |\hat{D}_{dest}|$ **and** $|\hat{D}_{src}| > 10$ **then**
            Remove all occurrences of $\hat{D}_{src}$ from D'
        else if $|\hat{D}_{dest}| \geq 10$ **then**
            Remove all occurrences of $\hat{D}_{dest}$ from D'
        end if
    until $|\hat{D}_{src}| \geq 10$ **and** $|\hat{D}_{dest}| < 10$
end

accurate points would be incorrectly assumed to be from the participant travelling between states. A filtering method given in algorithm 5.1 above was developed to filter the recorded data D. This algorithm relies on two observed properties of the low-accuracy data points: a) they are geographically located far from the high-accuracy data while being temporally close, thus resulting in a high $\frac{d}{dt}D$; and b) they were commonly found to have exactly the same location coordinates. The algorithm therefore filters the data by detecting data points that had $\frac{d}{dt}D > 100\,\mathrm{km\,h^{-1}}$, either from the previous point, or to the next point. Where there were unique locations that had many such points detected, they were removed from the data set.

The sample data in figure 5.1 were filtered using algorithm 5.1 to extract the true paths shown in figures 5.1(c) and (f).

5.4 Data down-sampling

The unevenly sampled filtered data were down-sampled to a sample rate of 12 samples per hour using a median filter to remove any remaining spurious values. The whole data stream was first split into 1 h epochs. If the standard deviation of the recorded data in both latitude and longitude in an epoch was less than 0.01 km then all samples within the hour were set to the mean of the recorded data. If the standard deviation was greater than 0.01 km then a 5-minute median filter window was applied to the recorded latitude and longitude in the epoch. The reduction in noise resulting from this process improves the results of the extraction of stationary locations (described below).

5.5 Data imputation

Many participants were found to have sections of missing data in the locations recorded. This may be because the phone was switched off (either intentionally or because the battery was depleted) or because the phone was unable to obtain a GPS location or contact the Google Location Server. This is not uncommon in previous studies where smartphone-recorded location data have been analysed. For example Saeb et al. (2015a) excluded over 50 % of participants (22 out of 40) from their analysis of GPS locations due to insufficient data. Likewise, Gruenerbl et al. (2014) reported that location data for a mean of 31 % of days from their 12 participants were insufficient.

The reliability of many of the extracted features (described fully in section 5.10 on page 127) are impacted by the levels of missing data. For example, a feature such as the percentage of time spent at home (HS feature) will be skewed if the phone is switched off overnight, or if there are no data recorded when the participant is away, perhaps at work, during the day. In many of these cases, a reasonable guess as to the location of the participant can be made. The absolute accuracy of the guess is often not important. For example, consider the location of the participant shown in figures 5.1(a) to (c). It can be seen that there is about a 1.5 h section of data missing

when the participant leaves home (this is most likely because the participant has GPS switched off and was relying on the network location provider - if there was no Internet connectivity available then the location was not available). It is impossible to accurately fill in this data, but nevertheless it is possible to say that it was very unlikely that the participant was at home at this time.

Missing data were therefore imputed to minimise the level of missing data as far as possible. Data were imputed to make reasonable estimates of a) stationary locations; and b) transitions between stationary locations.

5.5.1 Related work

Many of the previous studies described in section 2.2.4.3 simply ignored sections of missing data (e.g. Canzian & Musolesi, 2015; Grünerbl et al., 2015; Saeb et al., 2015a). However, as discussed above, the quality and reliability of the derived features is highly dependent on the quality of the underlying data so this was not considered acceptable.

There has been other interest in imputing missing geographic location data, indicating the importance of this step. Barnett & Onnela (2016) developed a statistical method to impute geolocation coordinates between the start and end of missing sections of geolocation data based on an analysis of human movement by Rhee et al. (2007, 2011). Rhee et al. found that human mobility traces can be modelled as Levy walks, a form of random walk. Barnett & Onnela used this to impute missing geolocation coordinates based on the recorded trajectories at the start and end of the missing section. While this method may be useful for high-resolution imputation of short sections of missing data, the features described later do not require this level of "accuracy" on the imputed data. Additionally, in the data analysed here, the missing sections of data were observed to have common patterns that may not fit with the method by Barnett & Onnela, which assumes that the missing sections are randomly

distributed.[1] Certain assumptions could also be reasonably made about the accuracy of the available data that could further guide imputation. Both of these considerations are discussed in greater detail in section 5.5.3 after the imputation methods are described.

Sobrado (2018) took this a step further and used all the data available from an individual to build a "personalised map" of locations that the individual has been recorded at, which could then be used to perform personalised imputation of missing data. This method was able to build highly accurate maps of regular commute paths, for example. In a multimodal sensor system, Jaques et al. (2017) describe a model based on denoising autoencoders for estimating missing sensor data, including location, using data from the remaining available sensors.

Yue et al. (2017) also recently described a method for fusing geolocation data with wireless network access point associations and activity data. This data fusion improved the data coverage and the quality of the extracted features, demonstrated through improved significance levels and classification results.

[1] Firstly, it was common to find an individual recorded for a long time in a given location, followed by a section of missing data, followed by a long time recorded in another location. In this situation, the individual would have travelled between the two locations at some point, and therefore linear interpolation (with a given minimum speed of travel), at some point in the missing section is likely to be appropriate. A random Levy walk generated using the method by Barnett & Onnela would impute sections of travelling time and sections stationary at random locations which may actually be less accurate.

Secondly, it was also common to find a section of travelling, followed by a missing section, followed again by travelling, such as the location traces of the individual shown in figures 5.1(a) to 5.1(c) with approximately 1.5 h of missing data between 10:00 and 11:30. Imputing a random walk between the start and end of the missing section is no more likely to be accurate than assuming that the individual stopped somewhere in between the two where the location could not be determined. The method by Barnett & Onnela, however, may assume that the individual is travelling in an arc due to the trajectories of the data before and after the missing section, which is less likely to be accurate because individuals usually travel to specific places.

Finally, where the individual was recorded at the same location before and after the missing section, there is no reason to believe that the user is anywhere other than at the same point for the whole period (up to a given time). A random walk around the two points is not likely to be accurate.

5.5.2 Data imputation methods

In order to impute missing data for this work, a number of heuristic methods were developed based on the characteristics of the missing section and the surrounding data. Three different methods were developed for different scenarios. The methods are summarised here with full details given in appendix A. The appropriate method is chosen for each section of missing data based on the properties of the missing data section according to the rules in table 5.1 on the following page.

Midpoint imputation The first method fills missing data with the midpoint of the latitude and longitude coordinates on either end. Additionally it fills a path from the starting coordinates to the midpoint coordinates at the start of the missing section; and a path from the midpoint coordinates to the ending coordinates at the end of the missing section. The imputed paths between the coordinates are linearly filled along a direct vector between the two endpoints travelling at $80\,\mathrm{km\,h^{-1}}$. The method is detailed fully in algorithm A.2 in appendix A.

This method was used primarily where both the start and end coordinates were not the home location or both the start and end coordinates were very close to home (less than 250 m).

Travelling to home The second method is used in situations where the starting coordinates are far from home and the ending coordinates are close to home (usually home itself). In this situation it is assumed that the end of the missing section (close to home) more accurately identifies when the individual returns home and that they are most likely away during the missing section.

To account for the uncertainty in these assumptions, up to 2 h of data are first filled at the start of the missing section at the starting coordinates. Up to a further 2 h of data are then filled at the end of the missing section at the ending coordinates. Additionally a path is filled linearly between the starting and ending coordinates travelling at $80\,\mathrm{km\,h^{-1}}$. This path is imputed either at the end of the max 2 h of data

Table 5.1: Missing data imputation method rules.

Distance criteria	Time criteria	Imputation method
$\mathbf{d}^{\Delta} < 1\,\mathrm{km}$	$t^{\Delta} \leq 6\,\mathrm{h}$ **or** $t^{-} \geq 9\,\text{P.M.}$ **and** $t^{\Delta} \leq 12\,\mathrm{h}$	Midpoint imputation (algorithm A.2)
$\|\|\mathbf{d}^{-}\|\| > 750\,\mathrm{m}$ **and** $\|\|\mathbf{d}^{+}\|\| > 750\,\mathrm{m}$	$t^{\Delta} \leq 6\,\mathrm{h}$ **or** $t^{-} \geq 9\,\text{P.M.}$ **and** $t^{\Delta} \leq 12\,\mathrm{h}$	
$\|\|\mathbf{d}^{-}\|\| < 250\,\mathrm{m}$ **and** $\|\|\mathbf{d}^{+}\|\| < 250\,\mathrm{m}$	$t^{-} \geq 9\,\text{P.M.}$ **and** $t^{\Delta} \leq 18\,\mathrm{h}$	
$\|\|\mathbf{d}^{-}\|\| > 750\,\mathrm{m}$ **and** $\|\|\mathbf{d}^{+}\|\| < 750\,\mathrm{m}$	$t^{\Delta} \leq 6\,\mathrm{h}$ **or** $t^{-} \geq 8\,\text{P.M.}$ **and** $t^{\Delta} \leq 18\,\mathrm{h}$	Travelling to home (algorithm A.3)
$\|\|\mathbf{d}^{-}\|\| < 750\,\mathrm{m}$ **and** $\|\|\mathbf{d}^{+}\|\| > 750\,\mathrm{m}$	$t^{\Delta} \leq 6\,\mathrm{h}$ **or** $t^{-} \geq 8\,\text{P.M.}$ **and** $t^{\Delta} \leq 18\,\mathrm{h}$	Travelling from home (algorithm A.4)

$t^{-}\,/\,t^{+}$	Time of data point immediately before/after missing section.
$\mathbf{d}^{-}\,/\,\mathbf{d}^{+}$	Data coordinates immediately before/after missing section.
$\|\|\mathbf{d}^{-\,/\,+}\|\|$	Distance from home to coordinates immediately before/after missing section.
$t^{\Delta} := t^{+} - t^{-}$	Length of missing section.
$\mathbf{d}^{\Delta} := \|\|\mathbf{d}^{+} - \mathbf{d}^{-}\|\|$	Euclidean distance between locations before and after section.

filled at the start of the missing data, or half way through any remaining missing data. The method is detailed fully in algorithm A.3 in appendix A.

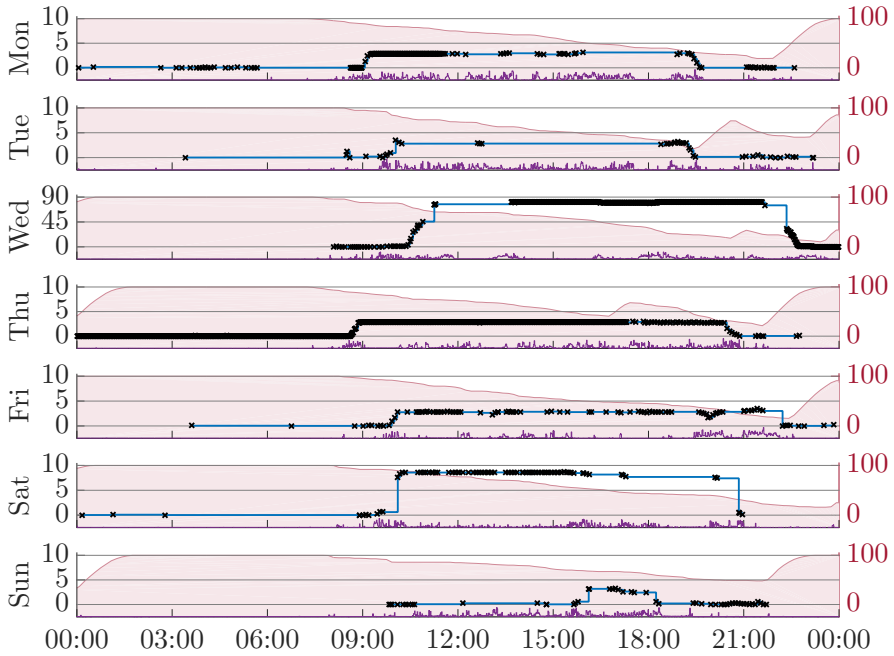
Travelling from home The final method is used in situations where the starting coordinates are close to home (usually home itself) and the ending coordinates are far from home. In this situation it is assumed that the start of the missing section more accurately identifies when the individual leaves home. As in the previous method, it is assumed that they are most likely away during the missing section. This method is therefore the opposite of the previous method, filling in first up to 2 h at the end of the missing section at the ending coordinates. The method is detailed fully in algorithm A.4 in appendix A.

5.5.3 Data imputation discussion

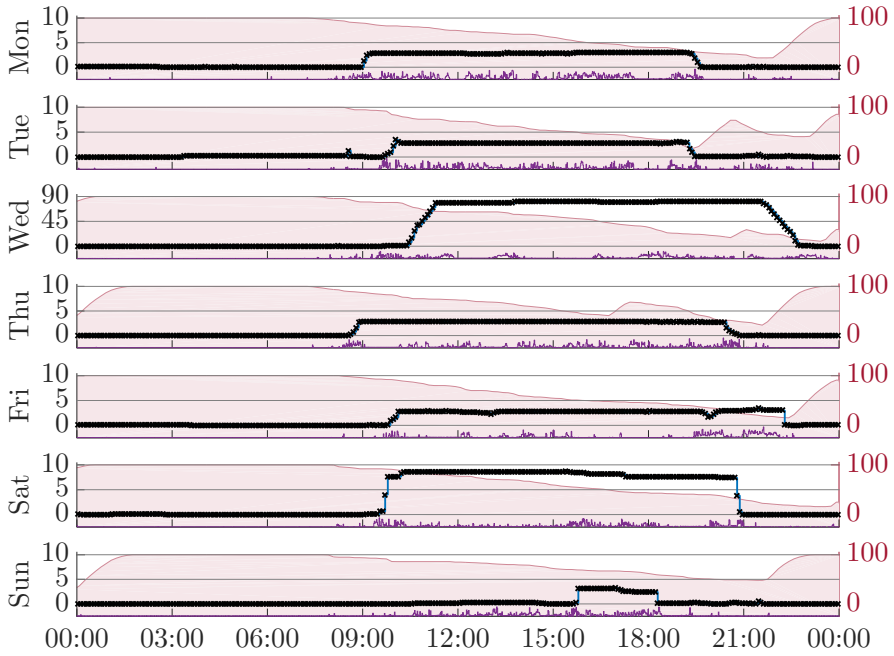
The data coverage in the raw data before preprocessing was described in section 4.3.4.2 and figure 4.5 on pages 80 and 81. While this gave an indication of the level of missing data (only 35 % of days from participants had $\geq$ 95 % data coverage) more detailed examination of the data is useful to understand the implication of this, and the impact of the data imputation methods developed.

Figure 5.2 on the next page shows a week of data from a participant (a) before preprocessing; and (b) after data imputation has been performed. The geolocation data are presented in the same way as the lower graphs in figure 5.1 as the distance in km from the assumed home location over the course of each day. This participant has particularly extreme levels of missing data in the week shown before preprocessing with only 37.7 % data coverage. After data imputation, this week has 100 % data coverage. Details of the overall data characteristics after preprocessing will be given in section 5.6.

Assessing the accuracy of the imputed data without separately collected reference data is difficult, but visual inspection of the data before preprocessing indicates that the imputed data are as expected. To aid in assessing the imputed data, the background shading in figure 5.2 shows the battery level recorded by the app, and the



(a) Before preprocessing (36.7 % coverage).



(b) After preprocessing (100 % coverage).

Figure 5.2: Examples of geolocation data (a) before; and (b) after preprocessing. The geolocation data are presented as the Euclidean distance of the participant from their assumed home location over the course of each day. The background shading indicates the battery level recorded by the app in % on the right-hand axis. The purple line at the bottom indicates the activity level of the phone recorded from the accelerometer (dimensionless).

purple line near the bottom of each graph shows the level of activity recorded by the accelerometer. This demonstrates that when the imputed data gives the participant as being at home, there are generally very low levels of activity, and the phone is normally plugged in. Travelling time is generally associated with higher levels of activity, and when the imputed data are away from home there are generally higher levels of activity. Additionally, there is no reason to believe that the participant is ever at home when they the imputed data show them as being away from home. Note that the imputed locations away from home carry little significance since none of the features described in the next section rely on specific places.

The data imputation methods developed make a number of assumptions about the missing data and characteristics of individuals. The primary assumption is that the phone will record time at home more accurately than time away from home. This is justified by the location data sources used by the Android operating system (described in section 5.1). One of the requirements of participation in the AMoSS study was that participants should have the phone connected to WiFi at home to enable data to be uploaded. Assuming that individuals leave WiFi switched on most of the time, this means that the location recorded at home will be relatively accurate (because WiFi-derived location is one of the most accurate location data sources). Considering the data from the participant shown in figure 5.2(a), when the participant leaves home or returns home there are generally data recorded at least in the transition. Figure 5.2(a) also indicates that higher levels of activity at home are generally associated with greater likelihood of having valid geolocation data. It is therefore likely that the Android operating system adapts the geolocation sample rate when there is little activity (presumably since it is unlikely that the location is changing). However, movement when leaving home or arriving home will trigger the operating system to provide location data if possible. At other times, the availability of geolocation data is dependent on other factors.

Extended periods of missing data where the individual is recorded at home before and after the missing section is most likely due to the phone being intentionally

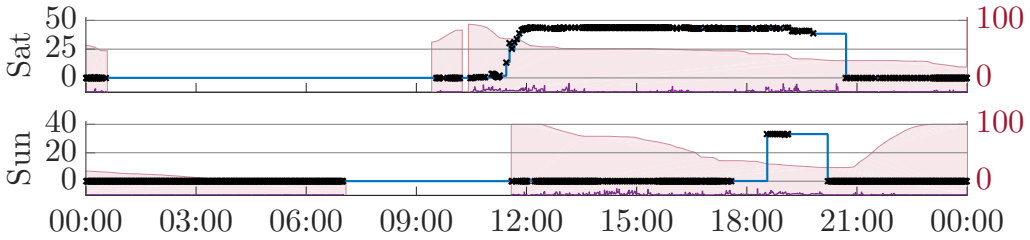
switched off (especially overnight). In this case the data can safely be imputed with the home location (or the midpoint or the locations either end).

In the other extreme where the individual is recorded away from home both before and after a missing section it can (within reason) be safely assumed that the individual does not return home in between. The choice of imputing this missing section with the midpoint between the locations recorded before and after the missing section is arbitrary, but is no less reasonable than using the start or end point.

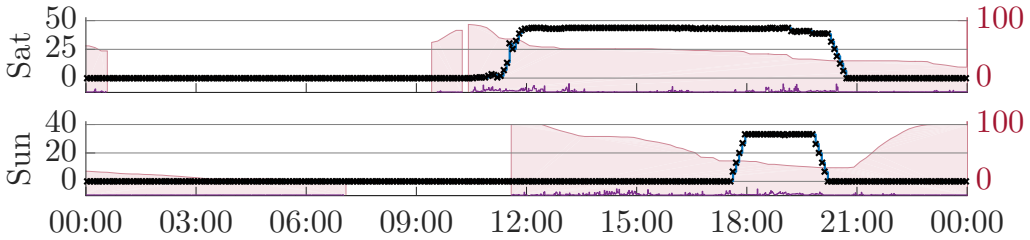
Finally, in the last two methods, one end of the missing data section is known to be far from home, and the other end is (close to) the home location. As above, the the critical assumption made here is that data recorded at home are reasonably accurate, and so when data are missing it is assumed that they are not at home.

The choice to include imputation of transitions between locations is optional but reasonable because individuals did indeed have to travel between the locations. An example of this can be seen with the imputed data at around 11 A.M. and 10 P.M. on Wednesday in figure 5.2. Clearly imputing transitions is reasonable here and ensures that features such as the total time spent travelling will accurately reflect the behaviour of the individual. A further example is given in figure 5.3 for two days from a different participant. Again, the imputed transitions on both days shown appear to be reasonable. Because this participant is travelling further from home than in figure 5.2, the imputed transitions will have a greater impact on the features calculated. The choice of imputing the travelling data at $80\,\mathrm{km\,h^{-1}}$ (approx 50 miles per hour) was a reasonable estimate given local speed limits.

→



(a) Before preprocessing (63.5 % coverage).



(b) After preprocessing (100 % coverage).

Figure 5.3: Further examples of geolocation data before and after preprocessing showing imputed transitions between location further from home on both days. Imputing transitions where the individual travels further from home will have greater impact on the calculated features.

5.6 Preprocessed geographic location data characteristics

This section describes the characteristics of the geolocation data after applying the preprocessing methods described in the previous sections in this chapter. The characteristics of the data before preprocessing were described in section 4.3.4.

5.6.1 Data coverage

The data coverage (percentage of each day with valid geolocation data) before preprocessing given in figure 4.5 showed that 35 % of days had 95 % or more coverage. Excluding the 23 % of days with no data, this means that 46 % of days with data had 95 % or more coverage.

The data coverage after the preprocessing described above is shown in figure 5.4. This shows that the number of days with over 95 % of coverage has increased to 61 %, or 79 % if days without any data are excluded.

5.6.2 Data properties

The data properties before preprocessing presented in figure 4.6 showed that 69.7 % of the recorded data were from when participants were within 500 m of their assumed home location, as determined by the method given in section 5.2.

The data properties after preprocessing are given in figure 5.5, which shows that preprocessing has decreased this slightly to 67.6 %. This is due to the midpoint imputation of data when participants are away from home, for example at work. As discussed previously it was often found that location data were recorded travelling to and from a location, with data missing between, when it is very likely that the individual was not at home.

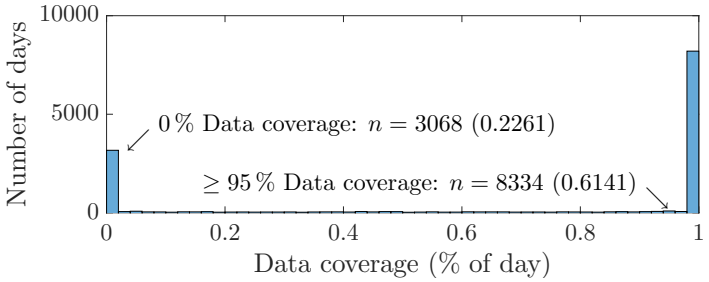


Figure 5.4: Data coverage of the geographic location data after preprocessing.

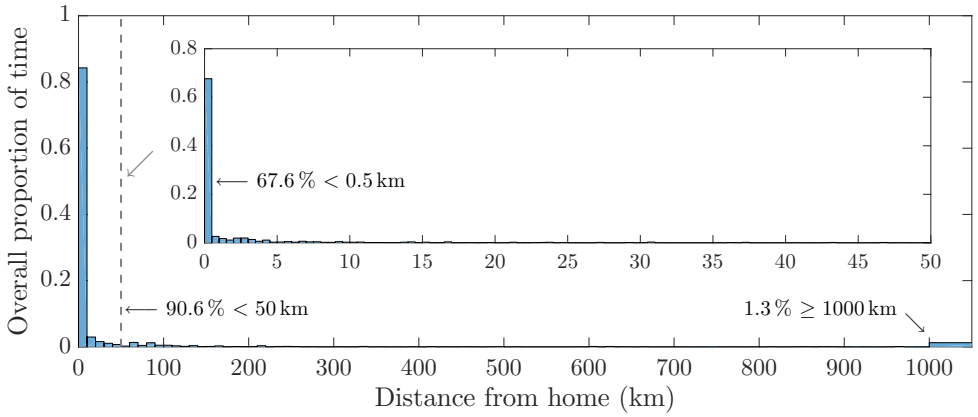


Figure 5.5: Overall distribution of the distance participants spent from home after preprocessing of the raw recorded geolocation data.

5.7 Data available for analysis

After data preprocessing and imputation, the location data were inspected and calendar weeks where there were sufficient data available were selected for analysis. As discussed in section 4.3.2, technical issues with the initial implementation of the app relating to geolocation were resolved in an update released on 1st May 2015. Therefore the data available for analysis were collected between May 2015 and April 2016.[2] Figure 4.3 in section 4.3 showed the periods of data available from all participants using the updated version of the app. Focussing on the participants who were using the upgraded app, figure 5.6 shows the calender weeks selected for analysis. Figure 5.6(a) shows the weeks with sufficient location data available for analysis, and Figure 5.6(b) shows the weeks with sufficient location data and questionnaire score labels available.

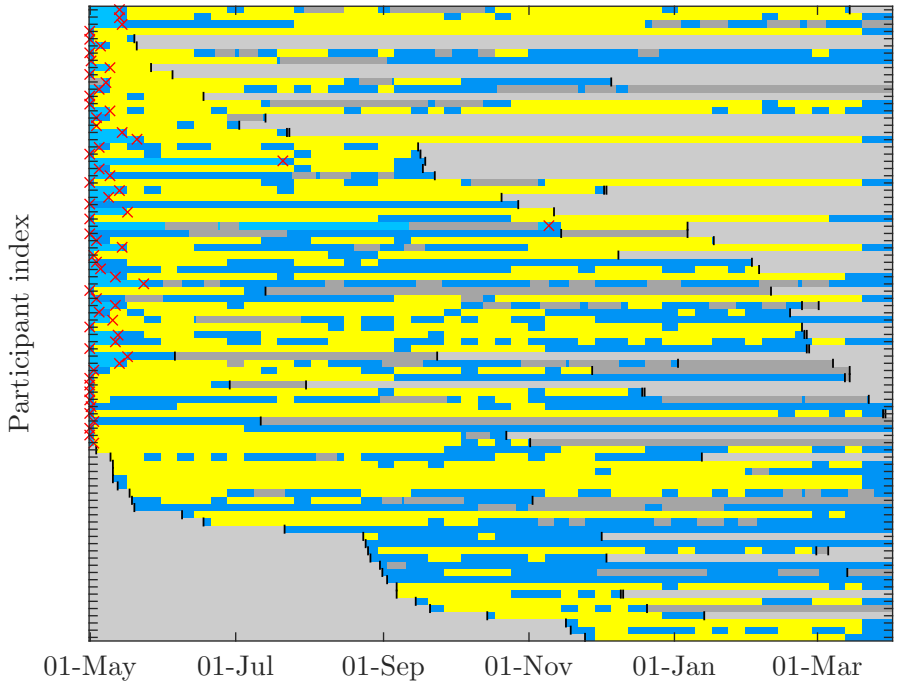
Details of the weeks of data available for analysis are given in table 5.2 on page 114 along with relevant participant demographics.

5.7.1 Questionnaire labels

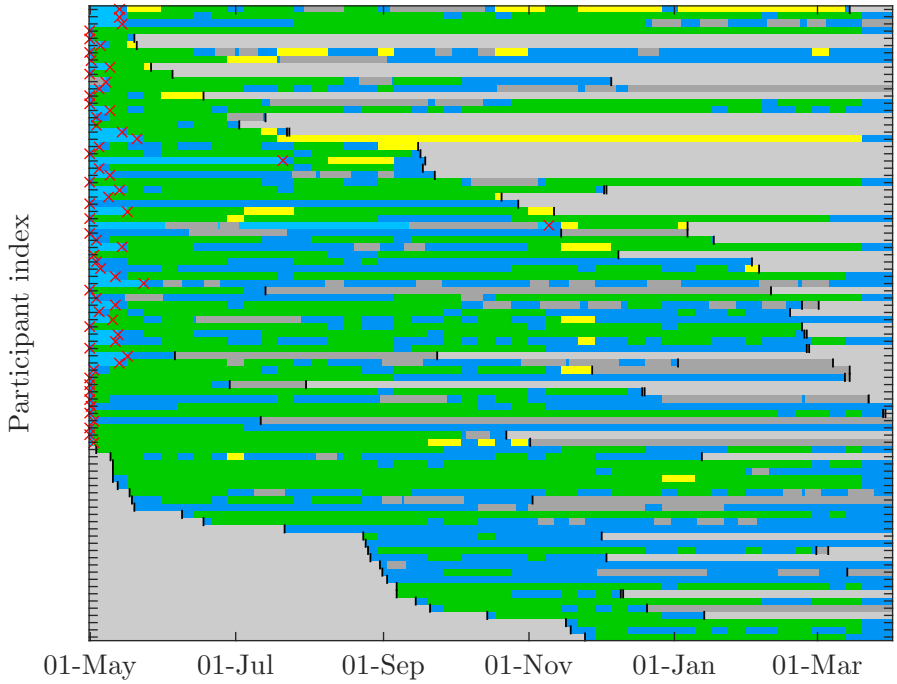
Labelling of calendar weeks using the clinically validated psychiatric questionnaires was described in section 4.4.4. Figure 5.6(b) shows the weeks for participants with sufficient geolocation data for analysis and associated labels from the questionnaires. Table 5.2 gives the number of labelled weeks available for analysis with summary participant demographics.

Section 4.4 described the questionnaires collected. The main three questionnaires were ASRM for assessing mania in bipolar disorder; GAD-7 for assessing generalised anxiety disorder; and QIDS for assessing depression. Due to the way that the questionnaires were administered to participants, generally all questionnaires were filled in at the same time. Figure 5.7 on page 110 shows the distributions of questionnaire responses for the three main questionnaires in the weeks with sufficient geolocation data available for analysis. Distributions are broken down by cohort, and the last row

[2] The data presented in this thesis were recorded between Monday 4th May 2015 and Sunday 20th March 2016 using the updated version of the study app.



(a) Weeks with sufficient geolocation data available for analysis shown in yellow.



(b) Weeks with geolocation data and associated questionnaire score labels shown in green. Weeks in yellow have geolocation data but no questionnaire scores.

Figure 5.6: Profile of calender weeks with sufficient (a) geolocation data; and (b) questionnaire scores for analysis between May 2015 and March 2016. Only the participants from figure 4.3 who had upgraded their phone app (see section 4.3.2) to the latest version are shown. The app update was released on 1st May 2015 and the red crosses show when participants upgraded.

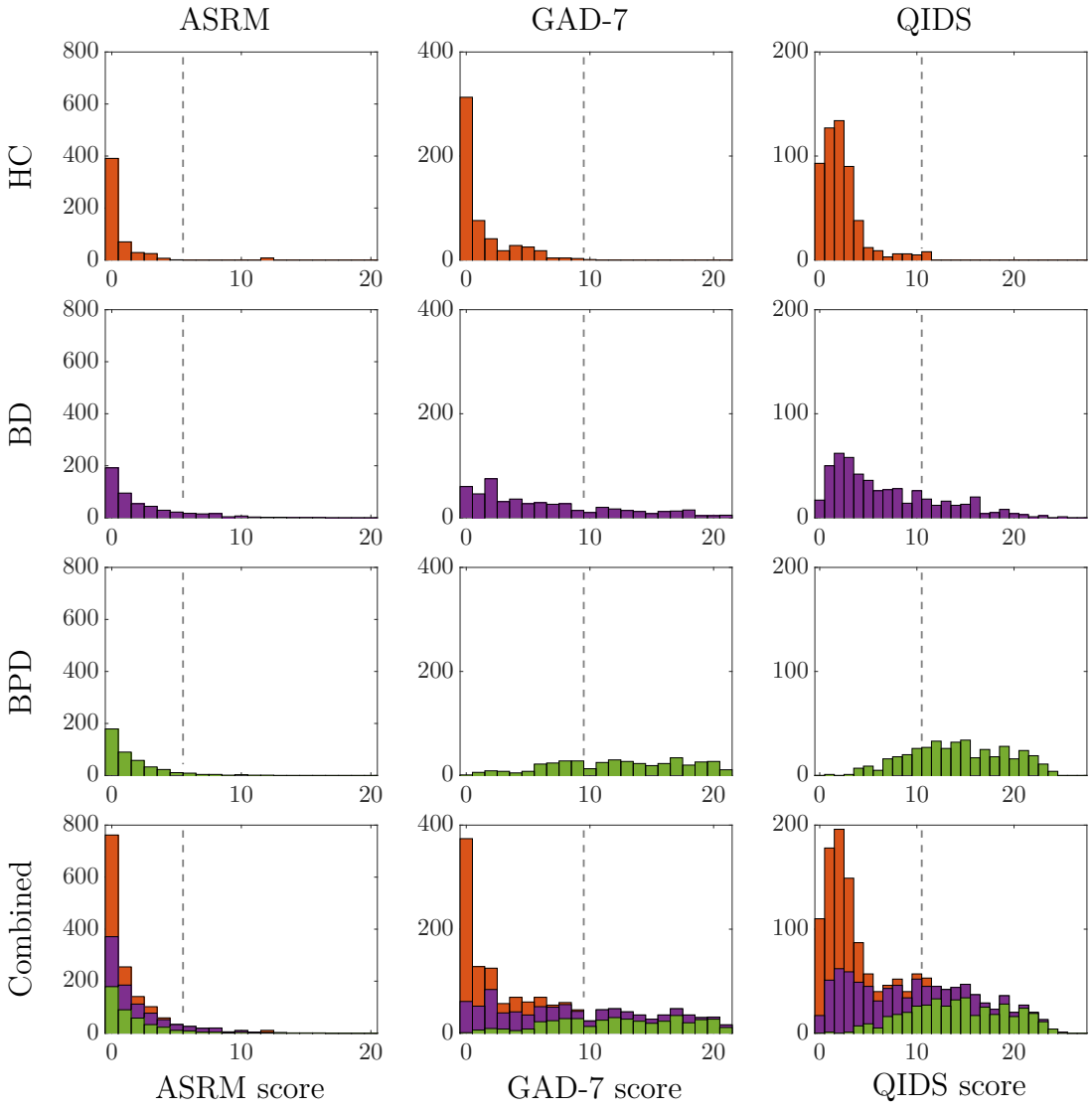


Figure 5.7: Distributions of clinically validated questionnaire responses by cohort. The first column shows ASRM; the second shows GAD-7; and the last shows QIDS. The grey dashed line in each column shows the clinically accepted thresholds for mania; generalised anxiety disorder; and depression respectively. The first three rows show the questionnaire response distributions for the three cohorts in the AMoSS study, and the last row shows stacked distributions for all cohorts. The height of each bar shows the total number of weeks labelled and the shading indicates how many are from each group.

shows the distributions for the three cohorts stacked. The relevant clinically accepted diagnostic thresholds for the three questionnaires (discussed in section 4.4) are shown as the dashed grey vertical lines. Further details of the number of weeks above and below the thresholds for each questionnaire and cohort are shown in table 5.3 on page 115.

The healthy controls were generally sub-threshold for all questionnaires. Participants with bipolar disorder were predominantly sub-threshold with 22 % of weeks labelled as depressed and 12 % labelled as manic. This roughly matches episode analysis in the literature with Solomon et al. (2010) reporting 219 bipolar I patients spending mean 31 % of time ill with a mood episode during a followup period of up to 25 years. The patients diagnosed with borderline personality disorder were predominantly above-threshold in QIDS and GAD-7 but sub-threshold for ASRM.

The limitations of assessing mental health state using these types of questionnaires were described in sections 2.2.1 and 4.4.3. At best, the QIDS questionnaire is a subjective proxy of the level of depressive symptomatology in the individual. However, these questionnaires are carefully worded in such a way as to reduce the inter-individual variability as much as possible. This contrasts to the Likert scale used in the custom daily *Mood Zoom* mood questionnaire implemented on the app (see section 4.4.2) which relies on the interpretation of the scale by the individual.

Examples of the questions included in the QIDS-SR_{16} questionnaire are shown in figure 5.8 on the next page. This demonstrates three different techniques used to reduce variability between individuals. The first technique is to use descriptive phrasing to describe the characteristics of the symptom. For example, instead of asking about sleep quality directly, the QIDS questionnaire uses the easily relatable descriptions of sleep shown in figure 5.8(a). The next technique is to use understandable quantitative metrics. For example, instead of asking individuals to rate how sad they feel, they are asked about how often then feel sad within specific bands as shown in figure 5.8(b). While the definition of 'sad' is still fundamentally subjective, individuals likely know when they consider themselves to be sad, so only have to identify how often they feel that way. Finally, symptoms may be rated relative to their usual levels. For example,

2. Sleep During the Night:
    0. I do not wake up at night.
    1. I have a restless, light sleep with a few brief awakenings each night.
    2. I wake up at least once a night, but I go back to sleep easily.
    3. I awaken more than once a night and stay awake for 20 minutes or more, more than half the time.

(a) Descriptive phrasing of symptomatology.

5. Feeling Sad:
    0. I do not feel sad.
    1. I feel sad less than half the time.
    2. I feel sad more than half the time.
    3. I feel sad nearly all of the time.

(b) Quantitative phrasing of symptomatology.

13. General Interest:
    0. There is no change from usual in how interested I am in other people or activities.
    1. I notice that I am less interested in people or activities.
    2. I find I have interest in only one or two of my formerly pursued activities.
    3. I have virtually no interest in formerly pursued activities.

(c) Relative phrasing of symptomatology compared to usual.

11. View of Myself:
    0. I see myself as equally worthwhile and deserving as other people.
    1. I am more self-blaming than usual.
    2. I largely believe that I cause problems for others.
    3. I think almost constantly about major and minor defects in myself.

(d) Combined phrasing including both descriptive and relative options.

Figure 5.8: Example questions from the QIDS-SR_{16} questionnaire by Rush et al. (2003) demonstrating different techniques for phrasing responses to reduce inter-individual variability as much as possible. The numbering of the responses indicates the contribution of the response to the final QIDS score.

when asking about general interest in people or activities, the responses are phrased in terms of the relative level of interest compared to usual as shown in figure 5.8(c). Relative and descriptive phrasing may also be combined as shown figure 5.8(d).

These techniques have led to the QIDS questionnaire becoming an accepted self-reported indicator of depression. While it is accepted that it may not be the ideal metric to use for labelling mental state in individuals, it is considered to be a suitable proxy.

5.8 Extracting stationary locations

After the data were filtered, the remaining location points could be split into stationary locations and transitioning / travelling points from the differential of the data. 10 min moving averages were calculated for each sample at time t as follows:

$$\bar{x}_t^{\text{cen}} = \frac{1}{|\tau|} \sum_{\tau} \frac{d}{d\tau} D(\tau) \qquad t - 10\,\text{min} \leq \tau \leq t + 10\,\text{min} \tag{5.1a}$$

$$\bar{x}_t^{\text{back}} = \frac{1}{|\tau|} \sum_{\tau} \frac{d}{d\tau} D(\tau) \qquad t - 10\,\text{min} \leq \tau \leq t \tag{5.1b}$$

$$\bar{x}_t^{\text{fwd}} = \frac{1}{|\tau|} \sum_{\tau} \frac{d}{d\tau} D(\tau) \qquad t \leq \tau \leq t + 10\,\text{min} \tag{5.1c}$$

Thresholds were applied where the location at time t was considered to be a transitioning location if

$$\left(\bar{x}_t^{\text{cen}} > d\right) \textbf{ and } \left(\bar{x}_t^{\text{back}} > d\right) \textbf{ and } \left(\bar{x}_t^{\text{fwd}} > d\right) \textbf{ and } \left(\frac{d}{dt} D(t) > d'\right). \tag{5.2}$$

It is possible to adjust the specificity and sensitivity of the extraction by adjusting d and d'. For this work, thresholds of $d = 1.5\,\text{km}\,\text{h}^{-1}$ and $d' = 0.4\,\text{km}\,\text{h}^{-1}$ were applied.

A further step of filtering was applied where stationary samples immediately preceded and followed by transitioning states were considered to be transitioning.

Table 5.2: Geolocation data statistics and participant demographics.

	Healthy controls	Bipolar disorder	Borderline personality disorder	Total
Enrolled in the study between January 2014 and November 2015				
Participants	53	54	33	140
Using upgraded app between 4th May 2015 and 20th March 2016				
Participants	35	29	24	88
Total weeks	932	817	749	2498
With calendar weeks of data selected for analysis				
Participants	30	27	21	78
Total weeks	572	543	436	1551
With labelled weeks of data available for analysis				
Participants	30	27	21	78
Total weeks	531	505	417	1453
Weeks per participant*	16.5 ± 18	18 ± 23	20 ± 21.5	17 ± 18
Demographics for participants with labelled data available for analysis				
Gender	10 male; 20 female	9 male; 18 female	1 male; 20 female	20 male; 58 female
Age*	40.5 ± 15.0	44.0 ± 17.75	37.0 ± 11.5	39.5 ± 16.0
BMI*	24.11 ± 5.60	27.22 ± 4.44	29.76 ± 10.90	26.05 ± 7.57
Employment status	16 full-time employed; 5 part-time employed; 4 unemployed; 3 students; and 2 retired	11 full-time employed; 9 part-time employed; 4 unemployed; 2 students; and 1 unknown	2 full-time employed; 6 part-time employed; 11 unemployed; 1 student; and 1 unknown	29 full-time employed; 20 part-time employed; 19 unemployed; 6 students; 2 retired; and 2 unknown

* Data presented as median $\pm$ iqr

Table 5.3: Breakdown of clinically validated questionnaire responses.

	Healthy controls	Bipolar disorder	Borderline personality disorder
Total labelled weeks	531	505	417
Total participants with labelled weeks	30	27	21
Labelled weeks with ASRM score < 6	523 (0.98)	443 (0.88)	399 (0.96)
Labelled weeks with ASRM score ≥ 6	8 (0.02)	62 (0.12)	18 (0.04)
Participants with ASRM score ≥ 6	1	12	8
Labelled weeks with GAD-7 score < 10	531 (1.00)	376 (0.74)	145 (0.35)
Labelled weeks with GAD-7 score ≥ 10	0 (0.00)	129 (0.26)	272 (0.65)
Participants with GAD-7 score ≥ 10	0	11	17
Labelled weeks with QIDS score < 11	526 (0.99)	395 (0.78)	116 (0.28)
Labelled weeks with QIDS score ≥ 11	5 (0.01)	110 (0.22)	301 (0.72)
Participants with QIDS score ≥ 11	1	16	19

5.9 Clustering stationary locations

The stationary locations extracted above were clustered to determine the number of unique locations visited by an individual. Clustering of locations is a common task when working with geolocation data and various different methods have been proposed and used.

Saeb et al. (2015a) used a method based on the common K-means clustering algorithm that attempts to extract K optimal cluster centres from multidimensional data. K-means requires specifying the number of clusters, K, to extract, but because the expected number of clusters is generally unknown for geolocation data, Saeb et al. used a simple method where where K is set to 1 and incremented while the Euclidean distance between all the clusters remains above a threshold distance l apart. While this may work in simple cases, it is unsatisfactory for a number of reasons. Firstly the clusters obtained by K-means are highly dependent on the initial (random) cluster locations. One attempt to improve performance, known as K-means++ (Arthur & Vassilvitskii, 2007), chooses the initial cluster centres to be evenly distributed throughout the training data set. Even with this modification, it is usually necessary to run the algorithm several times to avoid local optima (Jain, 2010), which in turn leads to the problem of how to select the 'best' result. Secondly, using the stopping criteria of a threshold between points may fail when there are two genuine clusters that are closer than the threshold, but weaker clusters may not yet have been identified.

Farhan et al. (2016a) used the DBSCAN clustering algorithm developed by Ester et al. (1996) for clustering of GPS coordinates. DBSCAN is a density-based clustering method that allocates data into arbitrary sized and shaped clusters. For each data point P, the algorithm finds the other points S within a maximum distance ϵ, and if there are at least m points, then it creates a new cluster containing P and S. This process is repeated for each point in S, enlarging the cluster until there is no data point in the cluster with m points within distance ϵ that are not already allocated to the cluster. The main advantages of DBSCAN are that it does not require the number of clusters to be specified in advance; it efficiently handles outliers; and it forms arbitrary

sized clusters based on the density of points in the dataset. While these advantages are useful in many applications, when clustering geographic coordinates where the clusters are expected to be roughly Gaussian, parametrising by m and ϵ is not as natural as parametrising by the size of location clusters. This is especially relevant where many distinct places may be close together with noisy measurements making them all appear to be in the same cluster using DBSCAN. This is demonstrated in section 5.9.2.

de Montjoye et al. (2013) were vague in how locations were identified, simply stating that they "empirically defined places by grouping together the GPS points of a user that are less than 50 m apart and by defining their center of mass as the coordinates of the place". It is not clear whether this was automatic or manual.

Ashbrook & Starner (2003) used an alternative modification of the K-means algorithm to group GPS points. Their algorithm chooses a random starting mean value and allocates all points with a given radius r of that mean value to the cluster. Similar to K-means, a new mean for the cluster is set to the mean of the allocated GPS coordinates. This process is repeated for this cluster until convergence, after which the cluster mean is fixed and the allocated points are removed from consideration. The whole process is then repeated until there are no more unallocated GPS coordinates remaining. Like K-means this is a greedy algorithm and given the dependence on the fixed value of the cluster radius r it is likely that once the main clusters have been allocated there will be remaining points that will end up in new clusters. It moves the parametrisation problem of selecting the number of clusters K to selecting the cluster radius r. In the paper, Ashbrook & Starner plot the radius against the number of clusters found and choose a value of r based on a "knee" in the graph.

Non-parametric clustering methods such as the Dirichlet process discussed in section 3.3 initially seem appealing, but may not work well in practice, mainly because GPS data are not naturally suited to representation as a Bayesian generative model. As demonstrated in figure 5.5, individuals tend to spend most of their time relatively close to home, but occasionally spend time far (sometimes very far) from home – presumably on holiday or work trips. This makes it difficult to define a suitable prior

to apply to the location coordinates, other than that they could be anywhere in the world!

5.9.1 Hybrid clustering method

An efficient hybrid method was therefore developed that uses the modified K-means by Ashbrook & Starner (2003) to generate initial clusters which are then used in a DP-inspired model to non-parametrically *combine* the extracted clusters, rather than using a DP model in the usual sense to create new clusters from G_0. Because this method does not allow the creation of new clusters it is not a Dirichlet process model. However, it uses the CRP prior (equation (3.53)), which is also used in the DP, to collapse clusters.

A number of other modifications were also made to the DP algorithms to handle specific features of geolocation data. The whole hybrid clustering method is described in algorithm 5.3 on page 122. There are three main parts to the algorithm described in the following subsections.

5.9.1.1 Initial clustering of locations

The first step in the algorithm is to produce a set of m initial clusters using the algorithm described by Ashbrook & Starner (2003). The full algorithm is described in algorithm 5.2. The algorithm takes the n raw data coordinates in $\mathbf{D}$ and returns m cluster centres in $\mathbf{L}$ and a cluster index for each point in $\mathbf{c}$. The cluster index for each point can take values in $[1, m]$ for each cluster, or 0 if the point is identified as noise.

Although most transitioning locations are excluded from the clustering phase as described in section 5.8, some transitioning points may remain and these can be considered as noise. The clustering methods described above, especially K-means, can be quite sensitive to noise such as this (Zhou et al., 2007). For this reason, the algorithm by Ashbrook & Starner was modified to only create a new cluster if there are a minimum number of points allocated to the cluster.

Algorithm 5.2 Initial clustering algorithm based on Ashbrook & Starner (2003).

data: $\mathbf{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$ set of n recorded stationary locations.
result: $\mathbf{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_m\}$ set of m unique cluster location centres.
$\mathbf{c} = \{c_1, \ldots, c_n\}$ cluster id for each data point, $c_i \in [0, m]$ ($0 \Leftrightarrow$ noise).

function SeedLocations($\mathbf{D}$)
    $\mathbf{L} \leftarrow \{\,\}$
    $c_i \leftarrow 0 \;\forall\; i \in [1, n]$
    $\mathbf{D}' \leftarrow \mathbf{D}$
    while $|\mathbf{D}'| > 0$ **do**
        $\mathbf{D}^* \leftarrow \mathbf{D}'$ rounded to the nearest 1 km
        $\mathbf{d}^* \leftarrow \text{mode}(\mathbf{D}^*)$ ▷ Coarse grid search.
        $\mathbf{D}^\dagger \leftarrow \{\mathbf{d}'_i\}$ rounded to the nearest 100 m where $\mathbf{d}^*_i = \mathbf{d}^*$
        $\mathbf{c}' \leftarrow \text{mode}(\mathbf{D}^\dagger)$ ▷ Fine grid search in $\mathbf{D}^\dagger$.
        for $t = 1, 2, \ldots$ **do** ▷ Repeat until convergence.
            $\mathbf{D}^\star \leftarrow \{\mathbf{d}'_i\}$ where $||\mathbf{d}^*_i - \mathbf{c}|| < r$
            $\mathbf{c}' \leftarrow \text{mean}(\mathbf{D}^\star)$
        end for
        if $|\mathbf{D}^\star| > 5$ **then**
            Add $\mathbf{c}'$ to $\mathbf{L}$ at index j
            $c_i \leftarrow j$ where $\mathbf{d}_i \in \mathbf{D}^\star$
        end if
        Remove elements of $\mathbf{D}^\star$ from $\mathbf{D}'$
    end while
    return $\mathbf{L}, \mathbf{c}$
end function

Ashbrook & Starner do not describe how the starting coordinates of each cluster were chosen. A two-step deterministic method is used in algorithm 5.2 where a coarse grid search is first performed where all locations are rounded to the nearest 1 km and the mode is selected. All raw locations in the coarse mode location are then rounded to the nearest 100 m and the mode used as the starting location. This ensures that the areas of high density (which are likely to be in the same cluster) are clustered first.

The two parameters in this part of the algorithm are the radius r within which points are assigned to a cluster; and the minimum number of points required for a cluster to be created. Clearly these two are related because larger clusters will have more points assigned on average. Due to the collapsing of clusters performed in the

next two parts, r was set to be small ($r = 100\,\text{m}$) to generate more cluster centres than will be required. The minimum number of points required for a cluster to be created was fixed at 6 which provided reasonable removal of noise, and in the down-sampled signal corresponds to at least 30 min spent in a location. Locations where less than 30 min were spent are unlikely to be significant.

5.9.1.2 Collapsing of clusters

The main part of the hybrid clustering algorithm is based on the Dirichlet process sampling algorithm described in section 3.3.1. Unlike in a DP model that allows unlimited creation of new clusters, in the hybrid algorithm it is assumed that there are no more than the m clusters generated by the initial clustering (see section 5.9.1.1). These m clusters are split into set $\mathcal{C}$ for clusters that currently have data points assigned; and set $\mathcal{C}^*$ for clusters that do not currently have data points assigned. During sampling, each data point i, recorded at coordinates $\mathbf{x}_i$, is allocated to a cluster according to probability

$$p(c_i = c \,|\, \mathcal{C}, \mathcal{C}^*) \propto \begin{cases} \frac{\sum_{j=1}^{n,\, j\neq i} \mathbb{I}(c_j = c)}{n-1+\alpha} \mathcal{N}_2(\mathbf{x}_i \,|\, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}) & \forall\ c \in \mathcal{C} \\ \frac{\alpha}{|\mathcal{C}^*|} \cdot \frac{1}{n-1+\alpha} \mathcal{N}_2(\mathbf{x}_i \,|\, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}) & \forall\ c \in \mathcal{C}^* \end{cases} \tag{5.3}$$

where the standard multivariate normal mean $\boldsymbol{\mu}_c$ is the pair of coordinates at the centre of cluster c.

The covariance matrix of each cluster, $\boldsymbol{\Sigma}$, is one of the two parameters of this part of the model. While the covariance of the points currently assigned to each cluster could have been used directly, or sampled with a suitable prior, fixing this value gave most control over the clustering process. A diagonal covariance matrix with a value of $\boldsymbol{\Sigma} = \sigma^2 \cdot \mathbf{I}_2$ where $\sigma = 150\,\text{m}$ was found to provide reasonable separation between neighbouring locations. With this covariance, there is a 95 % confidence interval of points being within a radius of $\sigma = 300\,\text{m}$ of the cluster centre. The α parameter was set to $\alpha = 0.0001$ to encourage larger clusters to form, although generally the amount of data was found to be more significant for cluster formation than the value of α.

After all points have been allocated to clusters, the cluster centres are resampled based on the points currently assigned to each cluster. If $\mathbf{D}_c^*$ contains the points currently assigned to cluster c, then the new cluster centre $\boldsymbol{\mu}_c$ is sampled from $\mathcal{N}_2(\bar{\mathbf{D}}_c^*, \boldsymbol{\Sigma})$ where $\bar{\mathbf{D}}_c^*$ is a vector of the mean x and y coordinates of the set of points. The covariance $\boldsymbol{\Sigma}$ could use the covariance of the set of data points, i.e. $\boldsymbol{\Sigma} = \text{cov}(\mathbf{D}_c^*)$, but this was found to result in excessive instability in the model. Greater control was obtained by setting $\boldsymbol{\Sigma} = \sigma^2 \cdot \mathbf{I}_2$ where $\sigma = 10\,\text{m}$, which means that location centres were sampled close to the mean of the data points assigned to each cluster.

5.9.1.3 Additional collapsing of clusters

The sampling of clusters using equation (5.3) works well to combine clusters when there are two clusters close together and one has many more points allocated than the other. However, if there are clusters close together with similar numbers of points allocated to each one then the sampler will randomly assign all the points to one of these clusters with relatively equal probability, which will reach an equilibrium rather than combining. This arises from the unconventional use of the CRP prior to combine clusters rather than to create new clusters in a DP model. To overcome this issue, a final stage of combining clusters was performed after the cluster sampling in each iteration.

This stage involves first calculating the distance between all pairs of cluster centres, finding the pair with the minimum distance δ', and then merging them with probability $2\,\Phi(\delta' \,|\, 0, \sigma)$. If two clusters are merging, then the points assigned to the cluster with the smallest number of points are assigned to the cluster with most points. For simplicity, σ was set to the same value as used when sampling cluster centres, $\sigma = 10\,\text{m}$, to ensure that only clusters very close are merged.

5.9.1.4 Hybrid clustering method summary

The three parts of the hybrid clustering method described above are given in detail in algorithm 5.3 on the next page.

Algorithm 5.3 Hybrid clustering algorithm.

data: $\mathbf{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$ set of n recorded stationary locations.
result: $\mathbf{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_m\}$ set of m unique cluster location centres.
$\mathbf{c} = \{c_1, \ldots, c_n\}$ cluster id for each data point, $c_i \in [0, m]$ ($0 \Leftrightarrow$ noise).

begin

Initial clustering of locations ▷ See section 5.9.1.1 for details.

$\mathbf{L}, \mathbf{c} \leftarrow$ SeedLocations($\mathbf{D}$) ▷ See algorithm 5.2 for details.

for $t = 1, 2, \ldots$ **do** ▷ Repeat until convergence.

Collapsing of clusters ▷ See section 5.9.1.2 for details.

for $i = 1, \ldots, n$ **do** ▷ Loop over all data samples.
$\mathcal{C} \leftarrow \left\{\mathbf{l}_k : k \in [1, m] \ \wedge \ \sum_{j=1}^{n, j \neq i} \mathbb{I}(c_j = k) > 0\right\}$ ▷ Assigned locations.
$\mathcal{C}^* \leftarrow \left\{\mathbf{l}_k : k \in [1, m] \ \wedge \ \sum_{j=1}^{n, j \neq i} \mathbb{I}(c_j = k) = 0\right\}$
$c_i \sim p(c_i \,|\, \mathcal{C}, \mathcal{C}^*)$ ▷ Sample c_i using equation (5.3).
end for

$\mathcal{K} \leftarrow \left\{k : k \in [1, m] \ \wedge \ \sum_{j=1}^{n} \mathbb{I}(c_j = k) > 0\right\}$ ▷ Loop over all clusters with assigned points.
for $k \in \mathcal{K}$ **do**
$\mathbf{D}^* \leftarrow \{\mathbf{d}_i : i \in [1, n] \ \wedge \ c_i = k\}$ ▷ Get points assigned to cluster k and sample cluster centre.
$\mathbf{l}_k \sim \mathcal{N}_2(\bar{\mathbf{D}}^*, \mathbf{\Sigma})$
end for

Additional collapsing of clusters ▷ See section 5.9.1.3 for details.

$\mathbf{\Delta} \leftarrow \{\delta_{kl} : k, l \in [1, m]\}, \delta_{kl} = \begin{cases} \infty & \text{if } k = l \text{ or } k \notin \mathcal{K} \text{ or } l \notin \mathcal{K} \\ ||\mathbf{l}_k - \mathbf{l}_l|| & \text{otherwise} \end{cases}$

$\delta' \leftarrow \min(\mathbf{\Delta})$ ▷ Calculate Euclidean distance between all pairs of points and find minimum.
$p_{\delta'} \leftarrow 2\,\Phi(\delta' \,|\, 0, \sigma)$
$u \sim \text{Unif}(0, 1)$
if $u < p_{\delta'}$ **then** ▷ Merge clusters k and l with probability $p_{\delta'}$ by assigning all points to the cluster with the most currently assigned points.
Select any k, l where $\delta_{kl} = \delta'$ in $\mathbf{\Delta}$.
if $\sum_{j=1}^{n} \mathbb{I}(c_j = k) > \sum_{j=1}^{n} \mathbb{I}(c_j = l)$ **then**
$c_i \leftarrow j \ \forall \ i$ where $c_i = k$
else
$c_i \leftarrow k \ \forall \ i$ where $c_i = l$
end if
end if
end for
end

5.9.2 Locations clustering discussion

The introduction to this section described several alternative methods of performing clustering. One of the methods described was DBSCAN, which has been used in some previous work to cluster GPS coordinates. Without labelled data to compare to, it is not possible to objectively compare the proposed hybrid clustering method directly with DBSCAN (or any other method), but figure 5.9 on the following page shows examples of how both methods perform with the same parameters given different amounts of data from one of the participants in the study. All the data tested are from within ±2 km of the assumed home location of the participant. Using data close to home is useful to demonstrate how the methods perform because, as demonstrated in figure 5.5, participants tend to spend most of their time near home. Participants are also likely to visit more different places near to home, and this is reflected in the data shown in figure 5.9.

In figure 5.9, data points identified as noise are shown in grey, and the location clusters identified are shown in different colours with their centres shown with a black cross. With the fewest coordinates presented to the algorithms in the first row, both algorithms perform similarly. The data points identified as noise are similar, and a similar number of clusters are found. The clusters also appear plausible in both cases, with similar sizes.

With more data available in the second row, both methods find more clusters as expected. However, DBSCAN also creates some very large clusters, for example the light blue one towards the top of the plot. This is exacerbated as even more data are available in the bottom row, while the hybrid clustering algorithm continues to find similarly sized clusters.

This behaviour of DBSCAN is due to the way that it is parametrised, as described in the introduction to the section, and the way that it creates clusters purely based on density. To demonstrate this further, figure 5.10 on page 125 shows the effect of tuning the DBSCAN parameters when clustering the subset of data towards the top of the bottom row in figure 5.9. As the minimum distance between points in a cluster,

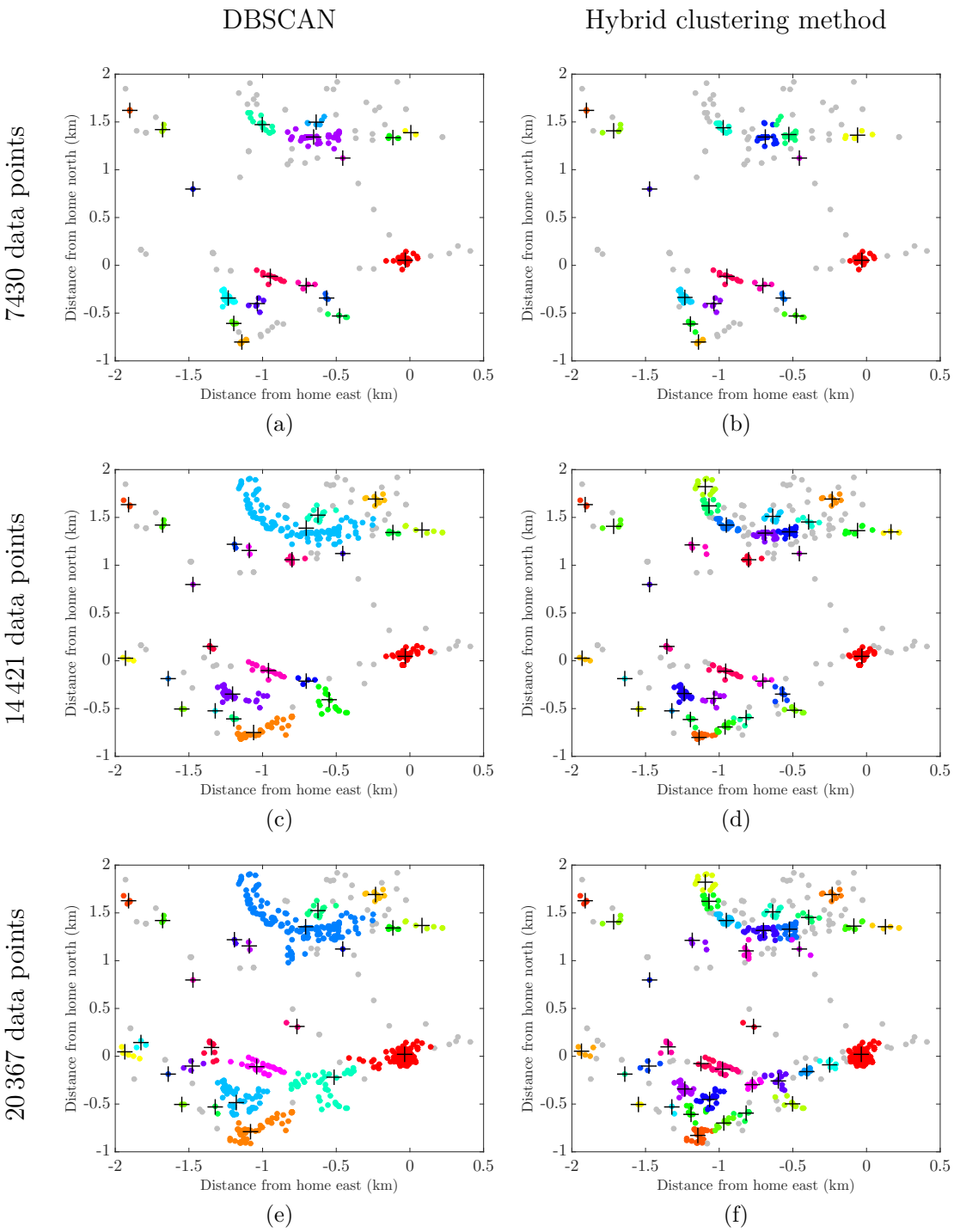


Figure 5.9: Examples of geolocation data clustered using DBSCAN and the hybrid clustering method described in section 5.9.1. The number of data points presented to each algorithm is varied between 7430 and 20 367 in each row, demonstrating the difference in performance depending on the amount of data used. Grey points are identified as noise, and individual clusters found are shown in different colours, with their centres shown with a black cross.

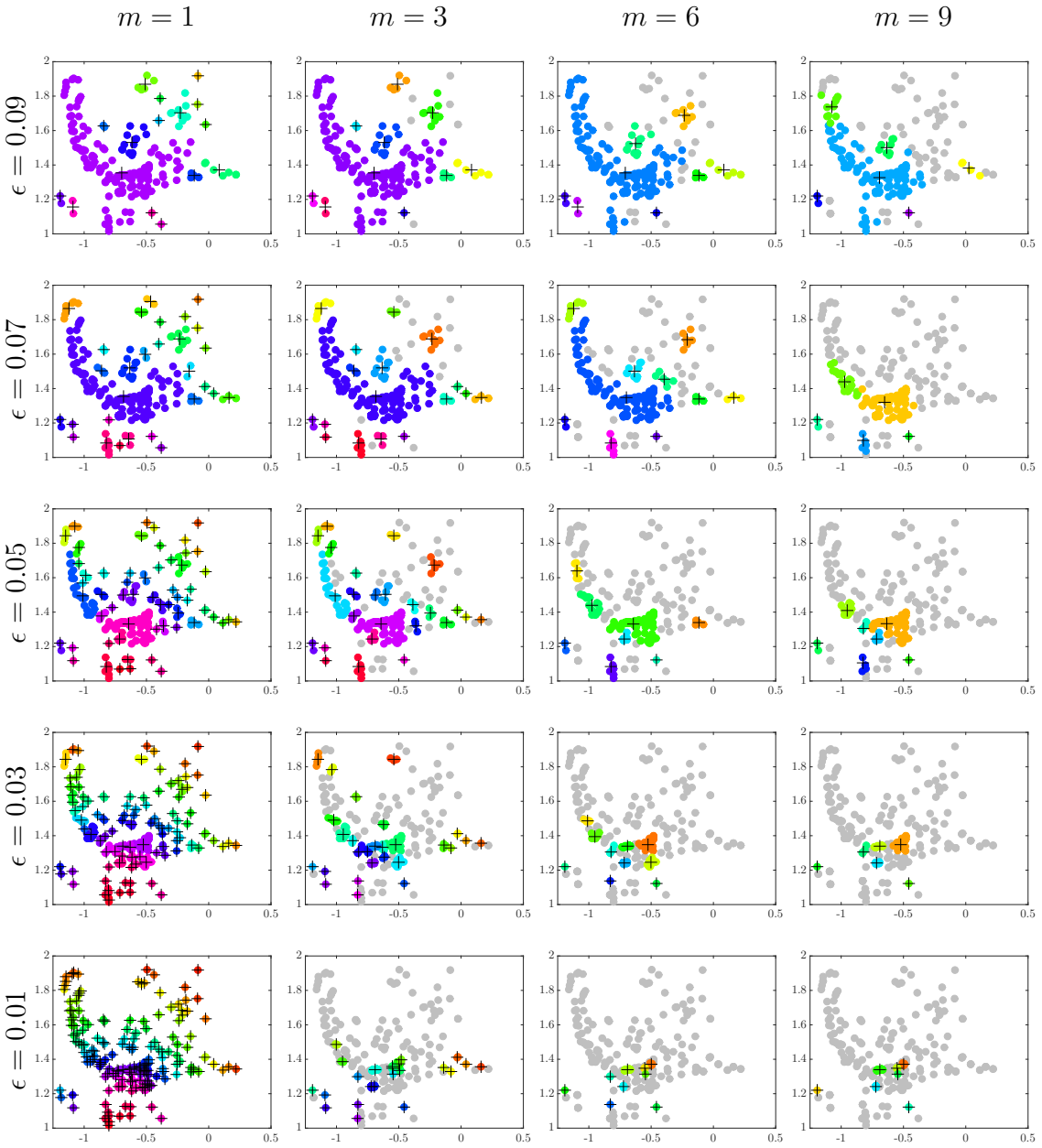


Figure 5.10: Demonstration of parameter tuning in DBSCAN using a subset of the data from the bottom row of figure 5.9 with 20 367 data points. The parameters tuned are ϵ, the minimum distance between point, on the rows; and m, the minimum number of points within a distance of ϵ required to create a cluster, on the columns.

ϵ, is reduced, more (and smaller), clusters are found, with more points identified as noise. Similarly, as the minimum number of data points required to create a cluster, m, is increased, more data points are also identified as noise. The parametrisation that appears to be most appropriate is when $\epsilon = 0.05$; and $m = 3$, but even then there are several clusters that appear unrealistic. DBSCAN therefore appears to be very sensitive to parameter tuning, and it is unclear how the clusterings would change when even more (noisy) data become available.

Conversely, the performance of the hybrid clustering algorithm is largely consistent and performs as expected regardless of the amount of data presented. Fundamentally, the challenge with applying DBSCAN to geographic coordinates is that the clusters can be expected to be roughly Gaussian, yet it is parametrised by two fairly abstract parameters. The strength of DBSCAN is to be able to create arbitrary shaped and sized clusters using only the density of the available data, which is very useful in some applications, but is not ideal for geographic location coordinates.

A further advantage of the hybrid clustering method is that it can naturally be adapted to handle new data. Although the ability to create new clusters once they have been collapsed is not available in the algorithm as presented, it would be straightforward to add this ability. Other methods, such as DBSCAN or k-means are not suited to online processing of data.

The hybrid method presented here, in common with all the methods reviewed in the introduction to this section, treats the data globally. This makes the clustering of locations more difficult, because over time more and more noise accumulates in the dataset, making locations that are clearly distinct over short periods of time appear to merge into single large clusters.

Additionally, while the methods discussed here are suitable for offline processing to demonstrate the principle of using geolocation-derived feature data as a predictor of mental state, for use in practice, data would need to be processed sequentially in an online manner. One way that online processing of geolocation data could be achieved is described as future work in section 8.3.4.

5.10 Feature extraction

Features were extracted for analysis from the preprocessed location data. As discussed in the introduction, regularity of routine and levels of activity are thought to be highly predictive indicators of depression so several features based on these principles were included as well as more general features.

5.10.1 Entropy (ENT)

The information-theoretical entropy has been proposed as a measure of variability in a signal (Shenkin et al., 1991), and was used by Saeb et al. (2015a, 2015b, 2016) and de Montjoye et al. (2013) to provide a measure of the variability in the time that participants spend in the different stationary location clusters extracted. The entropy feature is defined as

$$\text{ENT} = -\sum_{i=1}^{N} p_i \log p_i \tag{5.4}$$

where $i = 1, \ldots, N$ is the index of each location cluster extracted; and p_i is the percentage of time recorded at cluster i.

As an entropy measure, the maximum value is where the data is least predictable, i.e. where the proportion of time spent in each of the N locations is equal. This is counter-intuitive because although this would be indicative of a highly regular routine, from an information theory perspective the location as each time t would be a random draw from an N-dimensional categorical distribution with equal probabilities, i.e. the least predictable. In this case

$$\begin{aligned}
\text{ENT} &= -\sum_{i=1}^{N} \frac{1}{N} \log \frac{1}{N} && (5.5)\\
&= -N \frac{1}{N} \log \frac{1}{N} = -\log \frac{1}{N} && (5.6)\\
&= -(\log 1 - \log N) && (5.7)\\
&= \log N. && (5.8)
\end{aligned}$$

Considering another case, where time is split between N locations, but a fixed portion of the time, $q \in [0, 1]$, is spent in location 1 and the rest of the time is split between the remaining $N - 1$ locations. In this case,

$$\begin{aligned} \mathrm{ENT} &= q \log q + \sum_{i=2}^{N} \frac{1-q}{N-1} \log \frac{1-q}{N-1} && (5.9) \\ &= q \log q + (N-1) \frac{1-q}{N-1} \log \frac{1-q}{N-1} && (5.10) \\ &= q \log q + (1-q) \log \frac{1-q}{N-1}. && (5.11) \end{aligned}$$

Taking the limit as $q \to 1$, i.e. increasing the predictability of the outcome and reducing the entropy,

$$\lim_{q \to 1} \mathrm{ENT} = q \overbrace{\log q}^{0} + \overbrace{(1-q)}^{0} \log \frac{1-q}{N-1} = 0, \tag{5.12}$$

hence entropy will be in the range

$$0 \leq \mathrm{ENT} \leq \log N \tag{5.13}$$

where $\mathrm{ENT} = \log N$ indicates the greatest regularity in the data and $\mathrm{ENT} = 0$ indicates the least.

5.10.2 Normalized entropy (NENT)

As demonstrated above, the entropy measure is highly correlated with the number of clusters because its maximal value is a function of N. This correlation with the number of clusters can be removed by dividing the entropy by the maximum limit, $\log N$. The normalised entropy feature is therefore defined as

$$\mathrm{NENT} = \mathrm{ENT} / \log N, \tag{5.14}$$

resulting in an entropy measure in the range

$$0 \leq \mathrm{NENT} \leq 1. \tag{5.15}$$

5.10.3 Location variance (LV)

The variance in the recorded locations provides an indication of how much the individual is moving between different locations. The variance measure used is the sum of statistical variances in the latitude and longitude with a logarithm applied to compensate for the skewness in the distribution among participants. The location variance feature is therefore defined as

$$\mathrm{LV} = \log\left(\sigma^2_{\mathrm{lat}} + \sigma^2_{\mathrm{lon}}\right). \tag{5.16}$$

5.10.4 Home stay (HS)

The time spent at home is thought to be an important indicator of depression (Thielke et al., 2014), therefore the home stay feature is defined as the percentage of time spent at home in the data recorded. See section 5.2 for a description of how the home location is determined.

5.10.5 Transition time (TT)

The transition time feature is the percentage of all the time spent transitioning / travelling between locations in the data recorded.

5.10.6 Total distance (TD)

The total distance travelled feature is the sum of Euclidean distances between the consecutive location points recorded in the data.

5.10.7 Number of clusters (NC)

The number of clusters feature is the total number of stationary clusters extracted from the data using the hybrid clustering method described in section 5.9.1. This provides an indicator of how much the participant is moving between locations.

5.10.8 Diurnal movement (DM)

The importance of daily regularity has been discussed in section 2.2.4.5. The diurnal movement feature, developed by Saeb et al., aims to quantify this by using a frequency transform to get the power in the frequencies with wavelengths around 24 h.

Because the location data are unevenly sampled with missing data, the Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982) was used to perform the frequency transform, which fits sinusoids to data using a least squares fit. This method has advantages over the usual fast Fourier transform (FFT) because it enables extraction of frequency information from unevenly sampled data, and because it does not limit the transform to specific frequencies.

For the diurnal movement feature, the power spectral density (PSD) of the signal in selected frequencies with wavelengths between 23.5 h and 24.5 h is calculated and averaged as follows

$$E = \sum_{i=1}^{N} \mathrm{psd}\left(f_i\right) / N \tag{5.17}$$

where $\mathrm{psd}\left(f\right)$ is the power spectral density of the data at frequency f; and f_i for $i = 1, \ldots, N$ is the range of frequencies to use with wavelengths between 23.5 h and 24.5 h. This gives a measure of the energy in the spectrum with wavelengths around 24 h. For the results presented in this thesis, energies at 41 frequencies evenly spaced between 23.5 h and 24.5 h were used to calculate the DM feature.

The diurnal movement feature is formed by summing the energy values calculated for the latitude and longitude of the recorded data with a logarithm applied to compensate for the skewness in the distribution among participants. The diurnal movement feature is therefore defined as

$$\mathrm{DM} = \log\left(E_{\mathrm{lat}} + E_{\mathrm{lon}}\right). \tag{5.18}$$

The Lomb-Scargle periodogram is not affected by DC offset in the signal, but assumes stationarity and is severely affected by the magnitude of the signal.

5.10.9 Diurnal movement on normalised coordinates (DMN)

A limitation of the diurnal movement feature developed by Saeb et al. is that it is highly affected by outliers and by the distance travelled from home. For example, someone may have a highly regular diurnal pattern a short distance from home which would appear to have a low value of DM. Conversely, an individual who spends one day very far from home but the rest of their time at home will have a large DM due to the least squares fit overcompensating for the outlier day.

In order to counteract this, the diurnal movement on normalised coordinates (DMN) feature was developed to operate on a normalised set of coordinates, i.e. where the latitude and longitude were both normalised to have zero mean and unit variance within the period being classified.

5.10.10 Diurnal movement on the distance from home (DMD)

A further limitation of the diurnal movement feature developed by Saeb et al., also affecting the diurnal movement on normalised coordinates feature, is the sum of energy in latitude and longitude. This may hide important subtleties in the data. For example, consider an individual with a regular routine, but perhaps works in a job that requires a lot of travelling to different locations, often in different directions from home. The differences in direction mean that the power in either latitude of longitude will be low, even though there is clear regularity in their routine. Similarly, consider two individuals who both regularly travel the same distance from home each day for the same duration and at the same times, one travelling directly north, and one travelling north-east. The two individuals will show very different values in the DM and DMN features due to the sum of the power in latitude and longitude.

In order to counteract this, the diurnal movement on the distance from home (DMD) feature was developed to operate on the Euclidean distance from home, rather than latitude and longitude. As with the DMN feature, the distance from home was normalised to have zero mean and unit variance. In many cases, this will give a more

accurate representation of the diurnal movement of the individual because the time and distance is more important than the direction.

5.11 Feature calculation on data subsets

Saeb et al. calculated features over the whole data set being analysed, i.e. single calendar weeks in the results presented here. This may not accurately represent the characteristics of the individual. For example, someone who works Monday to Friday will most likely exhibit very different values for the normalised entropy feature during the week and at weekends.

For this reason, the features were all calculated over several different data subsets.

5.11.1 Base subset

The base subset calculates the features over the whole week being analysed.

5.11.2 Weekday subset

The weekday subset calculates the features over only the data recorded Monday to Friday. While it is likely that individuals have different work schedules and routines, overall this will provide a reasonable indication of the behavour of the participants on their working days, or at least when they are expected to be working.

5.11.3 Weekend subset

Conversely, the weekend subset calculates the features only over Saturday and Sunday. Again due to lifestyle reasons and work differences this will not accurately represent 'days-off' for all individuals, but overall will give a reasonable indication of the behaviour of individuals when they are not working.

5.11.4 Median subset

the median subset calculates the features over several further subsets of the data and then calculates the median value over these feature values. The median subset

features are calculated over the 10 feature subsets comprising of a) the base, weekend and weekday subsets described above. b) the full week with each of the 7 days removed in turn.

5.11.5 Optimised daily exclusion subset

As described in the DMN feature, the diurnal movement in particular is very sensitive to outlier days where the individual travels much further than on other days. This will also affect other features such as the location variance. In order to compensate for outliers in the recorded locations, the optimised daily exclusion subset was developed to exclude days where the individual travelled a greater distance from home than usual.

The general principle is that the maximum Euclidean distance reached from home, d_i, is calculated for each each day, $i = 1, 2, \ldots, 7$, in set $D = \{d_1, d_2, \ldots, d_7\}$. The standard deviation $\sigma_D = \text{std}(D)$ is calculated and any days where

$$d_i > \text{median}(D) + \alpha\sigma_D \tag{5.19}$$

are removed. The value of α was optimised per feature to maximise the classification accuracy using a logistic regression classifier.

5.12 Discussion

This chapter has described the pre-processing and feature extraction process applied to the raw recorded geolocation data. Several pre-processing steps were developed specifically for this application, and the features extracted from the pre-processed data were extended based on previous work by Saeb et al. (2015a) and others.

The data filtering was a crucial step required due to high levels of noise in the data. However, it is likely that if the accuracy of each location coordinate had been recorded, different techniques could have been applied. This was discussed briefly in section 4.6, and will be discussed in greater detail as future work in section 8.3.1.

The levels of data coverage presented in the previous chapter meant that imputation of missing data, where reasonable, was crucial to maximising the amount of data

suitable for accurate behavioural features to be extracted. The justifications behind the methods used were discussed in detail in section 5.5.3, but there are likely to be ways to improve the quality of the data imputation. As indicated by the battery and activity data shown in figures 5.2 and 5.3, increased accuracy of imputation could possibly be obtained by utilising the data from the other data sources collected, in a similar way to Yue et al. (2017). For example, the level of activity data recorded from the accelerometer could confirm the likelihood of someone travelling or not, and the battery level could confirm if the phone is switched off. It is unclear whether this would significantly improve results, but it is left as future work to investigate as detailed in section 8.3.1.

Clustering of stationary locations is also a crucial step for many of the features to be accurately calculated. A new method was developed for this which uses the CRP prior from the Dirichlet process to probabilistically collapse nearby data points into clusters. This was demonstrated to provide suitable and reliable clusters when applied to geolocation data, and is likely to be applicable to a wide range of geolocation data clustering tasks.

A range of features were then extracted from the pre-processed geolocation data to describe the behaviour of the individual. These features will be explored in greater detail in the next chapter, then used to estimate the mental state of the individual.

5.12.1 Alternative techniques

In many fields the traditional method of manually identifying and extracting features (as done here) is increasingly being replaced by more automated feature extraction methods, commonly based on variants of artificial neural networks (Goodfellow et al., 2016). Artificial neural networks, as implied by the name, are inspired by the neural architecture of the brain. The motivation is that the brain has evolved over millennia through natural selection to perform a wide range of tasks that traditional machine learning tools are still unable to match. If an architecture similar to that of the brain could be implemented *in silico* then that could enable computers to perform a much wider range of tasks.

At the most basic level, artificial neural networks are formed as a network of interconnected nodes, arranged in layers. The individual nodes are generally relatively simple, often sigmoidal, functions of the input values to each node. Data are generally connected to the lowest layer, with processing performed in the higher layers, and the result provided by the top layer. Each of the connections between nodes are weighted, and the optimisation of the interconnecting weights leads to the learning behaviour of the network as whole. Different data transformation can be performed in the different layers, and convolution operations are often applied. Although the individual nodes are relatively simple functions, the power of artificial neural networks comes from the use of extreme numbers of nodes per layer, with tens or hundreds of layers. State-of-the-art image recognition neural networks, such as the version of ResNet (He et al., 2016) that won the 2015 ImageNet competition (Russakovsky et al., 2015; UNC Vision Lab, 2015), had 152 layers and 60 million parameters (Sze et al., 2017). This demonstrates the computational challenge of implementing these networks in practice. Therefore, although the concept of artificial neural networks dates back to the 1960's when the neural architecture of the brain was being discovered, it was not until the second decade of the 21st century that advances in computational ability made them practicable tools for machine learning.

Both manual and automated methods have their advantages and disadvantages. Manual feature extraction arguably requires greater understanding of the data; greater understanding of the application domain in order to identify relevant features; and greater knowledge of the range of signal processing tools available to implement the identified features. Automated methods based on artificial neural networks place greater emphasis on knowledge of the design and configuration of network architectures, instead of the application domain (although an understanding of the application domain is helpful). Generally speaking, the main advantage of manual methods is that the features can be developed to be easily interpretable, while the main advantage of automated methods is their ability to identify features in the data that are difficult (or impossible) to extract manually.

Artificial neural networks have been demonstrated to provide state-of-the-art performance in many areas of signal processing and image analysis (Najafabadi et al., 2015). However they have seen relatively little application to the field of geolocation data processing. Common geolocation-related applications include estimating geographic location from sensory data such as photographs (Weyand et al., 2016); or estimating indoor location from wireless access points (Rahman et al., 2012). Due to their inherent ability to detect patterns in data, another area of interest is predicting the next location of a user from current (and past) location coordinates or visited places (Xu et al., 2017). Similarly, Sobrado (2018) used a neural network to estimate missing geolocation data by using a neural network to build a personalised map of locations that a user commonly visits. Jaques et al. (2017) also describe a system for estimating missing sensor data in a multimodal sensor system using deep denoising autoencoders trained on the remaining available sensors. On a larger scale, similar techniques can be used to predict crowd flows (Zhang et al., 2017) or traffic levels (Ma et al., 2015). Suhara et al. (2017) applied neural networks to mood prediction, but do not use raw sensor data as inputs. Similarly, Mikelsons et al. (2017) used geolocation-derived features (similar to those presented earlier in this chapter) with a neural network to estimate stress levels in the StudentLife data set (see section 2.2.4.1 for details). However, except for one upcoming paper by Mehrotra & Musolesi (2018) due to be presented and published later this year, there does not appear to be any previous work that directly uses neural networks on geolocation coordinates.

One of the reasons for this may be that geolocation data do not fit naturally into the neural network framework like other signal processing applications. When using neural networks for identifying objects in images, for example, the same object should, by definition, appear similar in all images. This is not the case for geolocation traces which are much more individualised.

An alternative method worth mentioning is using an end-to-end integrated Bayesian modelling approach to feature extraction and classification / regression. In chapter 7,

a Bayesian model is used to estimate depression levels from the feature values presented here, where the features are presented to the model as point-estimates. However, being a Bayesian approach, the location coordinates and features could instead be modelled as (conditioned) random variables, which would naturally have an element of uncertainty associated with them. This could eliminate the need to impute location coordinates because missing data could be incorporated into the distributions of the features. One potential way to design such a model is presented as future work in section 8.3.4.

Chapter 6

Global Models of Mental Health

The previous chapter described the features extracted from the geolocation data. This chapter explores the relationship between mental health state, primarily depression, and these geolocation-derived features. Mental health questionnaires such as QIDS or PHQ-9 are commonly used to assess patient depression levels. Objective measures based on sensors such as geographic location movements could be used to estimate the questionnaire score or to provide an objective indication of whether the patient is depressed or not. This has the potential to inform mental health professionals of the individuals currently most at risk without needing to wait for a routine appointment or for the patient or a relative to contact them. This could potentially improve outcomes since patients can get the treatment that they need more promptly.

This chapter is structured as follows. Section 6.1 first explores the properties of the geolocation-derived features described in the previous chapter and details several filtering steps applied to the raw feature values to improve predictive performance. Section 6.1.2.2 describes an adaptive Kalman filter to reduce noise in the feature space; and section 6.1.3 describes a method of excluding non-representative weeks of data from analysis. Three different automated identification tasks are then considered using global population-level models. Section 6.3 demonstrates binary classification of depression from the recorded geolocation data from the BD participants in the AMoSS study; and trinary classification of the three participant cohorts (HC; BD; and BPD). Finally, section 6.4 explores the ability to detect the level of depression in the BD participants using global regression models.

6.1 Feature properties and optimisation

This section explores the properties of the raw features calculated using the methods described above. To demonstrate the relevance of the features for classification, each feature is classified individually based on the output variable of interest. While combinations of features selected using an appropriate feature selection technique will generally perform better than features used individually, this is a simple way to demonstrate performance of the features.

The method used to demonstrate feature performance is logistic regression classification (see section 3.1.3) with leave-one-participant-out cross-validation and group equalisation (see sections 3.5 and 3.5.1). In higher dimensions, logistic regression is often outperformed by other classifiers with more complex decision boundaries, but in a single dimension, the way it is optimised means that it provides a reliable metric with reasonable performance.

6.1.1 Data subsets

As described in section 5.11, all the features were calculated on several different data subsets. The effect of calculating the features on the different data subsets can be seen by training a logistic regression classifier as described above to classify between depressed and non-depressed using the QIDS = 11 threshold. The results of the classification of each feature on the BD participants are shown in figure 6.1. This shows the distribution of classification accuracy for each of 100 iterations of group equalisation for each left out participant.

The first 10 groups show the classification accuracy for the model trained on each of the base features on the different data subsets. The last column in figure 6.1 shows the distribution of median classification performance for each feature. This provides a useful summary metric that will be used to compare these results in the following sections.

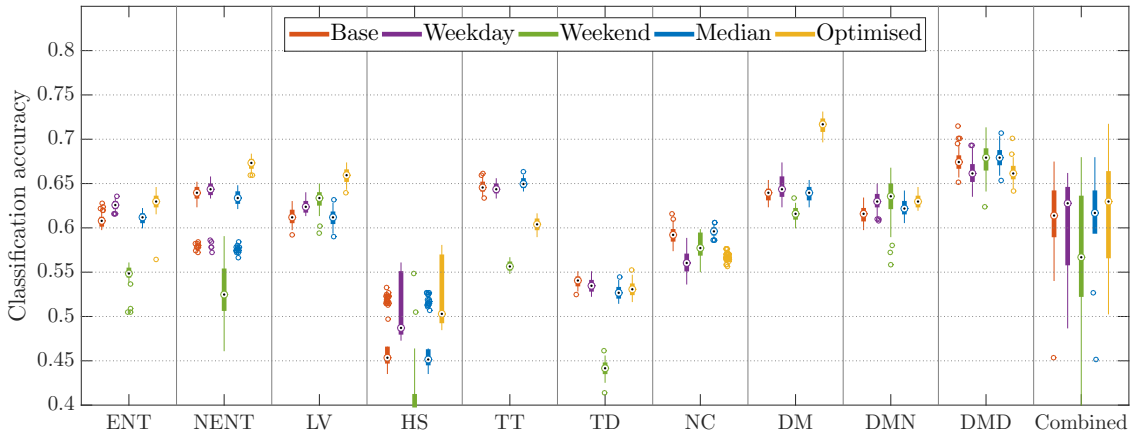


Figure 6.1: Feature performance for classification of depression (based on a threshold of QIDS scores = 11 on labelled calendar weeks of data) using features calculated on different data subsets for BD participants. The first 10 groups show the distribution of classification accuracy over 100 iterations of group equalisation for each left out participant and the last column shows the distribution of the median classification accuracy for each of the 10 features.

6.1.2 Feature smoothing

As discussed above, features are calculated over calendar weeks of data for the different data subsets. This means that the features are sensitive to small differences in behaviour between weeks. Depression on the other hand is a long-term psychiatric illness without rapid changes in state. For this reason, the features can be thought to have an element of unexplained variance between consecutive weeks, i.e. noise. The feature data can be smoothed to reduce the effects of this unwanted variance.

Farhan et al. (2016b) performed a similar smoothing on data from the StudentLife study (see section 2.2.4.1) using Haar wavelet filtering. However, as discussed below, this is an offline method that requires future data to compute.

6.1.2.1 Window filtering

The simplest method of smoothing the recorded feature data is to apply a window filter of length ω across the recorded data. Each data point is replaced with the mean value of it and the surrounding $\omega - 1$ data points. For example, for a window of length $\omega = 5$ weeks, a point at time t is replaced with the mean value of the points at times $t' = \{t-2, \ldots, t, \ldots, t+2\}$. The window does not need to be central, i.e. a window of length $\omega = 5$ weeks can also be defined as the mean value of the points at times $t' = \{t-4, \ldots, t\}$, however this can be recognised as a window centred on $t-2$.

The change in median classification accuracy for a range of different windows compared to the unfiltered data is shown in figure 6.2. The largest improvement in classification accuracy is obtained with a centred window of length $\omega = 7$, which provides a modest overall improvement over the unfiltered data for all data subsets. Note that the improvement in performance is least for the optimised daily exclusion subset (described in section 5.11.5) which is unsurprising given that the removal of days spent far from home can also be considered as a method of noise reduction. While performance does not improve for all features, there is an overall improvement from window filtering of any length.

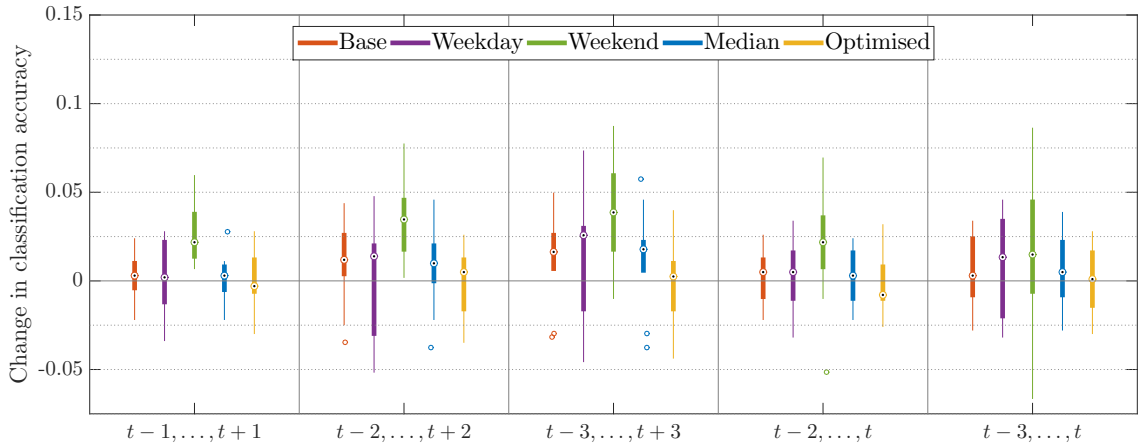


Figure 6.2: Change in median classification performance on classifying depression using individual window-filtered features calculated on different data subsets for BD participants compared to unfiltered features shown in figure 6.1.

6.1.2.2 Kalman filtering

Window filtering the data has a number of limitations. Firstly, as discussed above, centred windows make most sense conceptually but require future data. This makes them prohibitive for online use, similar to the wavelet filtering used by Farhan et al. (2016b). Secondly, in the case of a standard mean filter as used above, all samples within the window are treated equally. This means that explained variance in the sample may end up filtered out as noise. Finally, if there are missing points within the window then the filter will essentially apply a smaller window to the remaining points. This may end up with inconsistent results.

An alternative way to filter the data is using the Kalman filter (Kalman, 1960; Faragher, 2012), a classic smoothing and prediction method that attempts to predict the true value of a state from a (noisy) prediction from the last state and a (noisy) measurement of the current state. More formally, the Kalman filter models the value of the next state $\mathbf{x}_t$ as

$$\mathbf{x}_t = \mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{B}_t\mathbf{u}_t + \mathbf{w}_t \tag{6.1}$$

where $\mathbf{x}_{t-1}$ is the value of the previous state at time $t-1$; $\mathbf{u}_t$ is a control signal at time t; and $\mathbf{w}_t$ is the noise at time t which is zero-mean Gaussian distributed with covariance $\mathbf{Q}_t$. Matrices $\mathbf{F}_t$ and $\mathbf{B}_t$ define the effect of the previous state and control signal on the next state. Measurements $\mathbf{z}_t$ at time t can also be incorporated (if available) and modelled as

$$\mathbf{z}_t = \mathbf{H}_t\mathbf{x}_t + \mathbf{v}_t \tag{6.2}$$

where $\mathbf{H}_t$ is a linear mapping of the state values into the measurement domain (in some applications $\mathbf{H}_t = \mathbf{I}$); and $\mathbf{v}_t$ is the noise in each measurement at time t which is zero-mean Gaussian distributed with covariance $\mathbf{R}_t$.

The Kalman filter is implemented as follows:

$$\hat{\mathbf{x}}_{t\,|\,t-1} = \mathbf{F}_t\hat{\mathbf{x}}_{t-1\,|\,t-1} + \mathbf{B}_t\mathbf{u}_t \qquad \text{(state prediction)} \tag{6.3a}$$

$$\mathbf{P}_{t\,|\,t-1} = \mathbf{F}_t\mathbf{P}_{t-1\,|\,t-1}\mathbf{F}_t^\top + \mathbf{Q}_t \qquad \text{(state prediction covariance)} \tag{6.3b}$$

$$\mathbf{K}_t = \mathbf{P}_{t\,|\,t-1}\mathbf{H}_t^\top \underbrace{\left(\mathbf{H}_t\mathbf{P}_{t\,|\,t-1}\mathbf{H}_t^\top + \mathbf{R}_t\right)}_{\substack{\text{measurement prediction} \\ \text{covariance}}}^{-1} \qquad \text{(Kalman gain)} \tag{6.3c}$$

$$\hat{\mathbf{x}}_{t\,|\,t} = \hat{\mathbf{x}}_{t\,|\,t-1} + \mathbf{K}_t \underbrace{\left(\mathbf{z}_t - \mathbf{H}_t\hat{\mathbf{x}}_{t\,|\,t-1}\right)}_{\text{measurement residual}} \qquad \text{(updated state estimate)} \tag{6.3d}$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t\,|\,t-1} - \mathbf{K}_t\mathbf{H}_t\mathbf{P}_{t\,|\,t-1} \qquad \text{(updated state covariance)} \tag{6.3e}$$

The state prediction in equation (6.3a) is derived directly from the model definition in equation (6.1). The state prediction covariance in equation (6.3b) can be derived by noting that the covariance of a prediction $\hat{\mathbf{x}}_{t\,|\,t-1}$ from the true value $\mathbf{x}_t$ can be given by $\mathbf{P}_{t\,|\,t-1} = \mathbb{E}\left[\left(\mathbf{x}_t - \hat{\mathbf{x}}_{t\,|\,t-1}\right)\left(\mathbf{x}_t - \hat{\mathbf{x}}_{t\,|\,t-1}\right)^\top\right]$ and substituting in equations (6.1) and (6.3a). Equations (6.3c) to (6.3e) can be derived from the Kalman filter definition that the predicted state and measured value are modelled as two Gaussian distributions that are multiplied together to form another Gaussian distribution from which equations (6.3d) and (6.3e) can be read. The parameters can be represented as follows, with the first column showing the parameters of the predicted state distribution; the second showing the parameters of the measured value distribution; and the third showing the parameters extracted from the output distribution:

$$\mu_1 = \hat{\mathbf{x}}_{t|t-1} \qquad \mu_2 = \mathbf{H}_t\mathbf{z}_t \qquad \hat{\mathbf{x}}_{t\,|\,t} = \mu_{out} \tag{6.4a}$$

$$\sigma_1^2 = \mathbf{P}_{t|t-1} \qquad \sigma_2^2 = \mathbf{H}_t\mathbf{R}_t \qquad \mathbf{P}_{t\,|\,t} = \sigma_{out}^2 \tag{6.4b}$$

In this form, the Kalman filter can be said to give a maximum-likelihood estimation of the true value based on the distributions of the predicted and measured values. The Kalman gain $\mathbf{K}_t$ calculated in equation (6.3c) and used in equation (6.3d) determines how much influence the measurement has on the output.

The Kalman filter for smoothing the geolocation features is defined with $\mathbf{F}_t = \mathbf{I}$ (there is no underlying model of $\mathbf{x}_{t\,|\,t-1}$); $\mathbf{H}_t = \mathbf{I}$ (the measurements $\mathbf{z}_t$ are in the same

domain as the state $\mathbf{x}_t$); and $\mathbf{B}_t = \mathbf{0}$ (there is no control input to the model). This simplifies the full Kalman filter model in equations (6.1) and (6.2) to

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{w}_t \tag{6.5a}$$

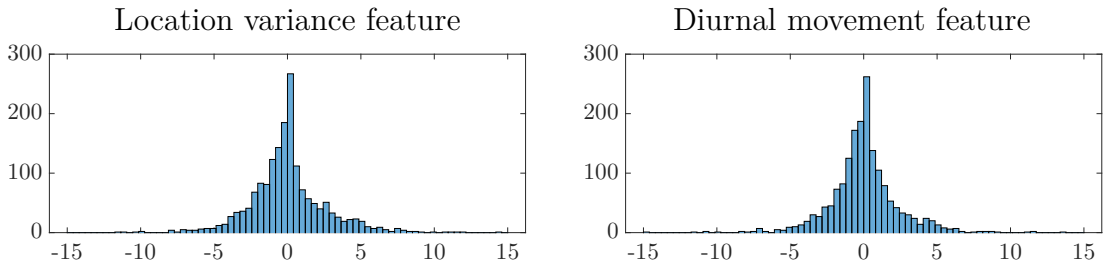
$$\mathbf{z}_t = \mathbf{x}_t + \mathbf{v}_t \tag{6.5b}$$

leaving the free parameters as $\mathbf{Q}_t$ (covariance of $\mathbf{w}_t$); and $\mathbf{R}_t$ (covariance of $\mathbf{v}_t$). These simplifications ignore any underlying dynamics in the feature values. It is unclear, however, whether incorporating a model of feature dynamics would improve results or impose unrealistic assumptions on the feature space. For example, if someone is tending to spend increasing amounts of time at home, it is unclear if this trend should be expected to continue. Incorporating such a dynamical model may negatively bias the results. Further investigation of the underlying dynamics in the feature space may however have clinical relevance and is left as future work as discussed in section 8.3.2.

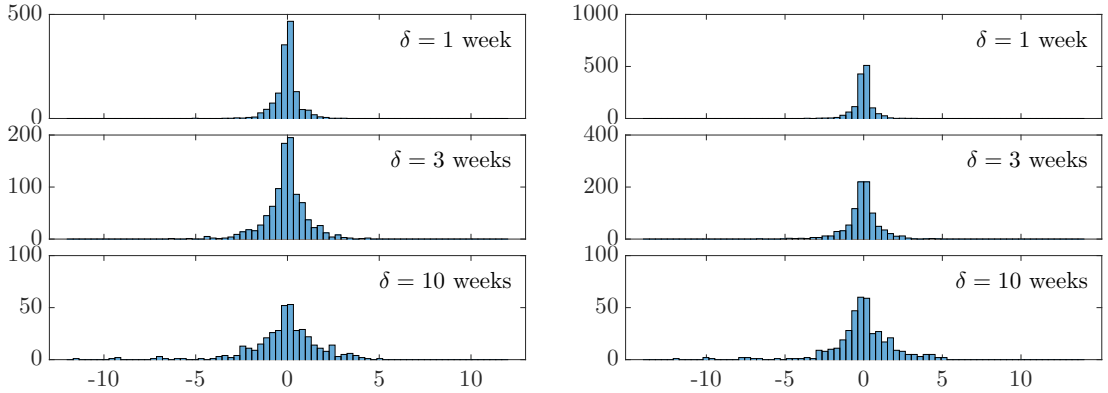
For simplicity, both $\mathbf{Q}_t$ and $\mathbf{R}_t$ were assumed to be diagonal meaning that each feature is treated independently. This ignores any interactions between the different features. The feature correlation coefficients given in appendix B show that there are strong correlations between many of the calculated features. These correlations could be incorporated in the smoothing model, however this may be counter-productive since noise in one feature would likely also show up in any correlated features. Smoothing each feature independently avoids unintentionally incorporating additional noise from correlated features in the output.

The Kalman filter provides a mechanism to emulate the window filtering described in section 6.1.2.1 without having to use future data. In order to do this, the values of the $\mathbf{Q}_t$ and $\mathbf{R}_t$ parameters are determined from window-filtered training data. The raw feature data were window-filtered with a 7 week centred mean filter.

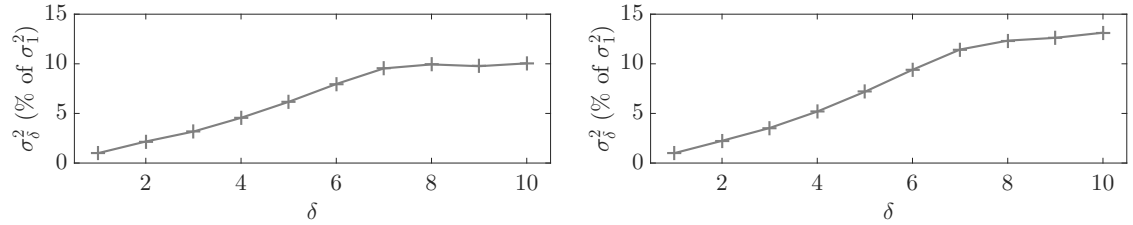
The value of the $\mathbf{R}_t$ parameter is estimated from the differences between points in the raw and filtered data. This is demonstrated in figure 6.3(a), which shows histograms of the difference between the raw filter data and the filtered data for the location variance and diurnal movement features. In this model, the $\mathbf{R}_t$ parameter is



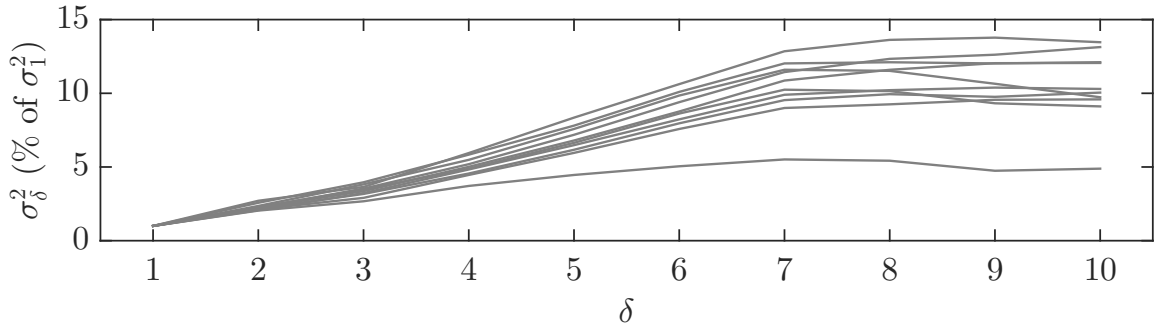
(a) Histogram of differences between filtered data and raw feature data for each data point. The variance of the differences for each feature i are used on the diagonal of parameter $\mathbf{R}_t$.



(b) Histograms of differences between consecutive data points on the top row; each pair of points separated by 3 weeks on the second row; and each pair separated by 10 weeks on the bottom row.



(c) Variances of differences between pairs of filtered points separated by δ weeks. This is used as parameter $\mathbf{Q}_t$ depending on the number of weeks between the measurements at t and $t-1$.



(d) Variances of differences between pairs of filtered points separated by δ weeks for 10 base features.

Figure 6.3: Kalman filter parameters derived from window-filtered feature data for LV feature in the left-hand column; and DM feature in the right-hand column. Feature data were filtered by a 7 week centred mean window function.

time invariant and takes the variance of the differences between the raw filter data and the filtered data for each feature individually along the diagonal.

The value of $\mathbf{Q}_t$ is estimated from the differences between consecutive filtered data points as demonstrated in the top row of figure 6.3(b). The diagonal of $\mathbf{Q}_t$ takes the variance of the consecutive differences for each feature individually. To enable handling missing data, different values of $\mathbf{Q}_t$ are calculated from pairs of points separated by δ weeks. This is demonstrated in the rows of figure 6.3(b) where the variance generally increases as δ (the number of weeks between pairs of points) increases. Figure 6.3(c) shows the variances for the two features for values of δ from 1 to 10, and figure 6.3(d) shows the variances of the 10 base features calculated over the whole calendar weeks of data. Both figures 6.3(c) and (d) show the variances for different δ as a percentage of the variance when $\delta = 1$. Again it can be seen that the variance generally increases as the distance between the pairs of points increases.

Figure 6.3(d) also shows that the relationship between δ and the variance of differences between points is fairly linear until $\delta = 7$ at which point the variance tends to flatten out (or in some cases decrease). It is also clear from figures 6.3(a) and (b) that the underlying distributions are not actually Gaussian, but appear to be more similar to Laplace distributions. In practice this is unlikely to have a major impact since they are both symmetrical unimodal distributions.

The change in median classification accuracy of the Kalman-filtered individual features compared to the unfiltered data is shown in figure 6.4 for different Kalman filters. The different Kalman filters were trained on window filtered data as described above with different configurations of window filter. The classification results are generally improved over the unfiltered data and are also improved over the window-filtered data for the same window used to train the Kalman filter. The best overall results were obtained from training the Kalman filter on a 7 week centred mean window filter, which is the same window filter that provided the largest improvement in classification accuracy in figure 6.2. Similar to the window filtered results in figure 6.2, not all individual features are improved by the filtering, but overall there

is a modest overall improvement. The performance of the individual features is also shown in figure 6.6(a) on page 153.

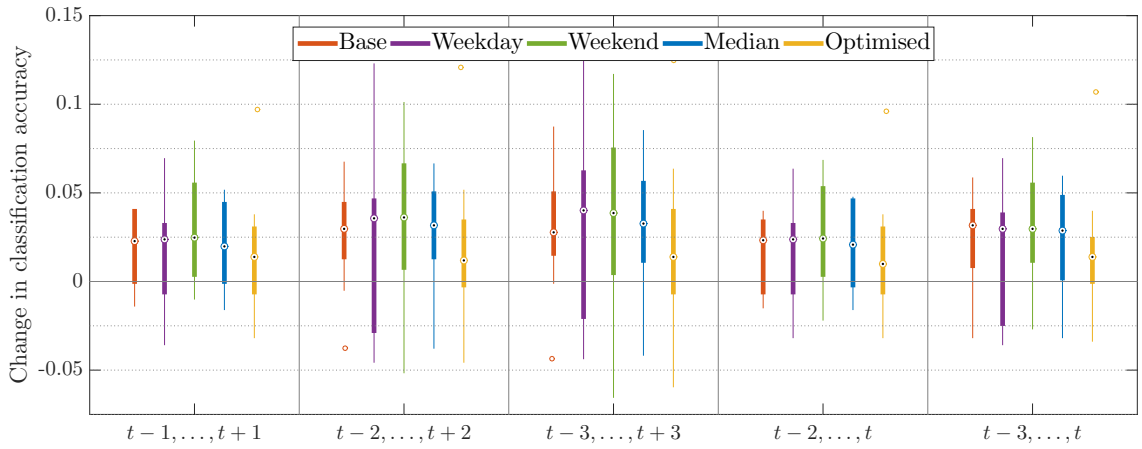


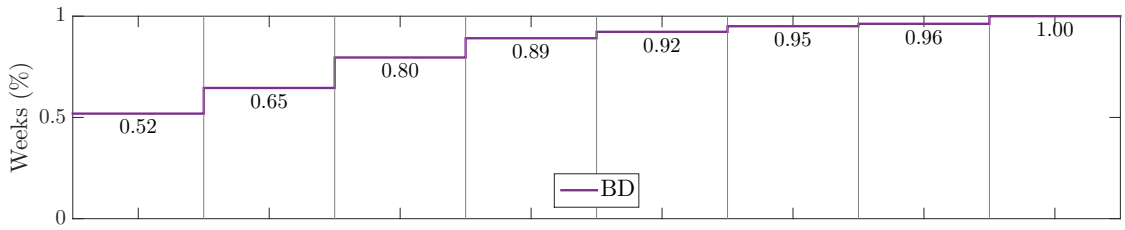
Figure 6.4: Change in median classification performance on classifying depression using individual Kalman-filtered features calculated on different data subsets for BD participants compared to unfiltered features shown in figure 6.1. Different Kalman filters were trained using mean window filters applied to the data with the bounds on the horizontal axis.

6.1.3 Filtering outlier data

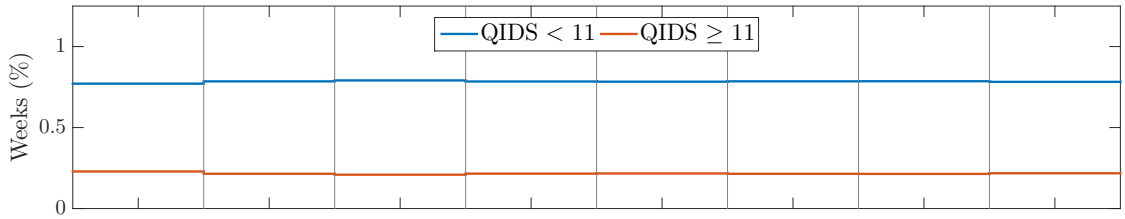
The features extracted are designed to characterise individuals in the community, in and around home. Weeks with large amounts of missing data, or where individuals are far away from home for extended periods (on holiday or work trips for example) can be considered as outliers that may distort the feature values leading to misleading results. Although the weeks for classification were selected based on whether they had sufficient data available for analysis, this was not a formal process. For these reasons, a number of tests were run to remove outlier weeks (before Kalman filtering the remaining weeks) based on a) the amount of missing data; b) the amount of time spent far from home (defined as $> 50\,\text{km}$ from home); or c) the total amount of time missing or spent far from home. It is accepted that there may be information in the weeks that are removed from analysis based on these factors. If a patient is spending a lot of time away from home, for example, this may be positive (being on holiday); neutral (being on a work trip); or negative (perhaps being hospitalised) with regard to mental state. However, because there are many different possible explanations for why the patient is away from home, interpreting the data is challenging without additional external information.

The most effective method to remove outliers was found to be filtering based on the sum of the amount of missing data and the amount of time spent far ($> 50\,\text{km}$) from home during each week. Results from removing outliers based on this metric are shown in figure 6.5. After removing outliers, the remaining data are Kalman filtered as previously described. Figure 6.5(c) shows the distribution of median classification results on individual features as outliers are removed at different thresholds. It can be seen that as the threshold becomes more restrictive (towards the left of the figure), the classification results improve for most data subsets. Figure 6.5(a) shows the proportion of all samples remaining after removing outliers while figure 6.5(b) shows the proportion of depressed and non-depressed samples remaining in the data set.

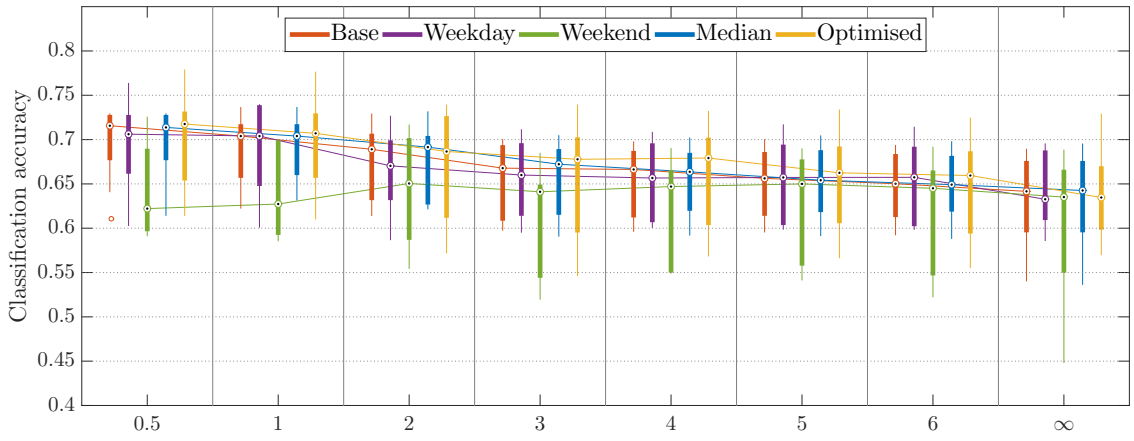
Even removing just the 5 % of weeks where the individual is away or has missing data for over 5 days in a week shows an improvement in classification performance.



(a) Percentage of labelled weeks remaining after filtering.



(b) Proportion of all labelled weeks in each group after filtering.



(c) Overall classification performance after filtering.

Figure 6.5: Feature performance for classification of depression (based on a threshold of QIDS scores = 11 on labelled calendar weeks of data) using Kalman-filtered features calculated after removing outlier weeks based on the total amount of missing data and time spent > 50 km from home.

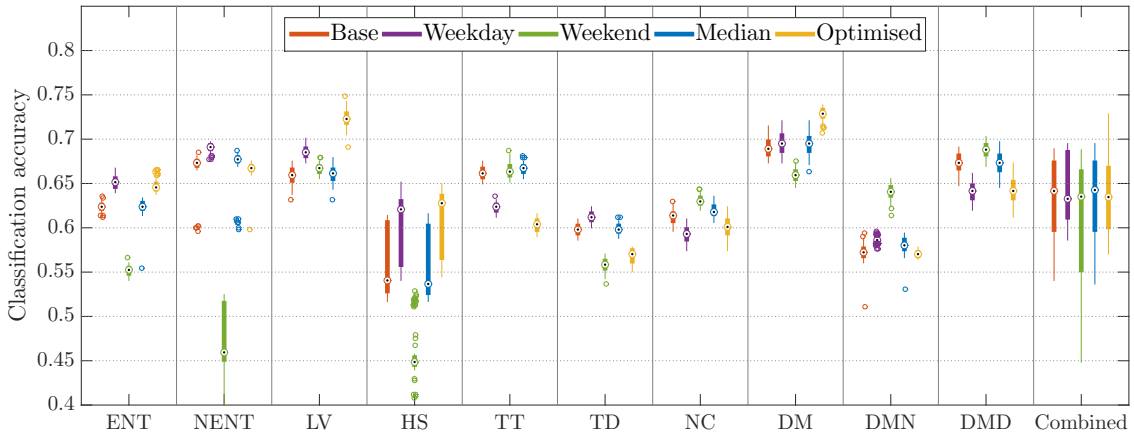
Classification performance continues to improve as the data are more heavily filtered. Note that throughout the thresholds tested, the proportions of depressed and non-depressed weeks (shown in figure 6.5(b)) do not substantially change, meaning that the algorithm is not biassed towards removing any one of the classes. Choosing the optimal threshold is a trade-off between maximising performance and not removing an excessive amount of data. A threshold of two days provides a substantial improvement in performance while only removing 20 % of the available data.

6.1.4 Feature optimisation results

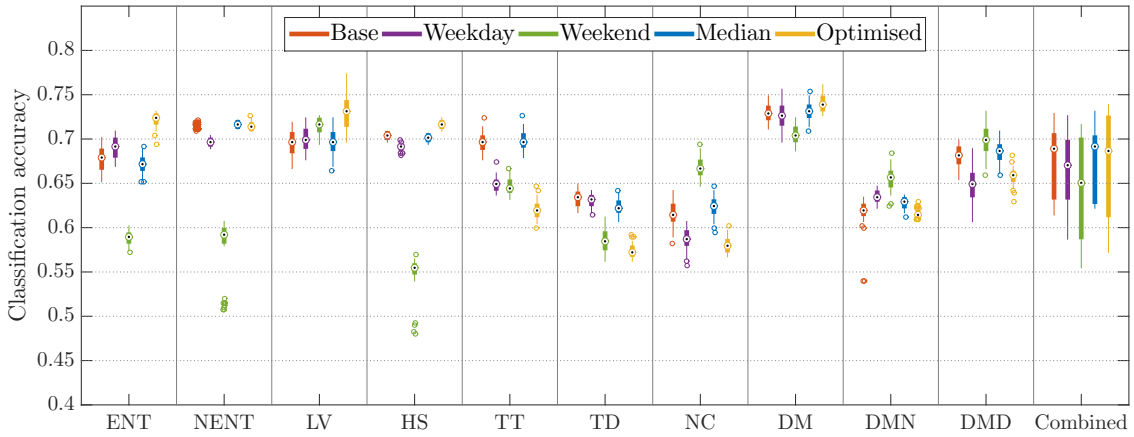
For a more detailed view of the results in figures 6.4 and 6.5(c), figure 6.6 shows the full set of classification results for all features calculated on the different data subsets under (a) the Kalman-filtered data; and (b) the Kalman-filtered data after outlier weeks where over two days of data are missing or spent far from home have been excluded. Additionally, figure 6.7 on page 154 shows the change in performance from the unfiltered results shown in figure 6.1.

The Kalman filtered data in figures 6.6(a) and 6.7(a) demonstrates that there has generally been an improvement in performance for most features over the unfiltered data. After removing outlier weeks in figures 6.6(b) and 6.7(b) the results improve further. Especially notable is the improvement in performance of the HS feature, which would be heavily affected by time spent away from home.

→

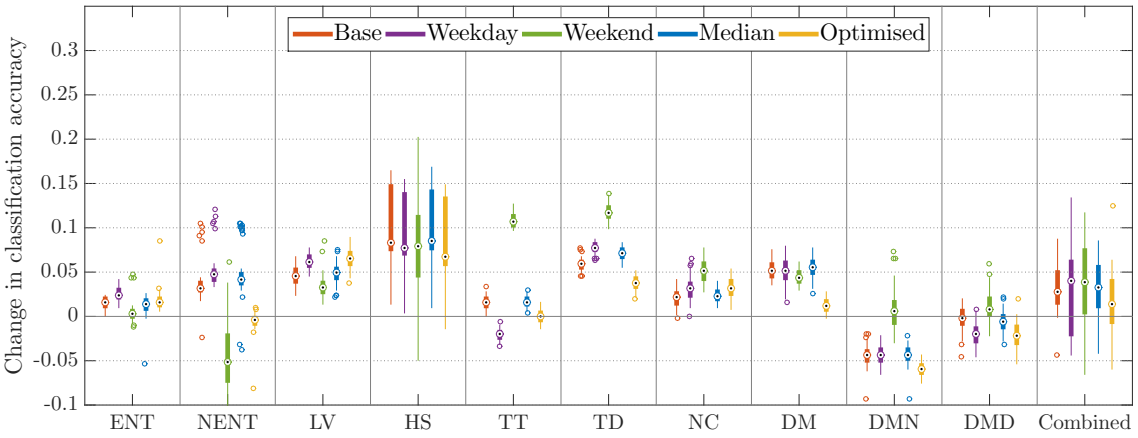


(a) Kalman-filtered data.

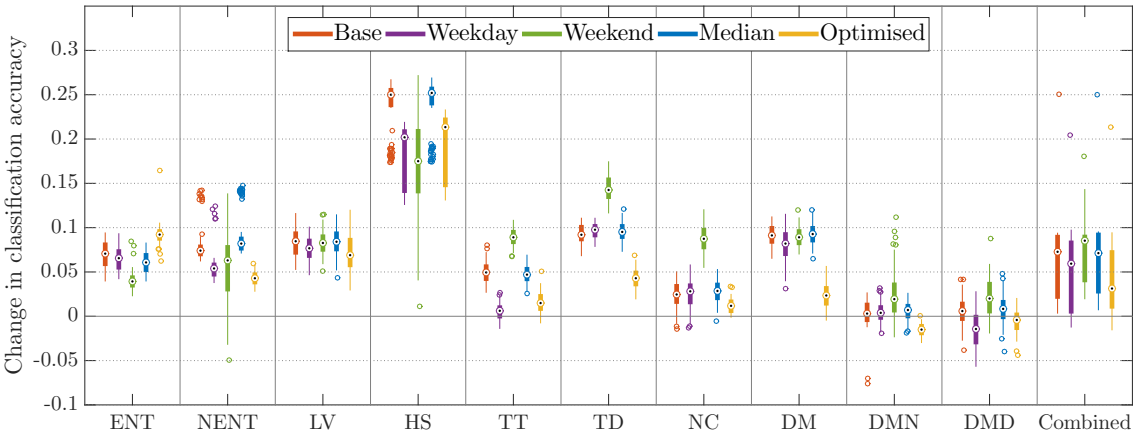


(b) Kalman-filtered data allowing a maximum of two days missing or > 50 km from home.

Figure 6.6: Feature performance for classification of depression (based on a threshold of QIDS scores = 11 on labelled calendar weeks of data) using Kalman-filtered features calculated on different data subsets for BD participants. The first 10 groups show the distribution of classification accuracy over 100 iterations of group equalisation for each left out participant and the last column shows the distribution of the median classification accuracy for the 10 features in each data subset. The change in classification accuracy from the performance of the unfiltered features shown in figure 6.1 is shown in figure 6.7.



(a) Kalman-filtered data.



(b) Kalman-filtered data allowing a maximum of two days missing or > 50 km from home.

Figure 6.7: Change in feature performance for classification of depression using Kalman-filtered features calculated on different data subsets for BD participants shown in figure 6.6 compared to unfiltered features shown in figure 6.1. The first 10 groups show the change in classification accuracy over 100 iterations of group equalisation for each left out participant and the last column shows the change in the median classification accuracy for the 10 features in each data subset.

6.1.5 Feature statistics

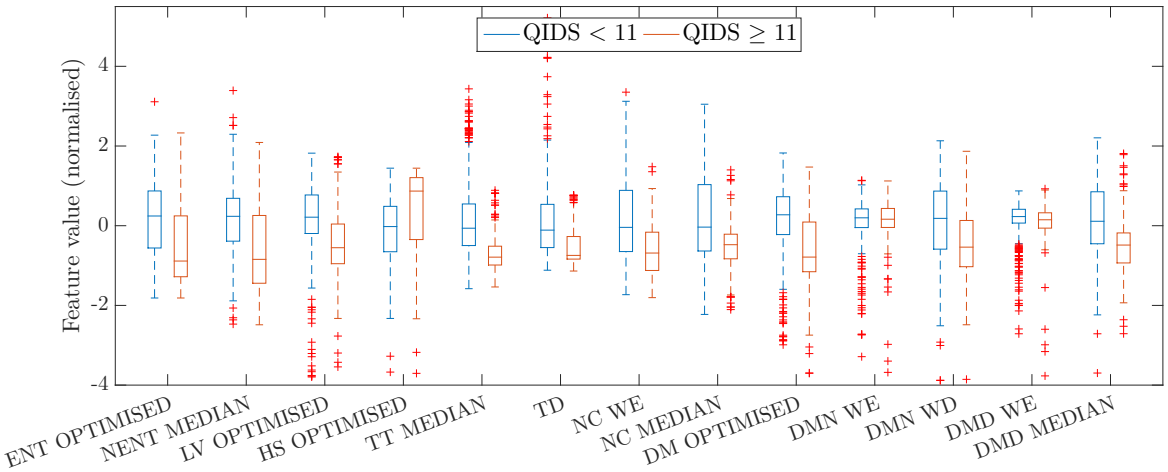
Figures 6.6 and 6.7 provide a useful overview of classification performance from the individual features calculated on the different data subsets. It can also be useful to explore the distributions of the individual features as box plots. This is shown in figure 6.8 on the following page where each feature is shown for BD participants from the data subset that provided the highest classification accuracy in figure 6.6(b). As with figures 6.6 and 6.7, features are shown calculated on (a) the Kalman-filtered data; and (b) the Kalman-filtered data after outlier weeks where over two days of data are missing or spent far from home have been excluded. Where the highest performing data subset is the weekend data subset, the next highest performing data subset is also shown. The raw (Kalman-filtered) feature values have been normalised to zero-mean, unit-variance for each feature over all samples for ease of display.

The general trends shown in figure 6.8 are as expected for both sets of data. The features focussing on daily regularity (ENT; NENT; DM; DMN; and DMD features) all demonstrate that depressed individuals have lower levels of regularity. Likewise, the HS feature demonstrates that depressed individuals spend more time at home and conversely the LV; TT; TD; and NC features demonstrate that depressed individuals travel less and therefore visit fewer locations. Comparing figures 6.8(a) and 6.8(b) shows that removing outlier weeks emphasises the differences in feature values between the two classes.

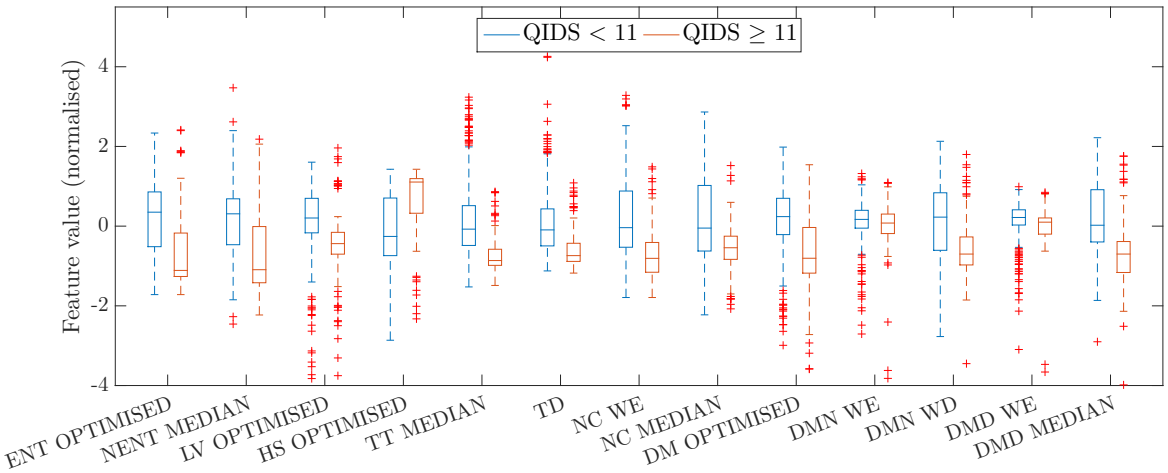
6.2 Feature data summary

To summarise the feature optimisations described in the previous sections, all results in the rest of this thesis are run on feature data where

a) the total amount of missing data (after preprocessing) and time spent $>$ 50 km from home in a week is less than two days (48 h); and
b) a Kalman filter trained on 7 week centred mean window filtered data has been applied to the remaining data.



(a) Kalman-filtered data.



(b) Kalman-filtered data allowing a maximum of two days missing or > 50 km from home.

Figure 6.8: Feature distributions of Kalman-filtered features calculated on BD participants. Each feature is shown calculated on the the data subset that provided the best performance in figure 6.6(b). Where the highest performing data subset for a feature is the weekend data subset, the feature calculated on the next highest performing data subset is also shown. Feature values have been scaled to zero-mean, unit-variance.

6.3 Classification tasks

The following classification tasks are explored in this section using the geolocation-derived features:

a) binary classification of depression in BD participants; and

b) trinary classification of the mental health conditions of participants in the AMoSS study (HC; BD; or BPD).

6.3.1 Methodology

6.3.1.1 Model validation

In addition to leave-one-participant-out cross validation, 10-fold and 5-fold cross-validation methods were used to validate the models. The labelled weeks from 10 % and 20 % of participants were used for testing while the model was trained on the labelled weeks for the remaining participants. Additionally, 3-fold cross-validation was used, but instead of leaving whole participants out, the data for each participant was split into the three partitions and cross-validation performed like usual, hence allowing data for a participant to be used for both training and testing.

6.3.1.2 Group equalisation

In most cases of cross-validation described above, the classes being classified were not equally sized in the training set. The training data were therefore adjusted by sub-sampling the over-represented class as described in section 3.5.1.

6.3.1.3 Classification models

Classification tasks were performed using logistic regression; naïve Bayes; and quadratic discriminant analysis models described in sections 3.1.3 to 3.1.5 respectively.

6.3.1.4 Feature selection

Most regression and classification models are sensitive to the combination of features presented to them as discussed in section 3.6. The greedy forward feature selection

method described in section 3.6.2.1 and algorithm 3.3 was used to incrementally select the best performing features combined with the previously selected features.

For two-class classification tasks, features were selected based on the F_1 score of each classification. For classification tasks with more than two classes, features were selected based on the mean of the F_1 scores calculated for each class.

6.3.2 Identification of depression

The first task is to investigate whether depression can be detected from the extracted geolocation features in the BD participants in the AMoSS study. This is a standard supervised binary classification problem. As described in section 4.4.1.1, the QIDS questionnaire has a standard clinically accepted threshold for moderate depression of 11. This is therefore used as the depression threshold for classification purposes. Analysis is limited to the BD participants because most HC participants in the study did not experience depression (see table 5.3); and self-reported depression questionnaire scores are not well researched or understood in BPD patients.

Figure 6.9 and table 6.1 show the results of depression classification with leave-one-participant-out cross-validation and 100 iterations of group equalisation for each left out participant. Classification results are shown using a logistic regression model and features were selected using the greedy forward feature selection method. Logistic regression was found to provide the best overall classification accuracy from the models tested. Section C.1 of appendix C shows the full results for all models tested.

→

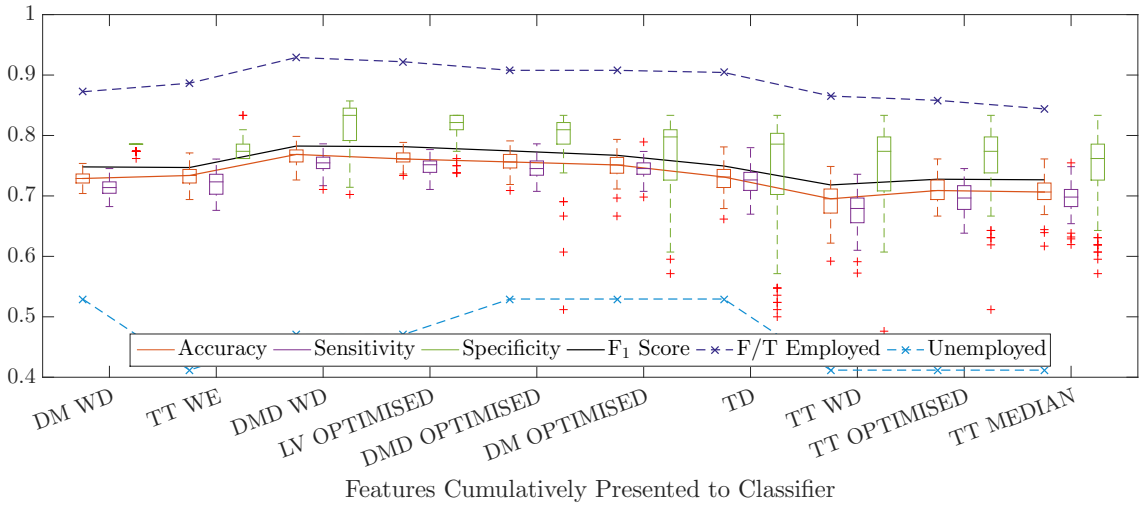


Figure 6.9: Identification of depression results using logistic regression classifier.

Table 6.1: Identification of depression results using logistic regression classifier.

Features	LOO CV Accuracy	10-fold CV Accuracy	5-fold CV Accuracy	3-fold CV Accuracy
1	0.729 ± 0.015	0.707 ± 0.016	0.699 ± 0.046	0.731 ± 0.022
2	0.734 ± 0.022	0.754 ± 0.035	$\mathbf{0.744} \pm 0.062$	0.789 ± 0.025
3	$\mathbf{0.769} \pm 0.020$	0.752 ± 0.022	0.724 ± 0.052	0.776 ± 0.020
4	0.761 ± 0.015	0.746 ± 0.022	0.719 ± 0.037	0.789 ± 0.022
5	0.756 ± 0.022	0.744 ± 0.025	0.729 ± 0.037	0.789 ± 0.022
6	0.751 ± 0.026	0.751 ± 0.060	0.714 ± 0.037	$\mathbf{0.801} \pm 0.022$
7	0.731 ± 0.030	$\mathbf{0.764} \pm 0.027$	0.714 ± 0.052	0.798 ± 0.022
8	0.695 ± 0.040	0.754 ± 0.030	0.716 ± 0.055	0.798 ± 0.022
9	0.709 ± 0.032	0.740 ± 0.027	0.729 ± 0.057	0.793 ± 0.021
10	0.707 ± 0.027	0.749 ± 0.042	0.741 ± 0.062	0.791 ± 0.025
11				0.789 ± 0.022
12				$\mathbf{0.801} \pm 0.022$

6.3.3 Differentiation of mental health conditions

The next classification task is to explore whether the features extracted from the geolocation data can be used to distinguish between the three cohorts of participants in the AMoSS study. The ability to distinguish BD from BPD has particular clinical relevance because it is known that these conditions can be challenging to differentiate due to overlap of diagnostic criteria (Baryshnikov et al., 2015) and similarity of symptoms, exacerbated by inaccurate retrospective recall (Saunders et al., 2015b). Figure 6.10 shows the results of a three-class classification using the quadratic discriminant analysis classifier with 20 iterations of group equalisation. The overall accuracy of the classification reaches (median $\pm$ iqr) 0.552 ± 0.003 after selecting three features for the model. The classification sensitivity for the three cohorts indicates that BPD and HC participants can be classified with much greater sensitivity than BD. Further insight can be gained from the confusion matrix in figure 6.11, which shows the sensitivity of HC and BPD classifications as 0.696 and 0.762 respectively, with the sensitivity of BD classifications as only 0.193. Section C.2 of appendix C shows the full results for all models tested.

Figure 6.12 shows the distributions of the features in figure 6.8 for the three cohorts. This indicates that there are clear differences between feature distributions of HC and BPD participants that enable classification between these two cohorts. The feature distributions for the BD participants is between the two, explaining the reduced classification sensitivity. This is also clear from the confusion matrix, where misclassified samples for BD participants are split between the other two cohorts, but with the majority misclassified as HC. Further insight can be obtained by splitting the data from BD participants by the QIDS threshold of 11 used for depressions classification earlier. This is shown in figure 6.13 on page 162 where it can be seen that, on the whole, the feature distributions for BD participants with QIDS score < 11 are very similar to HC participants. This is not surprising given the episodic nature of bipolar disorder. When individuals are not in an episode (i.e. they are euthymic) they should behave much like individuals without bipolar disorder (i.e.

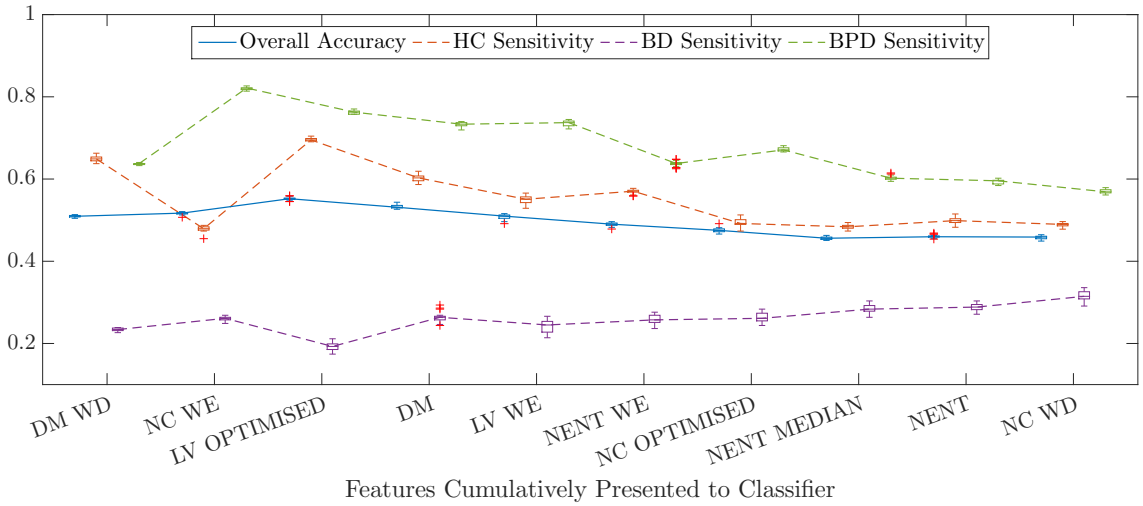


Figure 6.10: Differentiation of mental health conditions results using quadratic discriminant analysis classifier with leave-one-participant-out cross-validation.

	HC	BD	BPD
HC	**0.696**	0.134	0.170
BD	0.420	**0.193**	0.387
BPD	0.125	0.113	**0.762**

Figure 6.11: Confusion matrix of mental health condition classifications using first three features from figure 6.10. True values are shown on the rows and proportions of classifications on the columns. Bold values indicate proportion of 'correct' classifications.

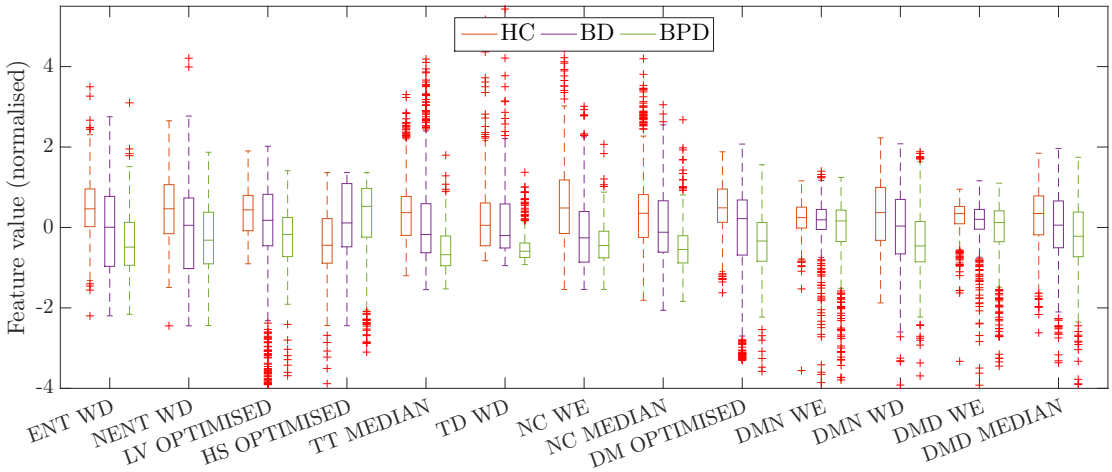


Figure 6.12: Feature distributions of filtered features calculated on HC, BD, and BPD participants.

HC participants in the study). This is exacerbated by the distributions of QIDS scores shown in figure 5.7 and table 5.3, which indicate that only 22 % of the labelled weeks of data from the BD participants are above the threshold for depression.

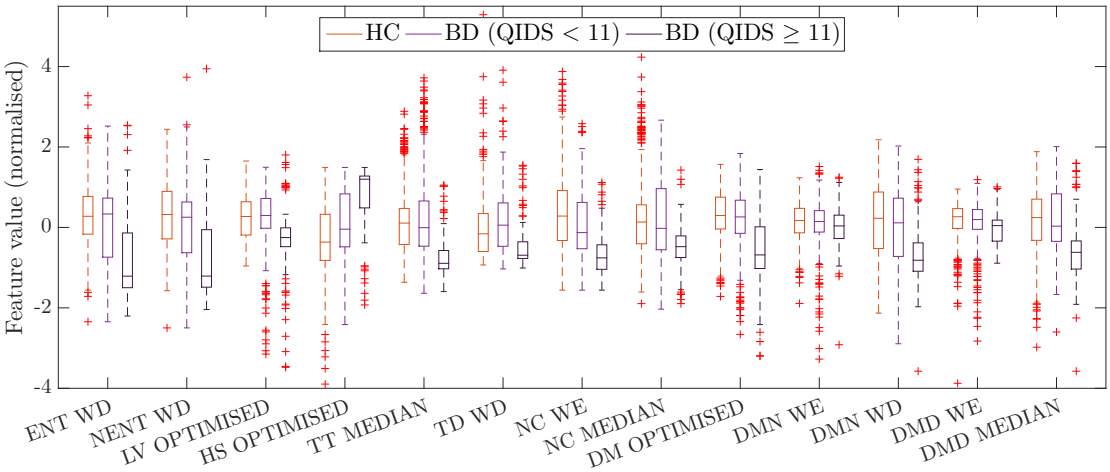


Figure 6.13: Feature distributions of filtered features calculated on HC and BD participants split by QIDS threshold of 11.

6.4 Regression tasks

The results in the previous section demonstrated that there are measurable differences in the geolocation-derived features between depressed and non-depressed individuals that can be used to train classification models that achieve reasonable accuracy in BD participants. However, in many cases simply being able to *detect* depression is not sufficient – psychiatric professionals need to be able to *quantify* the level of depression. This is a regression task that will be explored further in this section.

6.4.1 Methodology

6.4.1.1 Regression models

Standard linear regression models using equation (3.3) or higher order variants are useful in many situations. A first order linear regression model was used by Saeb et al. (2015a) for depression level estimation from geolocation data. However, simple linear regression models can be problematic because they are not bounded – QIDS scores are limited between 0 and 27, but linear regression can return values outside of this range. This can give misleading results when performing estimation and evaluating the model. Two models have been tested in this section:

a) the standard first-order linear model of equation (3.3); and
b) a quadratic linear model with non-linear extension of the form given in equation (3.20) with limits $a = 0$ and $b = 27$. This is equivalent to a quadratic generalised linear model (GLM) with a logistic link function if the QIDS scores are scaled to the range $[0, 1]$ by dividing by 27 and will be referred to as such.

6.4.1.2 Feature selection

Regression feature selection was performed using the regression models described above with leave-one-participant-out cross-validation. The same greedy forward feature selection method used for the classification tasks and described in section 3.6.2.1

and algorithm 3.3 was used to select features using the MAE of the QIDS score estimation.

Group equalisation is not applicable to regression tasks because there are no discrete groups to equalise.

6.4.2 Features

Box plots showing feature distributions for detection of depression were shown in figure 6.8 and discussed in section 6.1.5. Further insight can be gained by viewing the features in more detail.

Similar to how classification on individual features provided a useful indication of the predictive power of the different features and feature subsets in figures 6.1 and 6.6, training different regression models on individual features can provide a useful indication of how well they would perform in a regression task.

Figure 6.14 shows the results of training a standard linear regression model (indicated by a circle) and a quadratic GLM with logistic link function (indicated by a cross). A baseline model is also presented showing the result of using the mean QIDS score of the left out weeks as the prediction model. The regression models are generally similar to the baseline model with at least one of the models generally improving over the baseline. Full results are given in section C.3.1 of appendix C.

Figure 6.15 on page 166 shows scatter plots of six of the features for the BD participants calculated on the data subsets that provided the lowest error rates in figure 6.14. The same two regression models and the baseline model are shown overlaid, as well as the QIDS threshold of 11 used for depression classification. The same trends that were demonstrated in the feature distributions shown earlier can be seen in figure 6.15. Weeks where participants report higher levels of depression tend to have lower entropy values, higher proportion of time spent at home; and hence less time spent travelling and fewer locations visited. This matches the results in previous work by Saeb et al. (2015a).

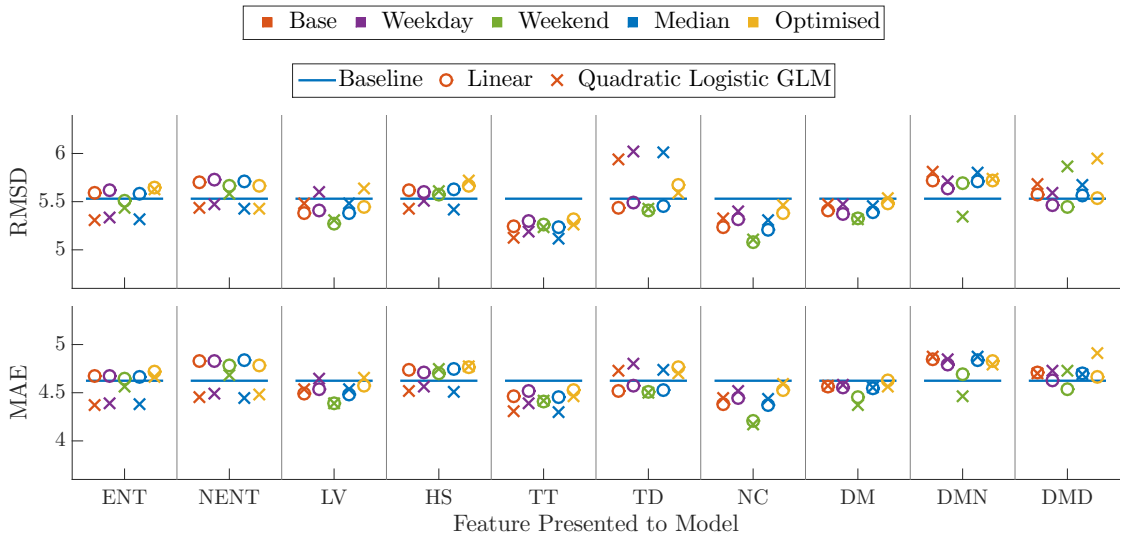


Figure 6.14: Feature performance for quantification of depression levels using individual geolocation features calculated on different data subsets for BD participants.

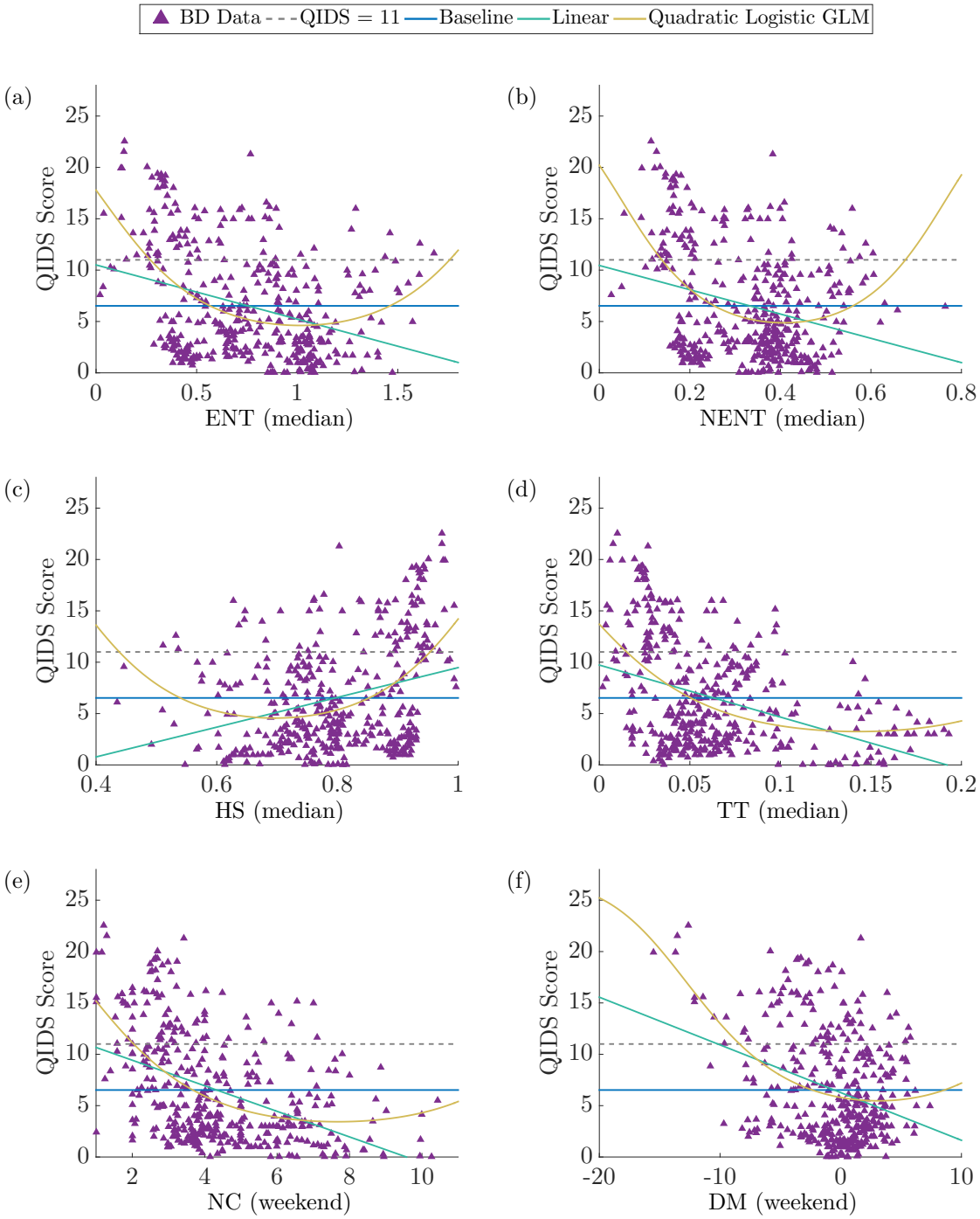


Figure 6.15: Selected geolocation features for BD participants with regression models overlaid.

6.4.3 Quantification of depression levels

Finally, figure 6.16 below shows the regression results from the two models trained on cumulative features selected using the greedy forward feature selection method. Full results including the features selected for each model are given in section C.3.2 of appendix C. Regression on multiple features improves results over individual features to a mean error rate of 3.929 with the linear model and 3.593 with the quadratic logistic GLM model using 9 and 16 features respectively.

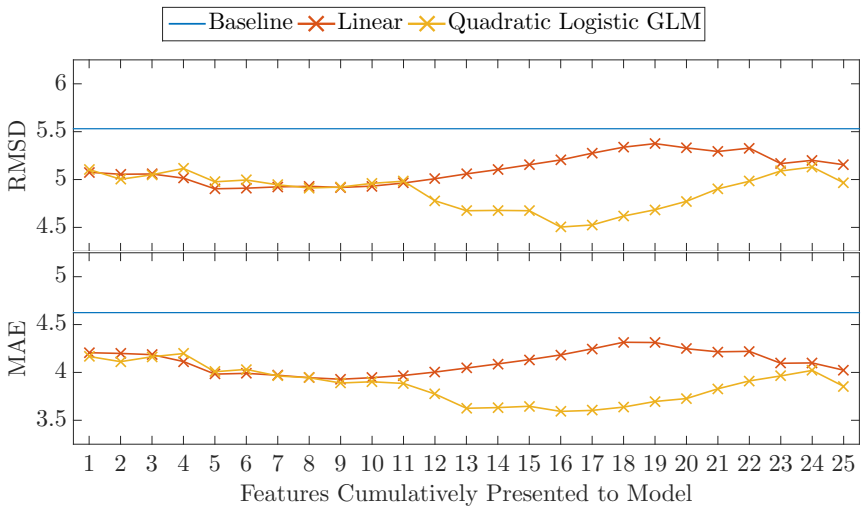


Figure 6.16: Quantification of depression level results for linear models trained on cumulative geolocation features presented to regression models for BD participants.

6.5 Discussion

This chapter has demonstrated the utility of geolocation-derived features as predictors of mental state. In particular, when combined in a logistic regression model, pooled accuracy of over 75 % was achieved in predicting depression from geolocation alone.

This chapter also demonstrated an online method for filtering of the extracted features using a Kalman filter trained to approximate window filtering. This was shown to improve the discriminative performance of the extracted features, demonstrating the importance of considering noise in the feature space, not just in the data space. It is believed to be the first time that this method has been applied to geolocation-derived features for mental state estimation. The method is likely to be applicable to other applications with noisy features.

In practice, direct application of classification scores may be useful to provide clinicians and community healthcare workers with an indication of which patients may need extra attention. However, a classification score alone may not be sufficient to fully understand the circumstances. This is why interpretable features such as the ones detailed in the previous chapter are useful. Healthcare workers can view the trends in the features to uncover what may have changed for their patient. Classification itself is only part of the process, but is useful nonetheless to identify patients for extra investigation.

As described in section 6.3.1.3, three classical models were tested for this application: logistic regression; naïve Bayes; and quadratic discriminant analysis. There are many more classification models available (Murphy, 2012; Hastie et al., 2015), both frequentist and Bayesian. Although only the results from one of the classification models tested are shown for each task in figures 6.9 and 6.10, the full results for all three models are given in sections C.1 and C.2 of appendix C, which show that the performance of all three models is similar. More complex models may provide marginally improved performance, but the feature distributions indicate that the performance limit is most likely to be the features characterising the patient, and not the classification model applied. This is supported by inspection of the raw feature values

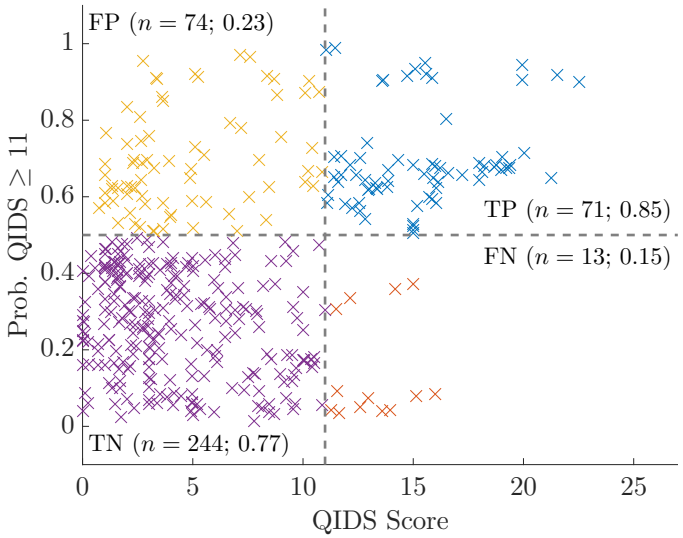


Figure 6.17: Probability of depression using logistic regression classifier with three geolocation-derived features found to provide optimal classification of depression in table 6.1. The vertical line shows the QIDS = 11 threshold for depression; and the horizontal line shows the standard $p = 0.5$ decision threshold applied to the logistic regression classifier to arrive at the results in table 6.1. Results were generated using leave-one-participant-out cross-validation over 100 iterations of group equalisation for each left out participant. The value for the probability shown is the mean of all iterations for each week of data tested.

shown in appendix D. The level of noise in the feature values makes this a challenging classification task. This also supports the level of data cleaning and preprocessing applied to the raw geolocation data in the previous chapter. It is crucial to have features that characterise the patient as well as possible to be able to make accurate inference about them. Large levels of missing data naturally mean that the features calculated from the raw values will not be an accurate representation of behaviour.

The logistic regression classifier used to generate the results in table 6.1 has two additional properties that may be advantageous in practice. Firstly, it can provide probability values to support the classification. These values can be considered as a risk score of being depressed. The mappings between QIDS scores and probability values for all weeks analysed are shown in figure 6.17 above. It can be seen that the probability values do not map directly onto the raw QIDS scores, but this is to be expected because the model is optimised for classification, not regression (in training, the model is only presented with binary labels, not the actual QIDS scores). However,

figure 6.17 also indicates that sensitivity for depression is high (0.85) meaning that the score is relevant for identifying patients, even if some non-depressed individuals are also identified. The strength of the risk score, combined with the classification itself, can provide additional decision support. Secondly, due to the linear model underlying logistic regression, an indication of which feature(s) contribute most strongly to the classification can be provided. This can be useful to give clinicians an indication of why a given prediction is made, again showing the importance of using interpretable features in the model. These properties combined provide clinicians with an additional source of behavioural data that they can use, together with their clinical judgement, to triage patients more efficiently.

The regression results given in figure 6.16 also deserve further investigation. Figure 6.18 shows the distribution of actual QIDS scores for ranges of scores estimated from the geolocation-derived features using the quadratic logistic GLM described in section 6.4.1.1 using the first 16 features shown to provide the optimal results in figure 6.16 and table C.8 of appendix C. This shows that generally the estimated QIDS scores provide a reasonable indication of the actual QIDS score. This can therefore form an additional metric that clinicians can use when assessing their patients. Each row is represented by the Gaussian distribution of the actual QIDS scores for each range of estimated scores, and the graph on the right of figure 6.18 shows the standard deviation of this distribution. The standard deviation is relatively low at lower and higher estimated QIDS scores, increasing to a maximum at estimated QIDS scores around 9. This indicates that estimates are more accurate at lower and higher levels of depression, with an area of lower confidence around the diagnostic threshold of QIDS = 11. Given the known noise in the QIDS score labels, discussed in section 5.7.1, this in not surprising.

Similar to the discussion on the choice of classification model earlier, several other regression models, with varying levels of non-linearity, are also available. However, in this application, linear regression-based models are a reasonable choice. The feature plots in figure 6.15, shown in greater detail in figure 7.1 in the next chapter with individuals highlighted, demonstrate the level of noise in the signal. Non-linearity

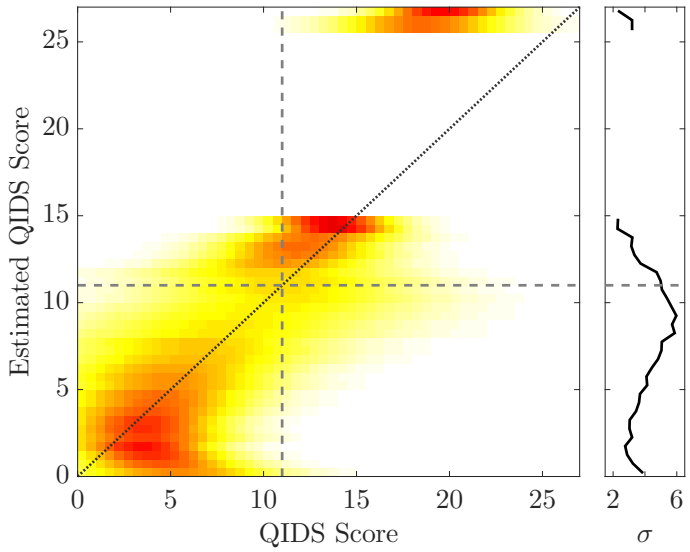


Figure 6.18: Distribution of of actual QIDS score for estimated scores calculated using the quadratic logistic GLM described in section 6.4.1.1 with the first 16 features shown to provide the optimal results in figure 6.16 and table C.8 of appendix C. The distribution in each row is calculated as a Gaussian distribution of the actual scores for the estimated scores in the given range. The graph on the right shows the standard deviation of the fitted Gaussian distribution for each range of estimated scores.

does not appear to be a major concern, rather inter-individual variability in model, which will be explored further in the next chapter. Having said that, the quadratic logistic GLM imposes a level of non-linearity to the model, which is demonstrated in figure 6.16 and section C.3.2 of appendix C to improve results with over 11 features included in the model. The application of the logistic GLM is also reasonable given the known bounds of the QIDS score.

In the case of both detection and quantification of depression, the use of the QIDS score as a subjective proxy for mental state has limitations, which have been discussed in detail in sections 2.2.1 and 4.4.3 and are accepted as limitations of this work.

The results in this chapter (both classification and regression) have been presented as independent classifications of weeks of geolocation data. As shown in the data examples in figure 7.2 in the next chapter, levels of depression follow a longitudinal course and are not independent week-to-week, indicating that there may be time-series dynamics in the data. Future work should explore incorporating these dynamics into the predictive models. Further details are given in section 8.3.2.

Incorporation of further modalities into the classification (or regression) models may also increase the performance presented using geolocation alone. However, the biggest limitation is likely to be inter-individual variability, where different individuals naturally have different lifestyles and routines that may not fit into the population-level models of the type presented here. In the next chapter, one potential way to handle inter-individual variability in longitudinal data is developed and explored in detail.

Chapter 7

Personalised Models of Mental Health

The previous chapter explored fitting global models of depression to features calculated from geolocation data recorded from participants with bipolar disorder. It was demonstrated that the feature data from all BD participants combined generally matched the trends that would be expected from previous work by Saeb et al. (2015a, 2015b, 2016) and the diagnostic characteristics of bipolar disorder. However, it was also noted that there was a large amount of noise in the data, making regression performance poor. Data from HC and BPD participants can also be included, which increases the noise further still. This chapter investigates the extent to which personalised models can improve over the results from the global models.

Further insight into the profile of the noise in the global feature can be seen by exploring the distribution of points from individual participants within the global distributions. This can be seen in figure 7.1 on the next page where several different participants have been highlighted within the global distributions for three of the geolocation features previously shown in figure 6.15. The vertical dashed line indicates the global logistic regression threshold around the QIDS = 11 threshold.

The first row, (a), shows two participants (in blue and orange) that appear to fit well with the global classification model of depression (based on a threshold QIDS score of 11). The participant in blue is mainly depressed and fits the global model of spending a high proportion of time at home; while the participant in orange is not

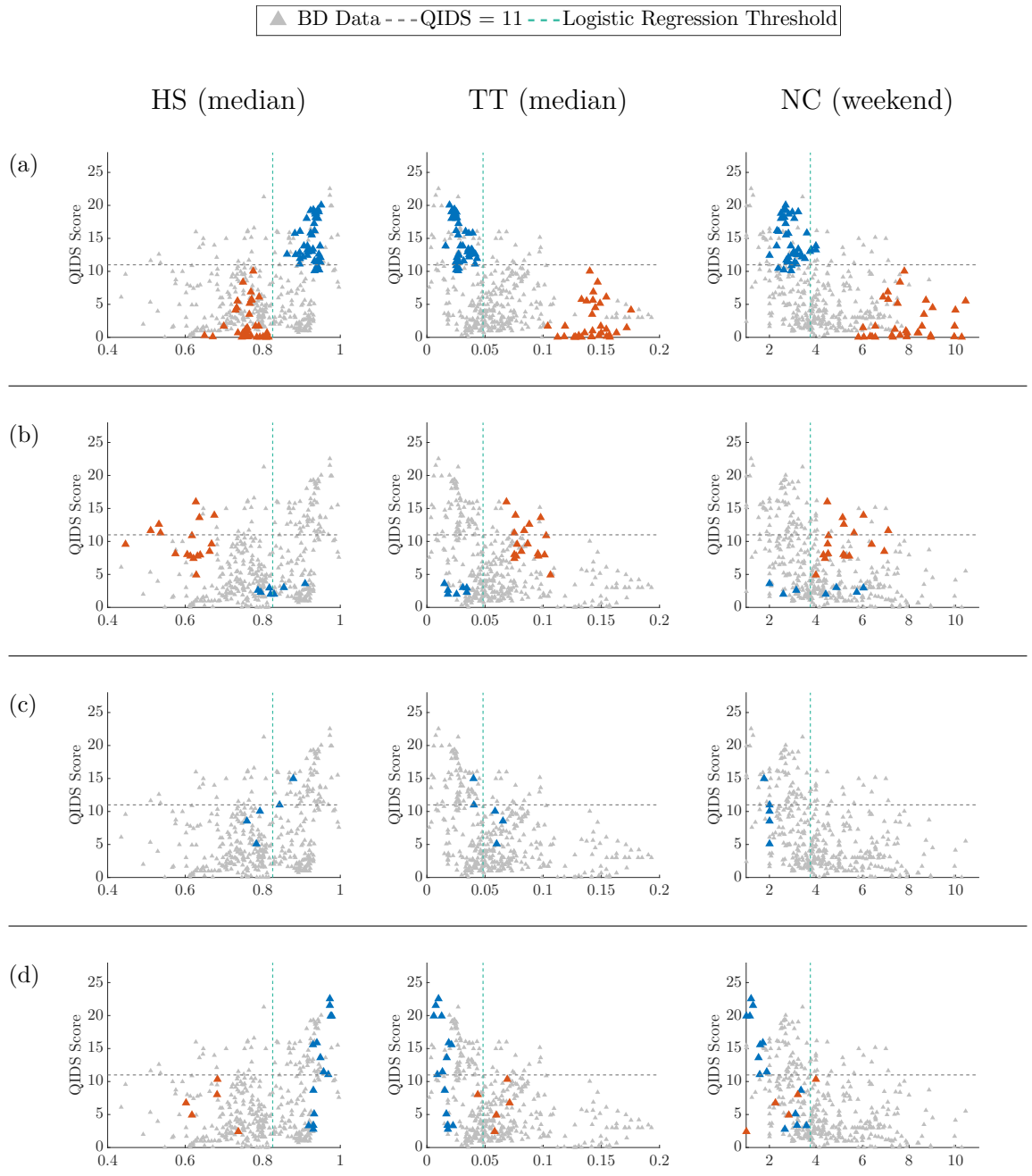


Figure 7.1: Selected geolocation features for BD participants with different individual participants highlighted in each row. The two individuals highlighted in row (a) appear to fit well with the global classification model of depression, while the two individuals in row (b) do not. The individual in row (c) has a useful trend in only two of the three features, while the blue individual in row (d) follows the expected trend, but appears to be much steeper than the global model. This explains why global models, trained over all available individuals, were shown in the previous chapter to perform poorly for left out individuals. This chapter explores how performance can be improved by training personalised models on individuals that appear similar.

Table 7.1: Spearman's rank correlation coefficients for highlighted participants in figure 7.1.

Row	Participant	n	Correlation coefficient		
			HS (median)	TT (median)	NC (weekend)
(a)	blue	43	0.0609	-0.4805^{**}	-0.2320
	orange	33	-0.2060	0.2700	0.1386
(b)	blue	7	0.4286	-0.1429	-0.0714
	orange	17	-0.0319	-0.2451	0.3333
(c)	blue	5	0.9000	-0.9000	-0.7071
(d)	blue	14	0.7407^{**}	-0.5868^{*}	-0.8549^{***}
	orange	5	-0.1000	0.2000	0.9000

Significance of correlation coefficients indicated by asterisks: $^{*} < 0.05$; $^{**} < 0.01$; $^{***} < 0.001$. Note that correlations are highly dependent on the number of data points available.

depressed and spends a high proportion of time travelling. These two participants would be easily classified as depressed or not using the global classification model indicated by the vertical dashed line.

The second row, (b), shows two participants that do not appear to fit the global classification model well. The participant in blue is not depressed, but spends very little time travelling (although the number of clusters visited at the weekend is more as expected from the global model). The participant in orange is consistently around the threshold for depression but spends a high proportion of time travelling and visits many clusters at the weekends. The blue participant would be incorrectly classified using the TT feature, and the orange participant would be incorrectly classified as not depressed when depressed.

Closer inspection of these four participants from a regression perspective reveals a more complex picture. None of the four participants exhibit strong correlations with the QIDS score. Spearman's rank correlation coefficients for the highlighted participants are given in table 7.1 above. The blue participant in row (a) only has a significant correlation between the TT feature and the QIDS score. A visual rela-

tionship can be seen with the HS feature, where the participant only exhibits higher HS values with lower QIDS scores, but it is not significant.

It is also clear from the four participants in the first two rows that pairs of QIDS scores and feature values for participants tend to cluster together. This is unsurprising given the different lifestyles of individuals. Where the QIDS scores for a participant cross the depression threshold of QIDS = 11 (only applicable to the blue participant in row (a) for three points, and the orange participant in row (b)), the feature values do not show a noticeable change making either classification or regression difficult.

The third row in figure 7.1, (c), shows a participant with a stronger correlation between the feature values and QIDS scores, as demonstrated by the correlation coefficients in table 7.1. This participant appears to be most well characterised by the HS and TT features. However, there are too few data points for the correlation coefficients to be significant. Classification between points above and below the QIDS = 11 threshold could be performed perfectly using the global classification model shown for the HS and TT features.

The final row, (d), shows two further individuals that exhibit stronger correlations between feature values and QIDS scores. The blue participant has a strong correlation with all three features but is most strongly characterised by the NC feature. It is also notable that the slope on the regression appears to be steeper than other participants, and steeper than the global trend. Using the global classification model shown, this participant would always be classified as depressed even when not depressed. The orange participant in row (d) also appears to be characterised by the NC feature, but inversely to the other participants and the global trend. There are also too few data points for this individual for the correlation coefficients to be significant.

Appendix D shows the same individuals highlighted on the raw, unfiltered feature data. The correlations between QIDS scores and features for the individuals shown do not appear to be reduced by the filtering and smoothing process.

It is clear from the participants shown in figure 7.1 that different individuals are characterised by different models. This indicates that global models will not apply

equally to all individuals and more personalised techniques are required. Mohr et al. (2017) call this the "curse of variability", from the well-known curse of dimensionality.

7.1 Overview of personalised models

To further explore the potential improvements that can be obtained using different types of personalised models, figure 7.2 across the next two pages shows the performance of regression models trained on different subsets of the complete set of individuals, and tested on the five selected individuals shown in rows (a) through (e). Each graph shows the recorded QIDS scores for the individual (where available), and the estimated QIDS scores from the regression model. The regression models used are all Bayesian Lasso models over multiple individuals, described fully in section 7.7.2, and the parameter values used are given in section 7.7.2.4. All the models were trained and tested on the 10 main features derived from the geolocation data (detailed in section 5.10) calculated on whole calendar weeks of geolocation data. For this section, data from all participants in the AMoSS study (HC; BD; and BPD) who provided more than 5 calendar weeks of labelled data were included. Where there are estimated QIDS scores shown in figure 7.2 but no actual QIDS scores, this means that the individual did not complete the questionnaire but did provide data for analysis – see section 5.7 for details.

Figure 7.2, across pages 178 and 179, shows four columns, where each column contains a different class of model.

The first column of figure 7.2 shows 'global' models trained using the available labelled data from all the individuals other than the one being tested, but including the first half of the data for the test participant, up to a maximum of 8 weeks. While this 'global' model is a common approach, performance is generally poor. The estimations given for all five selected individuals show that the output has fairly low variance. This indicates that the model is fitting mainly to the mean value of the features with an offset at the mean of the training data. Given the clustering of points for the individuals shown in figure 7.1, it is not therefore surprising that

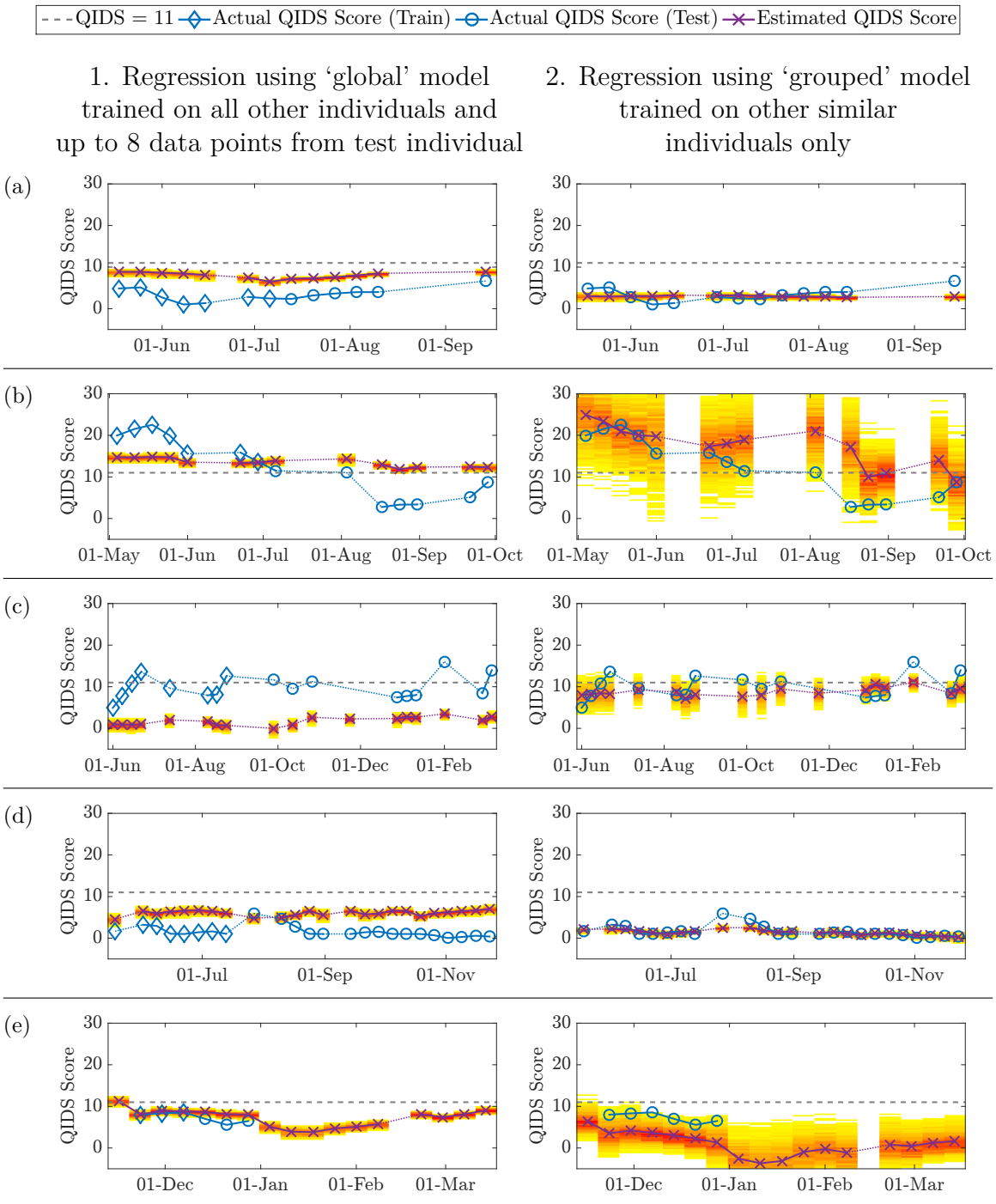


Figure 7.2: Estimation of QIDS scores using the 10 features detailed in section 5.10 calculated on whole calendar weeks of data for five selected individuals shown in rows (a) through (e). Bayesian Lasso regression models were trained on different subsets of individuals, shown in the four columns on this page and next. On this page, column 1 show estimation using models trained on all individuals except the test individual. Column 2 shows estimation using models trained on a subset of other individuals that appear 'similar' to the test individual. (continued on the next page)

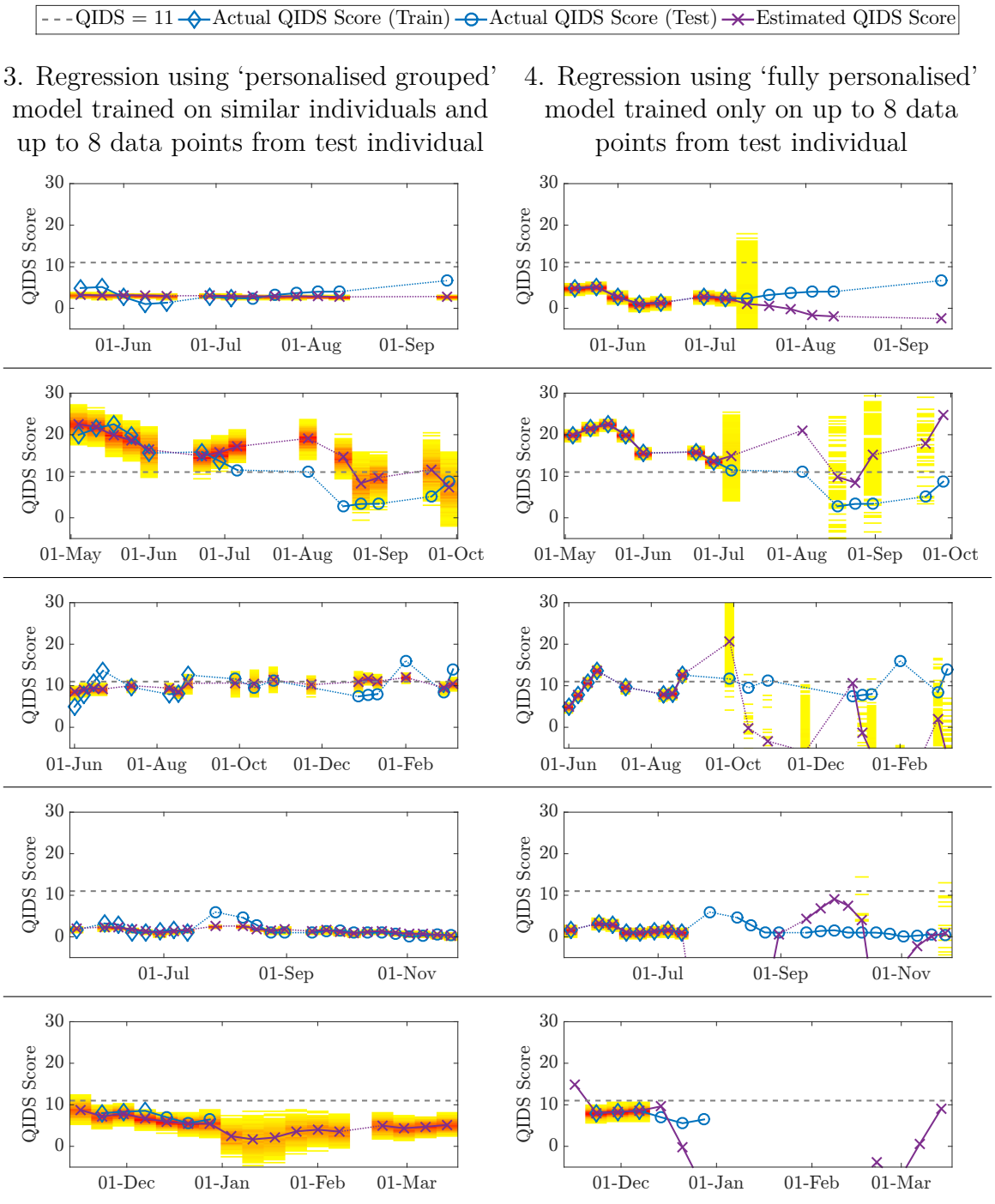


Figure 7.2 (cont.): Column 3 on this page shows estimation using regression models trained on the same subsets of 'similar' individuals used in column 2, with the addition of the first half of the labelled data available for the test individual, up to a maximum of 8 weeks weeks. The procedure used to determine the 'similar' individuals used to train the models in columns 2 and 3 is the main subject of this chapter. Finally, column 4 shows 'fully personalised' models trained only on the first half of the data available for the test individual, up to a maximum of 8 weeks.

some individuals exhibit estimated values consistently higher or lower than the actual values and that changes in the levels of depression are not accurately detected.

At the other extreme, the fourth column of figure 7.2 on page 179 shows 'fully personalised' models trained only on the first half of the data for the test participant, up to a maximum of 8 weeks. These models generally fit very well to the data used for training, but do not generalise well to the unseen test data. The variance on the estimations for the unseen data is also generally large. This indicates that the models are over-fitting the data used for training, but it also suggests that there may be potential relationships between the geolocation-derived features and the self-reported depression in the individuals tested, since the model fits well to the training data. While fully personalised models may not therefore be a practical approach due to the unpredictability of estimation on unseen data, it does demonstrate the potential for geolocation features to estimate depression levels.

For all five participants, the global model is outperformed on the training data by the fully personalised model where recorded QIDS scores are available. The participant in row (a) shows a close match for the fully personalised model although the range of QIDS scores is low. Row (b) shows the same participant as the blue highlighted participant in row (d) of figure 7.1, with significant correlations shown in table 7.1. The range of estimated QIDS scores are much more closely matched in the fully personalised model than in the global model, and even the estimations on the unseen test data show a similar trend to the actual data. Rows (c) and (d) show how the global model has a tendency to under- and over-estimate the QIDS scores for different individuals. This is not surprising given the clusterings of points seen in figure 7.1.

7.2 Related work

Personalised models have been explored in many ways in several different areas. This section reviews some of the related work in this broad field. Work directly related to grouping of individuals is reviewed in section 7.5.1.

The term 'personalised' is used in several different ways depending on the application. In healthcare, the concept of 'personalised medicine' commonly refers to the tailored selection of treatments or therapies for patients based on their biological and physiological profiles (Schleidgen et al., 2013). The means to this is commonly though genetic or genomic analysis, which was touched on briefly in section 2.2.2 in the context of mental health. The cost of performing gene sequencing has reduced dramatically (Zignol et al., 2018), which has facilitated the development of large genetic databases available for analysis by researchers (van Rooij et al., 2017). The general objective when analysing these data is to uncover subgroups (genetic variants) in high-dimensional data that predict the dependent variable of interest (disease risk factor; outcome; response to treatment; etc.). The extreme dimensionality of the data; the relatively low number of specimens; and the high level of natural inter-individual variability make this analysis particularly challenging (Dasgupta et al., 2011). This is essentially a feature selection task (Bolón-Canedo et al., 2014) as described in section 3.6. Early methods used included the Lasso extension to linear regression described in section 3.6.1, and the group Lasso (Bakin, 1999; Yuan & Lin, 2006), which treats groups of predictors as a single feature for selection purposes. Many more recent methods have been proposed (Bolón-Canedo et al., 2014) but these are fundamentally aimed at finding a subset of features across the whole dataset that explain variability, rather than splitting individuals into groups based on similarity. However, as noted by van Rooij et al. (2017), there are indications of variability between datasets from different populations, which suggests that grouping individuals here may reveal important (and clinically relevant) differences in genetic variants. While grouping by demographics is the most obvious approach, unsupervised grouping may reveal additional insight.

Another family of methods commonly applied in healthcare to form personalised physiological models are Gaussian processes (Rasmussen & Williams, 2006). Gaussian processes model output functions as a Gaussian distribution parametrised by mean and covariance functions of the feature values. By incorporating past data from an individual into the mean and covariance functions, personalised predictions of

expected future values can be made. Gaussian processes are particularly suited to time-series data, and in healthcare have been demonstrated to be useful for generating early-warning scores of patient degradation (Clifton et al., 2012; Alaa et al., 2018); detecting abnormal trajectories (Pimentel et al., 2013; Alaa et al., 2018); imputation and noise reduction in data recorded from wearable devices (Clifton et al., 2013); and post-surgical outcome prediction (Lee et al., 2016), among others. While the standard Gaussian process can fit a functional output specific to the patient, this function is itself parametrised from training on a larger population data and therefore assumes homogeneity of function parameters across the population. Alaa et al. (2018) extend this by incorporating a 'patient subtype' based on admission features, where each subtype has different model parameters. Similarly, Hensman et al. (2015) define a hierarchical Gaussian process with a Dirichlet process prior on the model parameters, allowing grouping of individuals into different subsets. As more data are available for analysis, these techniques may provide a complementary model of the interaction between features and mental state.

Outside of healthcare, personalisation is important for delivery of personalised services. The focus here is different, however, with the aim generally being to predict what action a user might take; what product a user might buy; or what film a user might like to watch; etc. based on their past actions (Bobadilla et al., 2013). These services are broadly known as 'recommender systems' and are enabled by a variety of machine learning models. Recommender systems fall generally into two categories: 'content-based filtering', which uses similarity in content between items to make suggestions; and 'collaborative filtering', which uses similarity in past interactions between users to make suggestions; or combinations of the two methods. Content-based methods broadly use the distance between attributes or properties of items to identify the one(s) with high levels of similarity (Pazzani & Billsus, 2007). The methods applied are often inspired by the closely related field of text search and include simple nearest neighbour methods; decision trees; linear classifiers; and naïve Bayes classifiers (Pazzani & Billsus, 2007). Collaborative filtering has additional challenges, often involving very large and very sparse matrices of users and items, where

most users have interactions with only a small subset of items, with the rest missing. Techniques used in collaborative filtering start with simple nearest neighbour methods based on the distance between non-missing attributes for pairs of users (Bobadilla et al., 2013). More advanced methods include non-negative matrix factorisation (Koren et al., 2009), which attempts to find lower-dimensional latent factors of the original sparse matrix that can recover the available entries in the matrix, and will also fill in the missing values to use for recommendation. More recently, deep learning-based approaches are also gaining interest (Singhal et al., 2017). While there are similarities between recommender systems and the present application, there are also important differences that mean that these methods are challenging to apply. Content-based methods find similarities in the feature space, and the aim in the present case is to find similarities in models linking features and outputs. Additionally, as shown in section 7.4, clustering in the model-parameter space is also challenging in this application. Collaborative filtering methods such as non-negative matrix factorisation are designed primarily for dimensionality reduction and filling missing feature values rather than for predicting output values.

A final area worth discussing is that of functional data analysis (Ramsay & Silverman, 2005; Ferraty & Vieu, 2006), which generalises the concept of considering discrete measurements of data as (noisy) observations of an underlying functional process. This method of data analysis is particularly (although not exclusively) suitable for application to time-series data where individual samples are not considered to be independent. Functional data analysis usually proceeds by extracting an approximation of the underlying process through fitting a smoothing function to the measured data, often a smoothing spline or Fourier transform (depending on application-specific assumptions of the data). After extracting the approximated functional form of the underlying process, this is used for analysis instead of the sampled data. Many standard statistical methods such as principal component analysis; correlation analysis; regression; and classification can be applied to the function domain rather than the data domain. A particularly useful feature of working in the function space is that derivatives can easily be computed, which may provide insight into subtle features

of the data that may otherwise be hidden. The raw geolocation data; the extracted features described in section 5.10; and the mental state data presented in figure 7.2 are all suitable for interpretation as functional data. While functional data analysis provides many useful statistical tools, most are based on population-level models. Dirichlet process extensions have been proposed, but these are focussed on clustering characteristically similar curves (Qin & Zhu, 2013) or piecewise clustering of curve segments (Nguyen & Gelfand, 2011; Boullé et al., 2014), not clustering by response of an output function to feature functions.

7.3 Grouped personalised models

The participants in rows (a), (d), and (e) of figure 7.2 highlight an additional problem with fully personalised models – because the participants exhibited relatively stable QIDS scores in the training data, it is difficult to know how they would behave when more or less depressed. Although it would be tempting to suggest asking individuals to complete a few weeks of QIDS questionnaires to initialise a model, this will not work well in practice because the behaviour of the individual under their whole range of potential mental health states will not be available for training. Further complicating the situation, the symptom profiles of different episodes, even for the same individual, may be different (Perlis et al., 2009), meaning that even a model trained on an episode for an individual may not pick up the next one.

One way to develop personalised models is therefore to apply appropriate priors to the coefficients of standard models to reduce the variance in fully personalised models. It is desirable to cluster similar individuals in the training data to build a library of potential 'prior models' that could apply to new individuals. New individuals would need to record a number of weeks of behavioural data whilst also completing the relevant questionnaires as calibration data, which would be used to assign them to a suitable prior model. This is similar to how a fully personalised model would be applied in practice, but would provide allocation to a known set of models, which would

reduce over-fitting to the limited calibration data available. The model identified can then be further tuned to the individual as more training data become available.

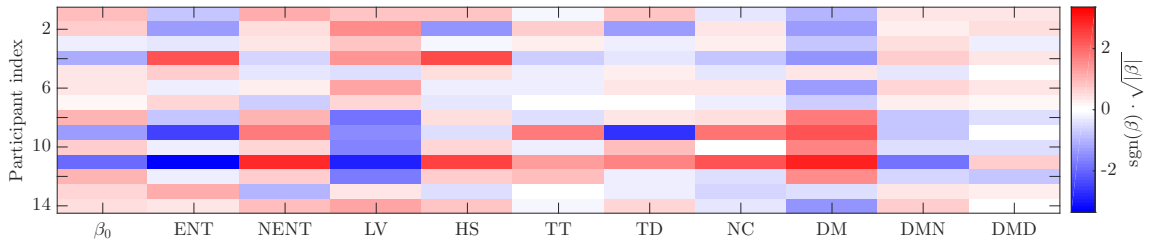
In the context of the use of technology in BD patients, Matthews et al. (2017) refer to a concept of *technology-mediated phenotyping* of different manifestations of BD. The diagnostic criteria of a depressive episode given in section 2.1.1.1 give wide scope for different individuals to have different combinations of symptoms, justifying the grouping of individuals into similar phenotypes.

Columns 2 and 3 in figure 7.2 show models trained on a subset of individuals in the whole data set that appear most 'similar' in characteristics to the test individual. In column 2, all the data from the test individual were left out and the model was trained on all the labelled data from the remaining individuals. In column 3, the first half of the labelled data for the test individual, up to a maximum of 8 weeks, were also included when training the model. It can be seen that the results in both columns 2 and 3 improve over the global and fully personalised models on the unseen test data. Including a limited sample of data from the test individual in column 3 reduces the variance of the sampled estimated values.

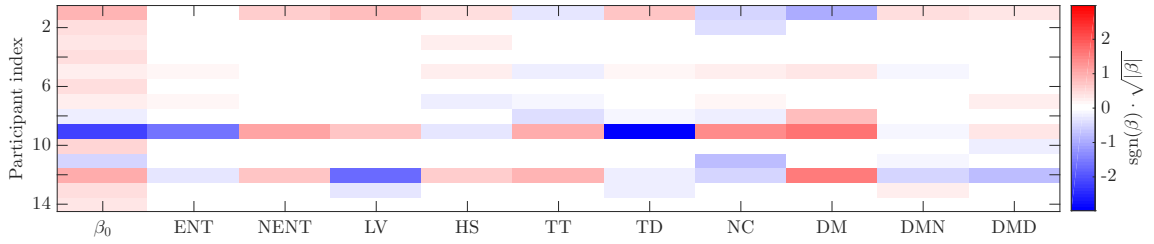
This leaves the question of how to determine the individuals that are 'similar' to the test individual. The remainder of this chapter describes a new model developed to dynamically group similar individuals.

7.4 Clustering individuals by model coefficients

One way to find individuals that have similar feature responses is to attempt to group participants based on the model coefficients in their fully personalised models. This is demonstrated in figure 7.3 on the following page where the coefficients of a fully personalised standard non-Bayesian regression model are shown for 14 BD participants who had 10 or more weeks of data available for training. Figure 7.3(a) shows a standard linear model of the form given in equation (3.3) while figure 7.3(b) shows a linear model with coefficients regularised using the Lasso (described in section 3.6.1). Both models were trained on all the available labelled data for each individual. In



(a) Standard linear regression model.



(b) Standard linear regression model regularised with the Lasso.

Figure 7.3: Examples of rescaled regression coefficients for fully personalised models trained on BD participants with 10 or more weeks of data. Models were trained only on the features calculated on full calendar weeks of data. For visualisation purposes, all coefficient values are scaled by square rooting the absolute value and multiplying the original sign. There are no clear patterns in the data, demonstrating the difficulty of finding individuals with "similar" relationships between predictors and output using only coefficient values from fully personalised models.

the regularised model, the regularisation parameter was chosen for each individual using 10-fold cross-validation on the available data. All models were trained on the 10 features calculated on the whole week of data, normalised to zero-mean, unit-variance over all participants. The columns show the coefficient values, and the first column shows the β_0 offset.

The range of coefficient combinations in the full linear model in figure 7.3(a) shows the difficulty of clustering individuals using fully personalised model coefficients directly. All of the participants appear to have very different regression coefficients. Participants 9 and 11 appear to be most similar but are inverted on the TD feature. This is not surprising given the over-fitting discussed above. Since these models are tuned so closely to the training data for the individual they do not generalise well for that individual, let alone another individual.

The Lasso is commonly used to regularise coefficients in linear regression (see section 3.6.1 for details). This has the effect of finding a subset of features that

best characterise the individual. This reduces performance on the training data, but often increases performance on unseen test data. Focussing on this subset may reveal subsets that are shared between individuals. The models in figure 7.3(b) were optimised using 10-fold cross-validation to select the number of features to retain. For the participants shown in figure 7.3(b), the regularised model does not reveal any more insight into participants that may group together. It is also clear that the number of features retained in the model varies widely between individuals, with all features retained for participant 9, and no features retained for participant 14 (i.e. the 'best' model is the constant baseline value).

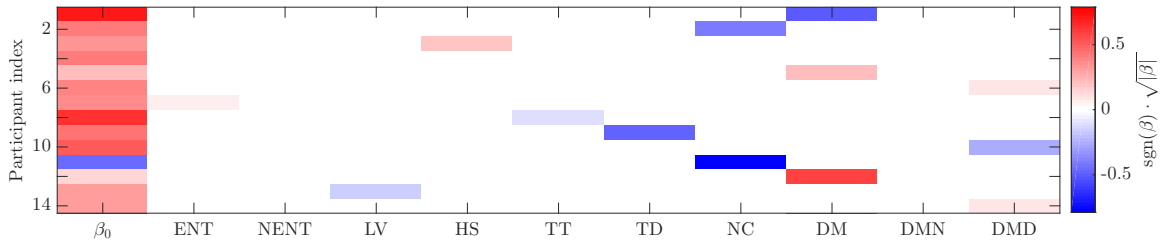
To further restrict the trained models, figure 7.4 on the next page shows coefficients for linear models regularised with the Lasso, but rather than choosing the best overall set of features through cross-validation, the number of features to allow in the model was restricted to a maximum of one in figure 7.4(a) to three in figure 7.4(c). As in the previous examples, even this increased restriction on model complexity does not reveal obvious similarities between individuals.

7.5 Dynamic grouping of individuals

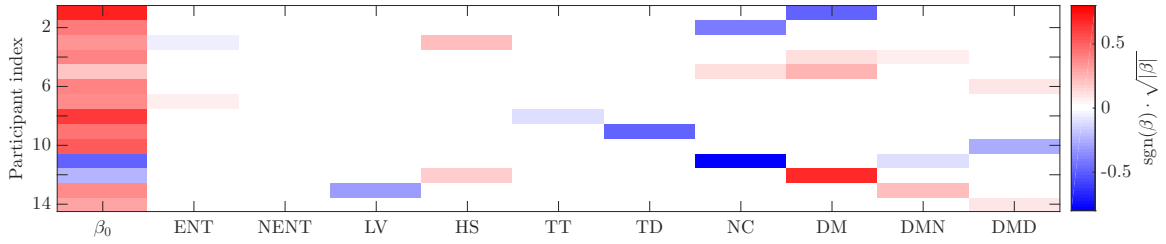
The previous section demonstrated the difficulties with using raw model coefficients from fully personalised models to group individuals. This is likely to be caused by a number of compounding factors.

Because the fully personalised models are optimised on each individual separately, the final model for each individual is (usually) the globally optimal model conditioned on the training data. This means that there may exist other (locally optimal) models that also perform well, but may have very different coefficients.

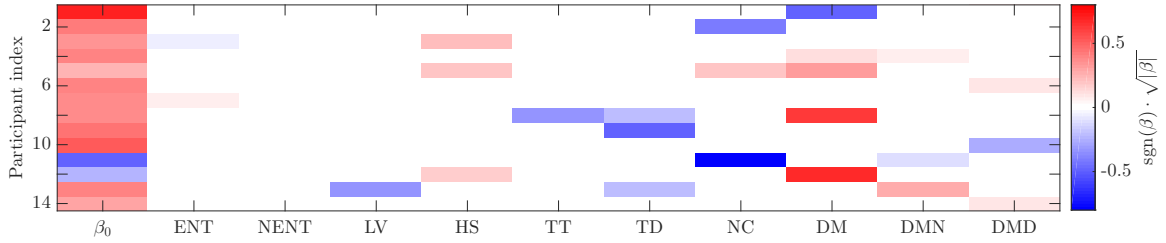
One of the reasons that this is particularly likely in this application is because of the highly correlated nature of some of the features. This can be seen from the global correlation coefficients shown in section B.1 of appendix B. Even within the main 10 features considered in the models above, there are several features that are



(a) Lasso regularised model allowing only one feature.



(b) Lasso regularised model allowing maximum of two features.



(c) Lasso regularised model allowing maximum of three features.

Figure 7.4: Examples of rescaled regression coefficients for Lasso regularised fully personalised models allowing a maximum of (a) one; (b) two; and (c) three features to be included in the model. Models were trained on BD participants with 10 or more weeks of data using only the 10 features from the base data subset. Similar to figure 7.3, no clear patterns are visible with regularised coefficients, again demonstrating the difficulty of finding individuals with "similar" relationships between predictors and output using only coefficient values from fully personalised models.

strongly correlated. Section B.1 also only shows global correlations, and participants may themselves have even stronger correlations between certain features.

It may therefore be the case that a globally optimal model for one individual is locally optimal for another, or more likely that a third model is locally optimal for both individuals. However, this grouping will not be revealed by viewing only the optimal models for the individuals separately.

A method is therefore required to dynamically cluster individuals into groups that have the same locally optimal model. Figure 7.5 provides a generic outline

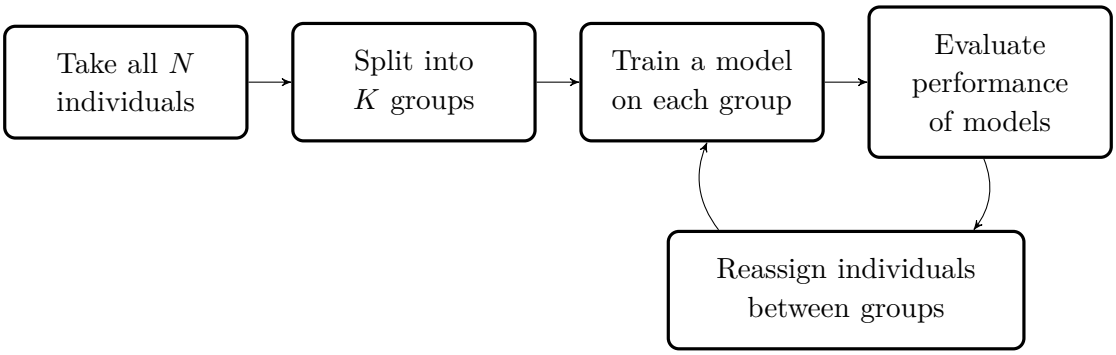


Figure 7.5: Generic method for dynamic grouping of individuals.

of the basic method required. The general principle is to assign all individual into one of K separate groups. A model is then trained per group, conditioned on all the individuals assigned to that group. The performance of all individuals under the model for each group is then evaluated, and individuals can be moved between groups as appropriate. At this stage, new groups can also be created, or groups can be combined, as appropriate. This process is repeated until convergence, when all individuals are assigned to the group with the model that best represents them.

7.5.1 Related work

Working on data from the StudentLife study (see section 2.2.4.1 for details), Farhan et al. (2016b) performed a similar clustering using a linear algebra technique by Sun et al. (2015). If $\mathbf{Z}$ is an $N \times p$ feature matrix, $\mathbf{Z}$ can be approximated using vectors $\mathbf{u}$ of length N and $\mathbf{v}$ of length p, where $\mathbf{Z} \approx \mathbf{u}\mathbf{v}^\top$. Subgroups of rows that have similar behaviour with respect to a subset of features can be found by optimising

$$\min_{\mathbf{u},\mathbf{v}} ||\mathbf{Z} - \mathbf{u}\mathbf{v}^\top||_F^2 \tag{7.1a}$$

subject to

$$||\mathbf{u}||_0 \leq s_u \tag{7.1b}$$

$$||\mathbf{v}||_0 \leq s_v \tag{7.1c}$$

where $||\cdot||_F^2$ denotes the Frobenius norm and $||\cdot||_0$ is the number of non-zeros in the vector. This has the effect of finding clusters of points that have similar behaviour

given a subset of features. The rows in $\mathbf{Z}$ corresponding to the non-zero elements of the optimised $\mathbf{u}$ can then be removed and the optimisation repeated until all clusters have been found. While this is an interesting approach that results in sparse feature allocations for each cluster, s_u and s_v must be appropriately optimised for it to work correctly.

Personalization by finding similar individuals in the population has long been considered in the area of activity recognition from sensor data (Lara & Labrador, 2013). For example, an approach to finding groups of individuals based on the similarity of their demographics, lifestyles and sensor data was developed by Lane et al. (2011) and extended by Abdullah et al. (2012). This method will be described in greater detail in section 7.11.1.6. An advantage of these types of methods that use only the demographic and sensor data for clustering is that they do not require any calibration data for new individuals. However, clustering on the distributions of the sensor data and not on the relationships between the sensor data and the model output (in the present case, mental state) may miss important differences or similarities between individuals.

7.6 Grouped Bayesian Lasso overview

The generic method described in the previous section, and shown in figure 7.5, is highly simplistic and there are several practical considerations. This section gives an overview of the grouped Bayesian Lasso, depicted in figure 7.6, which has been developed to dynamically group and find similar individuals in the data set.

Finding the clusters of similar individuals in the grouped Bayesian Lasso is performed as a two-pass process: first grouping the individuals based on the raw feature data; and then clustering them based on how commonly they are grouped in together in the first pass. Through the rest of this chapter, 'grouping' will be used to refer to the first pass, and 'clustering' will be used to refer to the second. The model is built up from several different components that are briefly described here, referenced into the following sections where they are fully described.

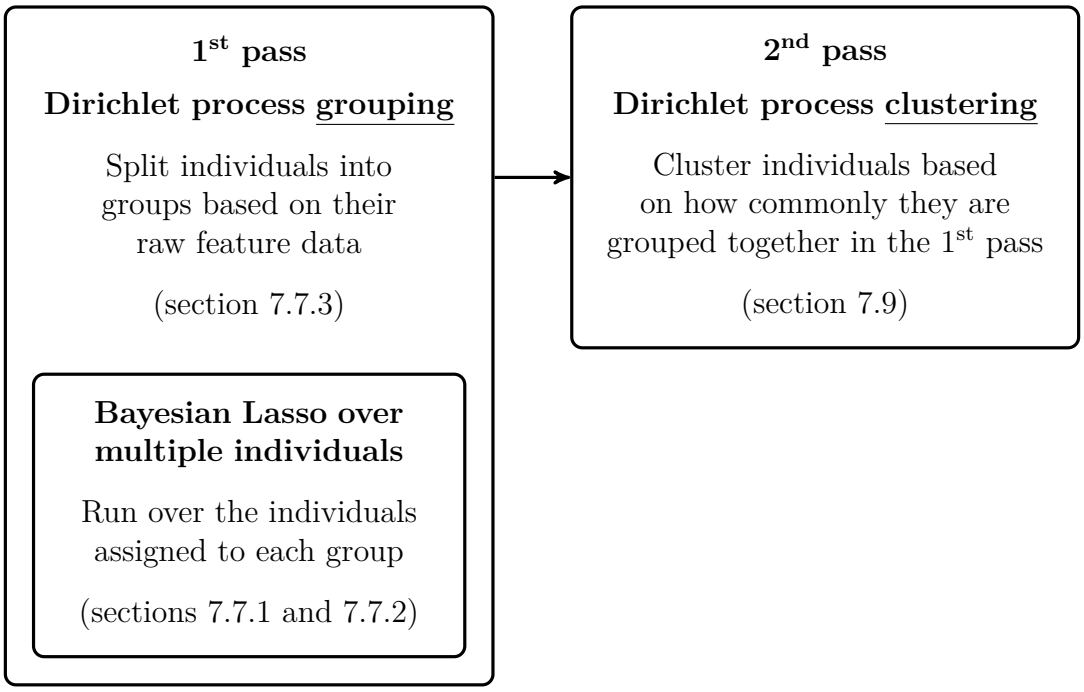


Figure 7.6: Overview of the grouped Bayesian Lasso developed to find clusters of similar individuals, performed as a two-pass process. An overview of the whole process is given in section 7.6; the 1st pass is described in detail in section 7.7; and the 2nd pass is described in section 7.9. Results from the 1st pass and justification for the two-pass process are given in section 7.8.

Ideally the grouping and clustering would be performed in a single pass as part of the grouping process. However, in practice it was found that the grouping model did not converge to stationary clusters, hence the need for the two-pass process. This will be explained in greater detail in section 7.8.

7.6.1 1st pass grouping of individuals overview

The first pass of the method groups individuals based on their raw feature data. This is the core part of the model that finds the groups of individuals based on their similarity. The need for the second pass will be explained in the next subsection.

Starting at the left in figure 7.5, the first consideration is how to split the N individual into groups, and how to chose the number of groups, K. Because the number of groups is difficult to estimate in advance, it would be beneficial to enable the number of clusters to adapt to the available individuals, while ensuring that

$K < N$. The Dirichlet process, described fully in section 3.3, provides an ideal base for grouping similar individuals due to the dynamic nature in which it allocates groups. The Dirichlet process is also particularly suitable because unlike many other clustering methods (k-means, DBSCAN, Gaussian mixture models, etc.) it does not require a distance metric between points, but instead wraps around a Bayesian model, where the 'distance' to each group is determined by the joint probability of the data point and the model parameters for the group. The difficulty of devising a distance metric between regression models for different individuals was shown in section 7.4.

The next consideration is the regression model to use to estimate the output, in this case level of depression, from the feature data. The Lasso, described in section 3.6.1, is a commonly used regression model that can improve test performance by reducing overfitting on the training data. It is particularly suitable for use for grouping individuals because it can pick out a subset of features that may characterise the group. A Bayesian version of the Lasso, developed by Park & Casella (2008), is used as the core model to group individuals, and combines naturally with the Dirichlet process. Thus the first pass of the method splits the individuals into groups that have similar regression models linking their geolocation-derived features to their level of depression.

7.6.2 2nd pass clustering of individuals overview

Ideally, the 1st pass grouping would converge to a steady state where all individuals are allocated to the group that matches them best. In practice the noisy geolocation data meant that individuals would often fit reasonably well into multiple groups, and would regularly switch group memberships. Additionally, when an individual joins or leaves a group, this changes the model that best fits the new group members. This ability for individuals to keep switching between groups is one of the key features of the Dirichlet process prior, but may not lead to useful and identifiable clusters of individuals. The effect of this will be depicted in greater detail in section 7.8.

A second pass is therefore used to cluster individuals based on which other individuals they most commonly occurred while sampling the 1st pass grouping. As

with the first pass, the correct number of clusters to find is unknown and a standard Dirichlet process is used once again. This 2nd pass clustering method is fully described in section 7.9

7.7 1st pass grouping of individuals

As described in section 7.6, the grouped Bayesian Lasso is broken into two passes, the first of which is to split all the individuals into groups that are represented by the same regression model linking their geolocation-derived features with their level of depression.

Regularised linear regression is used as the central regression model for depression estimation in the grouped model. Linear regression was shown in section 6.4 in the previous chapter to provide reasonable results for estimating depression in the BD participants so it is a natural choice to base the personalised model on. The regularisation of the linear regression means that only a subset of the full set of features will be included in each group identified. It has also been shown by (Leng et al., 2014) that the Bayesian Lasso model introduced below can be generalised to approximate a GLM applied to linear regression described in section 3.1.2. It should be straightforward to apply this generalisation to the grouped model presented here for additional non-linearity, but this is left as future work.

The core of the grouped method is therefore the "Bayesian Lasso" model by Park & Casella (2008), described fully in section 7.7.1. The Bayesian Lasso is extended to form the "Bayesian Lasso over multiple individuals" in section 7.7.2 by adding the ability to directly handle multiple individuals (section 7.7.2.1) and making a modification to the way that the offset β_0 parameter is sampled (section 7.7.2.2). Inference of the final Bayesian Lasso model over multiple individuals is described in section 7.7.2.3.

The grouping of individuals is performed by applying a Dirichlet process prior to the core model variables in the Bayesian Lasso model over multiple individuals. The full Dirichlet process model is defined in section 7.7.3. The standard Dirichlet

process sampling algorithm only allows new groups to be created through draws of model variables from the prior, but in this application it led to reduced mixing due to the low likelihood of drawing suitable regression model coefficients from the prior. A variation is therefore made to the sampling algorithm to enable creating new groups conditioned on the current model variables, described in detail in section 7.7.3.3.

7.7.1 The Bayesian Lasso

This section describes the background to the Bayesian Lasso, developed by Park & Casella (2008). This model, with the extensions described in section 7.7.2, forms the core model used in the 1st pass grouping process.

The standard Lasso model (developed by Tibshirani, 1996, and described fully in section 3.6.1) is an extension to linear models of the form

$$\hat{y} = f(\beta_0 + \boldsymbol{\beta}^\top \mathbf{z}) \tag{7.2}$$

where $\hat{y}$ is an output estimated from a set of p predictors, $\mathbf{z} = [z_1, z_2, \ldots, z_p]$; β_0 is the offset; and $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^\top$ is a vector of the regression coefficients. The model presented here considers a standard linear regression model of the form

$$\hat{y} = \beta_0 + \boldsymbol{\beta}^\top \mathbf{z}. \tag{7.3}$$

The Lasso applies ℓ_1 regularisation to the $\boldsymbol{\beta}$ parameter in equation (7.3) to minimise

$$\underset{\beta_0, \boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_i^N \left(y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{z}_i \right)^2 + \lambda \sum_j^p |\beta_j| \right\} \tag{7.4}$$

over all N data samples, which has the effect of encouraging some or all of the components of $\boldsymbol{\beta}$ to be exactly zero. The strength of the regularisation is controlled by the value of the λ parameter. Regularising the model coefficients in this way reduces over-fitting of the model to the training data; increases interpretability of the model (by only presenting the most relevant features); and in this application can find groups of features that are relevant across multiple individuals.

In a Bayesian setting, with $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top$ a matrix of features for all N training samples, the linear regression model of equation (7.3) can be represented as

the following hierarchical model, shown as the graphical model in figure 7.7(a) on page 197:

$$\mathbf{y} \,|\, \mathbf{Z}, \boldsymbol{\beta}, \beta_0, \sigma^2 \;\sim\; \mathcal{N}_N(\beta_0 \mathbf{1}_N + \mathbf{Z}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_N) \tag{7.5a}$$

$$\boldsymbol{\beta} \;\sim\; \pi(\boldsymbol{\beta}) \tag{7.5b}$$

$$\beta_0 \;\sim\; \pi(\beta_0) \tag{7.5c}$$

$$\sigma^2 \;\sim\; \pi(\sigma^2) \tag{7.5d}$$

Park & Casella (2008) extend this model to implement a Bayesian version of the Lasso by using zero-centred Laplace (double exponential) priors on the coefficients of $\boldsymbol{\beta}$ of the form

$$\pi(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{\lambda}{2} \exp\Big\{-\lambda|\beta_j|\Big\}. \tag{7.6}$$

Using this prior, the joint distribution of $\mathbf{y}$, $\boldsymbol{\beta}$, β_0, and σ^2 can be represented as

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\beta}, \beta_0, \sigma^2 \,|\, \mathbf{Z}) =& \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \beta_0\mathbf{1}_N - \mathbf{Z}\boldsymbol{\beta})^\top(\mathbf{y} - \beta_0\mathbf{1}_N - \mathbf{Z}\boldsymbol{\beta})\right\} \\ & \times \left[\prod_{j=1}^{p} \frac{\lambda}{2} \exp\Big\{-\lambda|\beta_j|\Big\}\right] \times \pi(\beta_0) \times \pi(\sigma^2) \qquad (7.7) \\ =& \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \beta_0\mathbf{1}_N - \mathbf{Z}\boldsymbol{\beta})^\top(\mathbf{y} - \beta_0\mathbf{1}_N - \mathbf{Z}\boldsymbol{\beta}) - \lambda\sum_{j=1}^{p}|\beta_j|\right\} \\ & \times \frac{1}{(2\pi\sigma^2)^{N/2}} \times \frac{\lambda^p}{2^p} \times \pi(\beta_0) \times \pi(\sigma^2) \qquad (7.8) \\ =& \exp\left\{-\frac{1}{2\sigma^2}\sum_{i}^{N}\Big(y_i - \beta_0 - \boldsymbol{\beta}^\top\mathbf{z}_i\Big)^2 - \lambda\sum_{j=1}^{p}|\beta_j|\right\} \times \cdots \end{aligned}$$

from which it is clear that maximising the joint in equation (7.8) is broadly equivalent to minimising equation (7.4).

The Laplace prior of equation (7.6) is convenient because it provides an equivalent to the Lasso in the joint density function. However, it is not conjugate to the other parameters in the model so it is not possible to construct a Gibbs sampler using it directly. To overcome this, Park & Casella instead utilise an alternative representation

of the Laplace distribution as a scale mixture of Gaussians with an auxiliary variable τ, first described by Andrews & Mallows (1974). Defining the hierarchical model

$$\beta \,|\, \tau^2, \;\sim\; \mathcal{N}(0, \tau^2) \tag{7.9a}$$

$$\tau^2 \,;\, \lambda \;\sim\; \mathrm{Exp}\left(\frac{\lambda^2}{2}\right) \tag{7.9b}$$

and marginalising the joint density over τ^2 results in

$$f_\beta(\beta) \;=\; \int_0^\infty \frac{1}{\sqrt{2\pi\tau^2}} \cdot \exp\left\{-\frac{\beta^2}{2\tau^2}\right\} \cdot \frac{\lambda^2}{2} \cdot \exp\left\{-\frac{\lambda^2\tau^2}{2}\right\} d\tau^2, \tag{7.10}$$

which is equivalent to $\frac{\lambda}{2}\exp\{-\lambda|\beta|\}$. This equivalence is demonstrated in section E.1 of appendix E.

The Bayesian Lasso is therefore defined by Park & Casella as the following hierarchical model, shown as the graphical model in figure 7.7(b):

$$\mathbf{y} \,|\, \mathbf{Z}, \boldsymbol{\beta}, \beta_0, \sigma^2 \;\sim\; \mathcal{N}_N(\beta_0 \mathbf{1}_N + \mathbf{Z}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_N) \tag{7.11a}$$

$$\boldsymbol{\beta} \,|\, \tau_1^2, \ldots, \tau_p^2, \sigma^2 \;\sim\; \mathcal{N}_p(\mathbf{0}_p, \sigma^2 \boldsymbol{D}), \quad \boldsymbol{D} = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2) \tag{7.11b}$$

$$\tau_1^2, \ldots, \tau_p^2 \,;\, \lambda \;\sim\; \mathrm{Exp}_p(\tfrac{\lambda^2}{2}), \quad \tau_1^2, \ldots, \tau_p^2 > 0 \tag{7.11c}$$

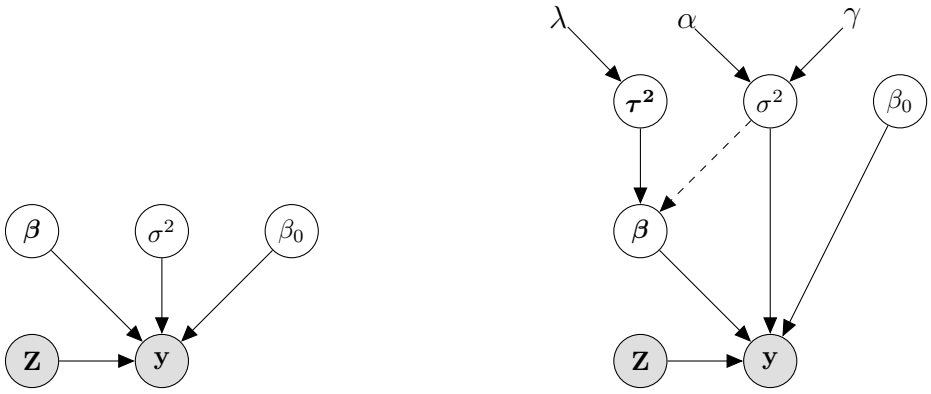
$$\beta_0 \;\sim\; \pi(\beta_0), \quad \pi(\beta_0) = 1 \tag{7.11d}$$

$$\sigma^2 \,;\, \alpha, \gamma \;\sim\; \text{Inv-Gamma}(\alpha, \gamma) \tag{7.11e}$$

The full joint density of the Bayesian Lasso is therefore

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\beta}, \beta_0, \sigma^2 \,|\, \mathbf{Z}) \;=\; & \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \beta_0\mathbf{1}_N - \mathbf{Z}\boldsymbol{\beta})^\top(\mathbf{y} - \beta_0\mathbf{1}_N - \mathbf{Z}\boldsymbol{\beta})\right\} && (\; p(\mathbf{y} \,|\, \cdot)\;) \\ & \times \frac{\gamma^a}{\Gamma(a)} \left(\sigma^2\right)^{-a-1} \exp\left\{-\frac{\gamma}{\sigma^2}\right\} && (\; \pi(\sigma^2)\;) \\ & \times \prod_{j=1}^{p} \frac{1}{(2\pi\sigma^2\tau_j^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\} && (\; \pi(\beta_j \,|\, \tau_j^2, \sigma^2)\;) \\ & \times \frac{\lambda^2}{2} \exp\left\{-\frac{\lambda^2\tau_j^2}{2}\right\}. && (\; \pi(\tau_j^2 \,;\, \lambda)\;) \end{aligned} \tag{7.12}$$

All parameters in the Bayesian Lasso are sampled in a Gibbs sampler. The conditioning of $\boldsymbol{\beta}$ on σ^2 in equation (7.11b) is optional and changes the double exponential



(a) Standard regression. (b) Bayesian Lasso regression.

Figure 7.7: Graphical model of (a) a standard linear regression model defined in equation (7.5); and (b) the Bayesian Lasso model by Park & Casella (2008) defined in equation (7.11). The dashed dependency from $\boldsymbol{\beta}$ on σ^2 indicates that the condition is optional – see text for details.

prior on $\boldsymbol{\beta}$ in equation (7.6) to the form

$$\pi(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left\{-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right\}. \tag{7.13}$$

This has the effect of making the value of λ scale invariant as well as reducing the possibility of multi-modal posterior distributions. The prior of equation (7.6) can be obtained by omitting the condition on σ^2 in equation (7.11b) of the model.

⟶

7.7.1.1 Tests on simulated data

To demonstrate the performance of the Bayesian Lasso model, sample data were generated as follows:

$$p = 10 \qquad \text{(number of predictors)} \tag{7.14a}$$

$$N = 100 \qquad \text{(number of data points)} \tag{7.14b}$$

$$\boldsymbol{\beta}\,;\, p \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p) \qquad \text{(global "true" model coefficients)} \tag{7.14c}$$

$$\boldsymbol{\epsilon}\,;\, p \sim \text{Inv-Gamma}_p(1, 0.1) \qquad \text{(noise level in individual coefficients)} \tag{7.14d}$$

$$\boldsymbol{\beta}_i \,|\, \boldsymbol{\beta}, \boldsymbol{\epsilon} \sim \mathcal{N}_p(\boldsymbol{\beta}, \text{diag}(\boldsymbol{\epsilon})) \qquad \forall \;\; 1 \leq i \leq N \qquad \text{(coefficients for point } i\text{)} \tag{7.14e}$$

$$\mathbf{z}_i\,;\, p \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p) \qquad \forall \;\; 1 \leq i \leq N \qquad \text{(predictors for point } i\text{)} \tag{7.14f}$$

$$y_i \,|\, \mathbf{z}_i, \boldsymbol{\beta}_i \sim \mathcal{N}(\mathbf{z}_i^\top \boldsymbol{\beta}_i, 0.2^2) \qquad \forall \;\; 1 \leq i \leq N \qquad \text{(data labels for point } i\text{)} \tag{7.14g}$$

A global "true" value of $\boldsymbol{\beta}$ is first drawn from a standard normal prior. For each of the N data points individually, additive Gaussian noise is applied to this prior with variance $\boldsymbol{\epsilon}$, drawn globally from an inverse gamma prior. Random feature values in feature matrix $\mathbf{Z}$ are drawn from a standard Gaussian distribution, and the model output, y_i, is set with a small amount of Gaussian noise for each sample.

The Bayesian Lasso model by Park & Casella (2008) was run over $\mathbf{Z}$ and $\mathbf{y}$, the task being to estimate the value of the $\boldsymbol{\beta}$ coefficients, denoted $\hat{\boldsymbol{\beta}}$. The λ parameter was varied between 1 and 100, and the Gibbs sampler was run for 10 000 iterations for each parametrisation. Figure 7.8 shows summary statistics of the 10 sampled $\hat{\beta}_i$ feature coefficients as λ is varied. The $\hat{\beta}_i$ coefficients are summarised by (a) the mean; (b) the median; (c) the mode[1]; and (d) the MAP estimate.[2] Both the mean and median smoothly shrink towards zero, whereas the mode appears to exhibit roughly linear shrinkage. There is noise in the mode estimates due to the way that they are

[1] The mode was calculated from the sampled data using the iterative half-sample mode estimator algorithm of Robertson & Cryer (1974) (described in Bickel & Frühwirth, 2006).

[2] As discussed by Park & Casella (2008), for a given σ^2, the value of $\hat{\boldsymbol{\beta}}$ that maximises the joint in equation (7.12) is a Lasso estimate. The MAP estimates shown in figures 7.8 and 7.9 are therefore obtained by selecting the value of $\hat{\boldsymbol{\beta}}$ when the sampled values of $\hat{\boldsymbol{\beta}}$, and $\hat{\sigma}^2$ maximise the joint in equation (7.12). The condition on σ^2 in equation (7.11b) means that value of $\hat{\boldsymbol{\beta}}$ resulting in the maximum posterior probability may be different for different values of $\hat{\sigma}^2$.

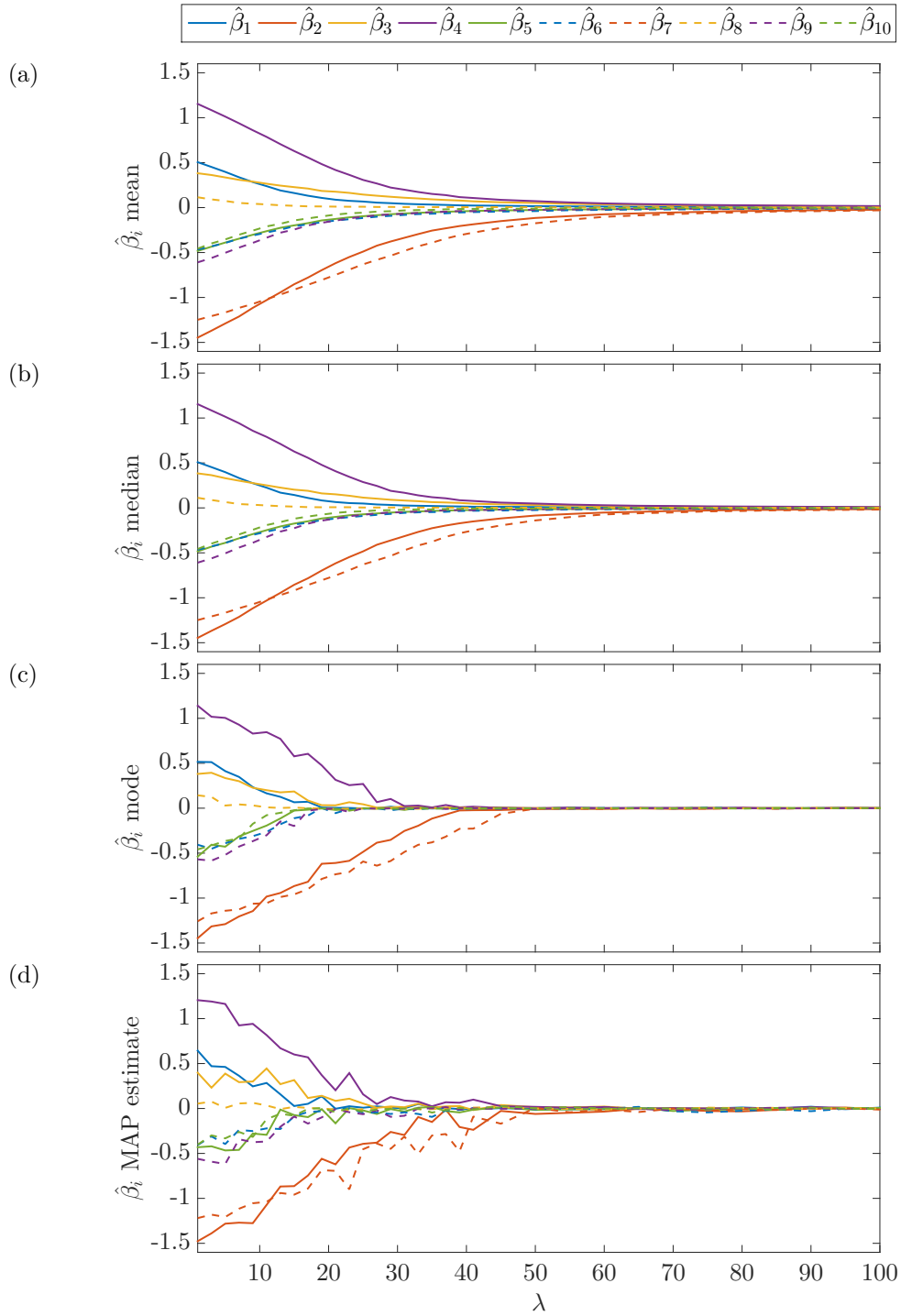


Figure 7.8: Variable shrinkage of $\hat{\boldsymbol{\beta}}$ for varying λ using the Bayesian Lasso model of Park & Casella (2008) run on sample data described in equations (7.14a) to (7.14g). The Bayesian Lasso was run for 10 000 iterations for each value of λ. The mode in (c) was calculated using the iterative half-sample mode estimator algorithm of Robertson & Cryer (1974) and the MAP estimate in (d) is the value of $\hat{\boldsymbol{\beta}}$ when the sampled values of $\hat{\boldsymbol{\beta}}$, and $\hat{\sigma}^2$ maximise the joint in equation (7.12).

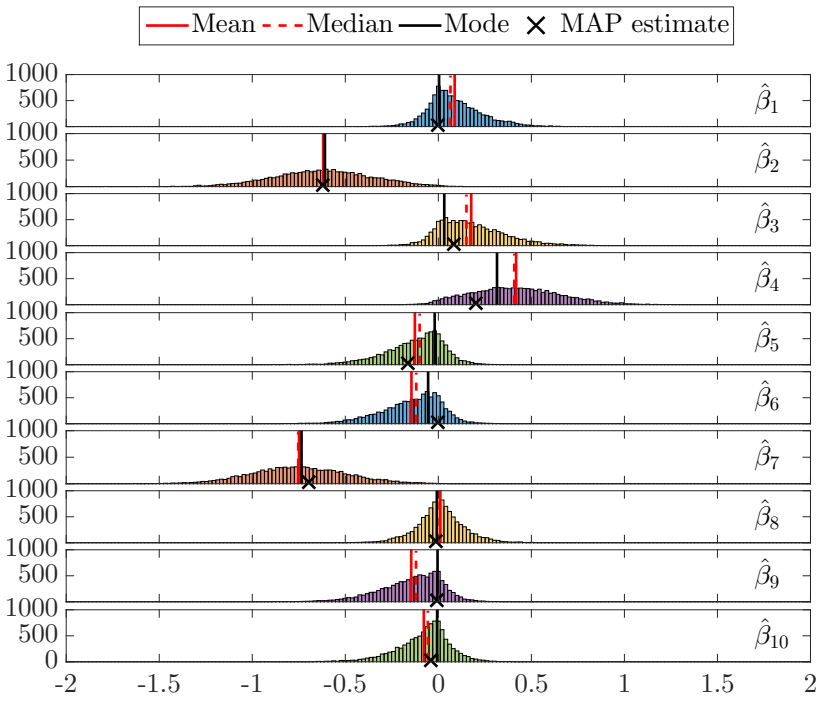


Figure 7.9: Samples of $\hat{\boldsymbol{\beta}}$ using the Bayesian Lasso model of Park & Casella (2008) run on sample data described in equations (7.14a) to (7.14g). The Bayesian Lasso was run for 10 000 iterations with $\lambda = 21$. Core statistics are shown overlaid. The mode was calculated using the iterative half-sample mode estimator algorithm of Robertson & Cryer (1974) (see also Bickel & Frühwirth, 2006). The MAP estimate is the value of $\hat{\boldsymbol{\beta}}$ when the sampled values of $\hat{\boldsymbol{\beta}}$, and $\hat{\sigma}^2$ maximise the joint in equation (7.12).

calculated. The MAP estimate adds further noise, which is not unexpected due to the dependency on $\hat{\sigma}^2$ discussed earlier.

Figure 7.9 above shows the distributions of the samples of $\hat{\boldsymbol{\beta}}$ when $\lambda = 21$, chosen because some coefficients appear fully regularised in figure 7.8 while some are still large. Summary statistics of the mean, median, mode and MAP estimate of each $\hat{\beta}_i$ coefficient are shown overlaid. The distributions for all coefficients appear unimodal, but often not Gaussian and not symmetrical. The mode values of $\hat{\beta}_1$, $\hat{\beta}_3$, $\hat{\beta}_5$, $\hat{\beta}_6$, $\hat{\beta}_8$, $\hat{\beta}_9$, and $\hat{\beta}_{10}$ are ≈ 0, but the mean and median are ≈ 0 only for $\hat{\beta}_8$, which also appears distinctly Laplace distributed. The median value of each coefficient is generally between the mean and the zero (for the asymmetrically distributed coefficients where the mode is ≈ 0), and although not immediately obvious, the median values of $\hat{\beta}_i$ in figure 7.8 also shrink slightly faster than the mean values.

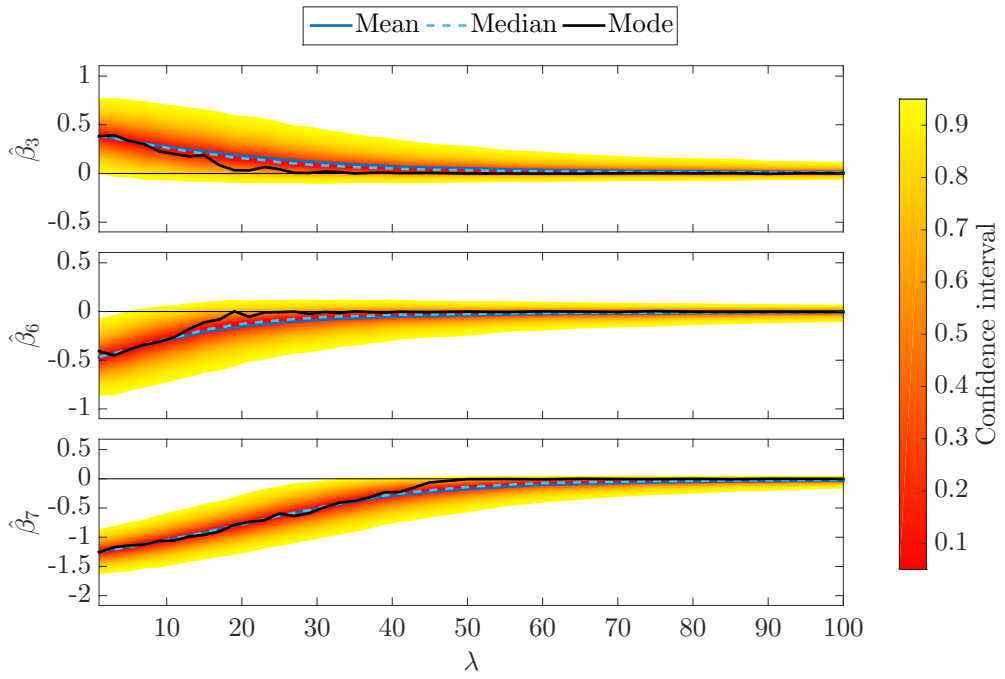


Figure 7.10: Distributions of samples of selected $\hat{\beta}_i$ coefficients for varying λ using the Bayesian Lasso model of Park & Casella (2008) run on sample data described in equations (7.14a) to (7.14g). The mean, median, and mode values are the same as shown in figure 7.8 and the graded background indicates the proportion of samples that were at the particular value of $\hat{\beta}_i$. This shows more clearly how the mean and median smoothly shrink to zero while the mode appears to shrink linearly. The distribution of samples at $\lambda = 100$ shows that the distribution of $\hat{\beta}_6$ is symmetrical at this point, while the distributions of $\hat{\beta}_3$ and $\hat{\beta}_7$ are more asymmetrical.

The distributions of the sampled β_i coefficients and summary statistics for the $\hat{\beta}_3$, $\hat{\beta}_6$, and $\hat{\beta}_7$ coefficients for varying λ are shown in figure 7.10 above. The $\hat{\beta}_7$ coefficient in particular shows that the distribution of the sampled coefficients appears to be symmetrical when λ is small (and the coefficient is far from zero) matching the result in figure 7.9. As λ is increased, and the coefficient approaches zero, the distribution is truncated, and thus skewed, close to the origin, resulting in the asymmetrical distributions discussed earlier. The variance of the sampled values is also fairly constant while lambda is small, decreasing as the distribution becomes more truncated. Figure 7.10 shows more clearly how the mean and median are affected by this asymmetry, while the mode estimate continues shrinking linearly towards zero.

7.7.2 The Bayesian Lasso over multiple individuals

The Bayesian Lasso by Park & Casella, described in detail in the previous section, is designed to work on a single set of predictors, $\mathbf{Z}$, and a single set of outputs, $\mathbf{y}$. This section describes two extensions to the basic model that enable appropriate performance for grouping individuals. The full Bayesian Lasso over multiple individuals model, including both extensions, is given in equation (7.21) in section 7.7.2.3 on page 205, and a graphical model is shown in figure 7.11 on page 209. Inference of the model is also described in section 7.7.2.3.

7.7.2.1 Handling multiple individuals

This section describes an extension to the basic model to directly handle a set of M indiviuals, where each individual has a set of predictors, $\mathbf{Z}_k$, and a set of outputs, $\mathbf{y}_k$.

In this case, it is convenient to define the model directly over the set of individuals, i.e. $\forall\ k \in [1, \ldots, M]$. To do this, equation (7.11a) in the Bayesian Lasso model is replaced with

$$\mathbf{y}_k \,|\, \mathbf{Z}_k, \boldsymbol{\beta}, \beta_0, \sigma^2 \ \sim\ \mathcal{N}_N(\beta_0 \mathbf{1}_{N_k} + \mathbf{Z}_k \boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_{N_k}), \tag{7.15}$$

resulting in the plate shown in the graphical model in figure 7.11 on page 209. This translates through to a product over all M individuals in the joint, or a sum in the exponential power. Conjugacy is not affected, and all parameters can still be sampled in a Gibbs sampler. This is included in the definition of the Gibbs sampler for the Bayesian Lasso over multiple individuals given in section 7.7.2.3.

7.7.2.2 Sampling the offset β_0

Park & Casella chose to place a non-informative flat prior on β_0 in the Bayesian Lasso, as shown in equation (7.11d) of the model definition. The conditional density

of $p(\mathbf{y}, \beta_0 \,|\, \cdot)$ is then decomposed as

$$p(\mathbf{y}, \beta_0 \,|\, \cdot) \;\propto\; \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \beta_0\mathbf{1}_N - \mathbf{Z}\boldsymbol{\beta})^\top(\mathbf{y} - \beta_0\mathbf{1}_N - \mathbf{Z}\boldsymbol{\beta})\right\} \tag{7.16}$$

$$\propto \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \bar{y}\mathbf{1}_N - \mathbf{Z}\boldsymbol{\beta})^\top(\mathbf{y} - \bar{y}\mathbf{1}_N - \mathbf{Z}\boldsymbol{\beta})\right\}$$

$$\times \exp\left\{-\frac{N}{2\sigma^2}(\bar{y} - \beta_0)^2\right\} \tag{7.17}$$

by using the identity $\mathbf{y} - \beta_0\mathbf{1}_N = \mathbf{y} - \bar{y}\mathbf{1}_N + \bar{y}\mathbf{1}_N - \beta_0\mathbf{1}_N$ in equation (7.16) and rearranging. Using the decomposition in equation (7.17), β_0 can be sampled in the Gibbs sampler from a Gaussian distribution with

$$\text{mean parameter } \mu'_{\beta_0} = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \;\text{ and} \tag{7.18a}$$

$$\text{variance } \sigma^{2\prime}_{\beta_0} = \frac{\sigma^2}{N}. \tag{7.18b}$$

It is also simple to integrate equation (7.17) over the β_0 parameter and run the model on $\mathbf{y} - \bar{y}\mathbf{1}_N$ directly. Park & Casella argue that the mean parameter "is generally of secondary interest" and in many situations that may be true. However, this only applies when the offset is the same for the whole population, and in personalised modelling this would not be the case because different individuals may have different baseline values outside of the available data. When grouping individuals, the offset will apply to the whole group, not the individuals themselves. It would instead be beneficial to sample different combinations of baseline value β_0 and offset coefficients $\boldsymbol{\beta}$ based on suitably chosen priors.

The Gibbs sampler for the β_0 parameter in equation (7.18), derived from equation (7.17), also has two drawbacks. Firstly, the distribution of β_0 is only conditioned on the recorded data $\mathbf{y}$ (via $\bar{y}$) and the sampled values of σ^2. Secondly, none of the other sampled parameters are conditioned on β_0, meaning that the sampled values of β_0 are not used later and they do not influence the final model in any way. The form of the variance $\sigma^{2\prime}_{\beta_0}$ also implies that $\sigma^{2\prime}_{\beta_0} \to 0$ as $N \to \infty$, meaning that the more data that are available for training the model, the more narrowly spiked the distribution of β_0 will become around $\bar{y}$.

The form of equation (7.17) indicates an alternative interpretation of the first part of the hierarchical model with $\bar{y}$ as a sampled parameter, defined as follows:

$$\mathbf{y} \,|\, \mathbf{Z}, \boldsymbol{\beta}, \bar{y}, \sigma^2 \;\sim\; \mathcal{N}_N(\bar{y}\mathbf{1}_N + \mathbf{Z}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N) \tag{7.19a}$$

$$\bar{y} \,|\, \beta_0, \sigma^2 \;\sim\; \mathcal{N}(\beta_0, \tfrac{\sigma^2}{N}) \tag{7.19b}$$

$$\beta_0 \;\sim\; \pi(\beta_0), \quad \pi(\beta_0) = 1 \tag{7.19c}$$

This interpretation is possible because $(\bar{y} - \beta_0)^2 \equiv (\beta_0 - \bar{y})^2$. This still leaves the variance of $\bar{y}$ as a function of the current sampled value of σ^2 and the number of samples, N, which may be unhelpful, especially where the number of samples is large. A custom prior can therefore be applied to the value of $\bar{y}$ that is used throughout the model definition. For consistency, the parameter $\bar{y}$ will be denoted β_0, and the model defined as follows:

$$\mathbf{y} \,|\, \mathbf{Z}, \boldsymbol{\beta}, \beta_0, \sigma^2 \;\sim\; \mathcal{N}_N(\beta_0\mathbf{1}_N + \mathbf{Z}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N) \tag{7.20a}$$

$$\beta_0 \,;\, \mu_{\beta_0}, \sigma^2_{\beta_0} \;\sim\; \mathcal{N}(\mu_{\beta_0,}, \sigma^2_{\beta_0}) \tag{7.20b}$$

Considered as a generative model, this definition makes intuitive sense. The baseline β_0 parameter is sampled from an appropriate Gaussian distribution, parametrised by μ_{β_0} and $\sigma^2_{\beta_0}$. The model output $\mathbf{y}$ is then conditioned on this β_0 variable.

Equations (7.5a) and (7.5c) in the definition of the standard Bayesian Lasso model are therefore replaced with equations (7.20a) and (7.20b). This forms the basis of the Gibbs sampler for the Bayesian Lasso over multiple individuals given in the next section and is shown in the graphical model in figure 7.11 on page 209.

7.7.2.3 Inference in the Bayesian Lasso over multiple individuals

The full model of the Bayesian Lasso over multiple individuals with inferred offset β_0 is therefore defined as follows, shown in figure 7.11 on page 209:

$$\begin{aligned}
\mathbf{y}_k \,\Big|\, \mathbf{Z}_k, \boldsymbol{\beta}, \beta_0, \sigma^2 &\sim \mathcal{N}_{N_k}(\beta_0 \mathbf{1}_{N_k} + \mathbf{Z}_k\boldsymbol{\beta}, \sigma^2\mathbf{I}_{N_k}) && \text{(7.21a)}\\
\boldsymbol{\beta} \,\Big|\, \tau_1^2, \ldots, \tau_p^2, \sigma^2 &\sim \mathcal{N}_p(\mathbf{0}_p, \sigma^2\mathbf{D}), \quad \mathbf{D} = \operatorname{diag}(\tau_1^2, \ldots, \tau_p^2) && \text{(7.21b)}\\
\tau_1^2, \ldots, \tau_p^2 \,;\, \lambda &\sim \operatorname{Exp}_p(\tfrac{\lambda^2}{2}), \quad \tau_1^2, \ldots, \tau_p^2 > 0 && \text{(7.21c)}\\
\beta_0 \,;\, \mu_{\beta_0}, \sigma_{\beta_0}^2 &\sim \mathcal{N}(\mu_{\beta_0}, \sigma_{\beta_0}^2) && \text{(7.21d)}\\
\sigma^2 \,;\, \alpha, \gamma &\sim \text{Inv-Gamma}(\alpha, \gamma). && \text{(7.21e)}
\end{aligned}$$

The full joint density for this model over all variables is written as

$$p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau^2}, \beta_0, \sigma^2 \,|\, \mathbf{Z}) = \left[\prod_{k=1}^{M} f(\mathbf{y}_k \,|\, \mathbf{Z}_k, \boldsymbol{\beta}, \beta_0, \sigma^2)\right] \cdot f(\boldsymbol{\beta} \,|\, \boldsymbol{\tau^2}, \sigma^2) \cdot \pi(\boldsymbol{\tau^2}) \cdot \pi(\beta_0) \cdot \pi(\sigma^2) \quad (7.22)$$

$$= \left[\prod_{k=1}^{M} f(\mathbf{y}_k \,|\, \mathbf{Z}_k, \boldsymbol{\beta}, \beta_0, \sigma^2)\right] \cdot \left[\prod_{j=1}^{p} f(\beta_j \,|\, \tau_j^2, \sigma^2) \cdot \pi(\tau_j^2)\right] \cdot \pi(\beta_0) \cdot \pi(\sigma^2) \quad (7.23)$$

$$\begin{aligned}
&= \left[\prod_{k=1}^{M} \frac{1}{(2\pi\sigma^2)^{N_k/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_k - \beta_0\mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0\mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right\}\right]\\
&\quad \times \left[\prod_{j=1}^{p} \frac{1}{(2\pi\sigma^2\tau_j^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\} \cdot \frac{\lambda^2}{2} \exp\left\{-\frac{\lambda^2\tau_j^2}{2}\right\}\right]\\
&\quad \times \frac{1}{(2\pi\sigma_{\beta_0}^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma_{\beta_0}^2}(\beta_0 - \mu_{\beta_0})^2\right\} \cdot \frac{\gamma^\alpha}{\Gamma(\alpha)} \left(\sigma^2\right)^{-\alpha-1} \exp\left\{-\frac{\gamma}{\sigma^2}\right\} \qquad (7.24)\\
&\propto \left[\prod_{k=1}^{M} \frac{1}{(\sigma^2)^{N_k/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_k - \beta_0\mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0\mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right\}\right]\\
&\quad \times \left[\prod_{j=1}^{p} \frac{1}{(\sigma^2\tau_j^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\} \cdot \exp\left\{-\frac{\lambda^2\tau_j^2}{2}\right\}\right]\\
&\quad \times \exp\left\{-\frac{1}{2\sigma_{\beta_0}^2}(\beta_0 - \mu_{\beta_0})^2\right\} \cdot \left(\sigma^2\right)^{-\alpha-1} \exp\left\{-\frac{\gamma}{\sigma^2}\right\}. \qquad (7.25)
\end{aligned}$$

A Gibbs sampler is constructed from the joint density in equation (7.24), sampling the $\boldsymbol{\tau^2}$, $\boldsymbol{\beta}$, β_0, and σ^2 variables repeatedly until convergence. The distributions used to sample the above variables are given in table 7.2 on the following page, and derivations for these distributions are given in section E.2 of appendix E.

Table 7.2: Gibbs sampler distributions for the Bayesian Lasso over multiple individuals defined in equation (7.21). Derivations are given in section E.2 of appendix E.

Variable	Distribution sampled
τ_j^2	$\eta_j^2 = \frac{1}{\tau_j^2}$ is sampled from an inverse Gaussian distribution with $\text{mean parameter } \mu'_{\eta_j^2} = \sqrt{\frac{\sigma^2\lambda^2}{\beta_j^2}}$ and $\text{scale parameter } \lambda'_{\eta_j^2} = \lambda^2$
$\boldsymbol{\beta}$	$\boldsymbol{\beta}$ is sampled from a multivariate normal distribution with $\text{mean parameter } \boldsymbol{\mu}'_{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{B}^\top$ and $\text{covariance parameter } \boldsymbol{\Sigma}'_{\boldsymbol{\beta}} = \sigma^2\mathbf{A}^{-1}$ where $\mathbf{A} = \mathbf{D}^{-1} + \sum_{k=1}^{M} \mathbf{Z}_k^\top \mathbf{Z}_k, \quad \mathbf{D} = \text{diag}(\tau_1^2, \ldots, \tau_p^2)$ $\mathbf{B} = \sum_{k=1}^{M} (\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k})^\top \mathbf{Z}_k$
β_0	β_0 is sampled from a normal distribution with $\text{mean parameter } \mu'_{\beta_0} = \frac{\sigma^2\mu_{\beta_0} + \sigma^2_{\beta_0}\sum_{k=1}^{M}(\mathbf{y}_k - \mathbf{Z}_k\boldsymbol{\beta})^\top \mathbf{1}_{N_k}}{\sigma^2 + \sigma^2_{\beta_0}\sum_{k=1}^{M} N_k}$ and $\text{standard deviation } \sigma'_{\beta_0} = \sqrt{\frac{\sigma^2\sigma^2_{\beta_0}}{\sigma^2 + \sigma^2_{\beta_0}\sum_{k=1}^{M} N_k}}$
σ^2	σ^2 is sampled from an inverse gamma distribution with $\text{shape parameter } \alpha'_{\sigma^2} = \left(\sum_{k=1}^{M} N_k/2\right) + p/2 + \alpha$ and $\text{scale parameter } \beta'_{\sigma^2} = \frac{1}{2}\sum_{k=1}^{M}(\mathbf{y}_k - \beta_0\mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0\mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})$ $+ \boldsymbol{\beta}^\top\mathbf{D}^{-1}\boldsymbol{\beta}/2 + \gamma, \quad \mathbf{D} = \text{diag}(\tau_1^2, \ldots, \tau_p^2)$

7.7.2.4 Results on geolocation data

The Bayesian Lasso over multiple individuals was used to produce all the estimations shown in figure 7.2 on pages 178 and 179 using the 10 main geolocation-derived features (detailed in section 5.10) calculated on whole calendar weeks of data. For the 'global' model shown in the first column on page 178 and the 'fully personalised' model shown in the fourth column on page 179, the model was run for 1000 iterations with parameter value $\lambda = 0.001$ to minimise regularisation of model coefficients. The 'global' model was trained on all of the data from the individuals not being tested, and up to 8 weeks from the test individual. The 'fully personalised' model was trained only on half the data from the individual being tested, up to a maximum of 8 weeks.

For the 'grouped' and 'personalised grouped' models shown in the second and third columns in figure 7.2, the model was run for 1000 iterations with parameter value $\lambda = 0.4$. This value of λ will be demonstrated to provide optimal performance in section 7.10. In both of these cases, the model was trained on individuals that appeared similar to the test individual using the methods in the following sections. In the 'personalised grouped' model, half the data from the individual being tested, up to a maximum of 8 weeks, were also included in training the model.

7.7.3 Dirichlet process grouping of individuals

The Bayesian Lasso over multiple individuals defined in the previous section provides the core model for a regularised regression model over multiple individuals. However, this model only performs regression over multiple individuals and does not find the distinct groups that are each represented by different regularised regression models. For this functionality, a further layer of processing is required.

A Dirichlet process (DP) model, described fully in section 3.3 on page 54, provides the core ability to find these groupings. The DP model for grouping individuals is defined in section 7.7.3.1 below. Inference in the DP using Algorithm 8 from Neal (2000) is recapped in section 7.7.3.2, and section 7.7.3.3 describes a variation to the standard sampler to improve mixing of groups in this application.

7.7.3.1 Model definition

The DP model is set up with the $\boldsymbol{\beta}$, β_0, and σ^2 parameters in the Bayesian Lasso over multiple individuals defined as grouped variables, indicated by a $^{(g)}$ superscript. The full DP grouping model is therefore defined as follows:

$$\mathbf{y}_k \Big| \mathbf{Z}_k, \boldsymbol{\beta}^{(g)}, \beta_0^{(g)}, \sigma^{2(g)} \sim \mathcal{N}_{N_k}(\beta_0^{(g)}\mathbf{1}_{N_k} + \mathbf{Z}_k\boldsymbol{\beta}^{(g)}, \sigma^{2(g)}\mathbf{I}_{n_k}) \tag{7.26a}$$

$$\boldsymbol{\beta}^{(g)}, \sigma^{2(g)}, \beta_0^{(g)} \Big| G \sim G \tag{7.26b}$$

$$G \Big| G_0 \sim \mathrm{DP}(\alpha_G, G_0) \tag{7.26c}$$

where G_0 is defined as a distribution over the main variables $\{\boldsymbol{\beta}, \beta_0, \sigma^2\}$ and auxiliary variables $\{\tau_1^2, \ldots, \tau_p^2\}$ such that

$$\begin{aligned} G_0^{\boldsymbol{\beta}} \Big| \tau_1^2, \ldots, \tau_p^2, \sigma^2 &= \mathcal{N}_p(\mathbf{0}_p, \sigma^2\mathbf{D}), \\ \tau_1^2, \ldots, \tau_p^2, \sigma^2 &\sim G_0^{\tau_1^2}, \ldots, G_0^{\tau_p^2}, G_0^{\sigma^2}, \\ \mathbf{D} &= \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2) \end{aligned} \tag{7.26d}$$

$$G_0^{\tau_1^2}, \ldots, G_0^{\tau_p^2}\,;\, \lambda = \mathrm{Exp}_p(\tfrac{\lambda^2}{2}), \quad G_0^{\tau_1^2}, \ldots, G_0^{\tau_p^2} > 0 \tag{7.26e}$$

$$G_0^{\beta_0}\,;\, \mu_{\beta_0}, \sigma_{\beta_0}^2 = \mathcal{N}(\mu_{\beta_0}, \sigma_{\beta_0}^2) \tag{7.26f}$$

$$G_0^{\sigma^2}\,;\, \alpha, \gamma = \text{Inv-Gamma}(\alpha, \gamma). \tag{7.26g}$$

A graphical model representation is given in figure 7.12 on the facing page.

7.7.3.2 Standard DP sampler for grouping of individuals

The 1st pass grouping model is sampled from a variant of Algorithm 8 from Neal (2000) as described in section 3.3.1. To recap, Neal's Algorithm 8 is a general algorithm for sampling over arbitrary non-conjugate DP mixtures (such as the model defined above). It defines a set of distinct groups in $\mathcal{C}$, each with parameters ϕ_c. In this case, ϕ_c is the set of sampled $\{\boldsymbol{\beta}, \boldsymbol{\tau^2}, \beta_0, \sigma^2\}$ variables defined above. Each individual k is assigned to one group $c_k \in \mathcal{C}$. On each iteration of the algorithm, individual k can either remain in the same group; be reassigned to one of the other existing groups in $\mathcal{C}$; or be assigned to a new group with initial parameter values drawn from G_0. The set of all existing groups are denoted $\mathcal{C}$, and a set of m new groups with parameters

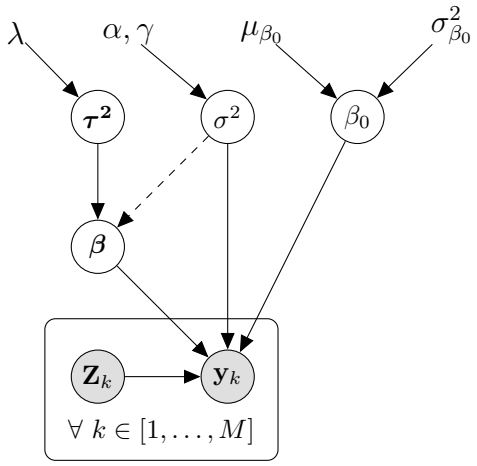


Figure 7.11: Graphical model of the Bayesian Lasso over multiple individuals.

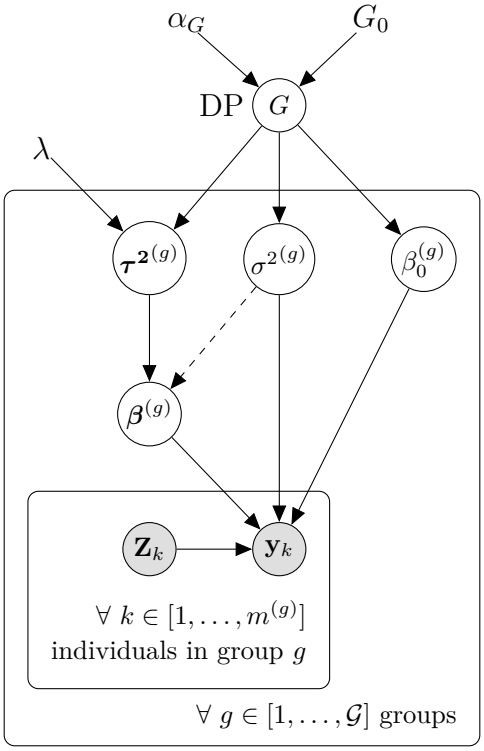


Figure 7.12: Graphical model of the 1st pass grouping of individuals based on raw feature data. Parameters of $\sigma^{2(g)}$ and $\beta_0^{(g)}$ have been omitted for clarity.

ϕ_c drawn from G_0 are denoted $\mathcal{C}^*$. If individual k is currently the only member of class c_k, then the class is removed from $\mathcal{C}$ and added to $\mathcal{C}^*$, which means that it is treated the same as a new draw from the prior. The group c_k for each individual is then sampled according to distribution

$$p(c_k = c \,|\, \cdot) = \begin{cases} b \cdot \frac{n_{-k,c}}{M-1+\alpha_G} \cdot p(\mathbf{y}_k, \phi_c) & \text{if } c \in \mathcal{C} \\ b \cdot \frac{\alpha_G}{|\mathcal{C}^*|} \cdot \frac{1}{M-1+\alpha_G} \cdot p(\mathbf{y}_k, \phi_c) & \text{if } c \in \mathcal{C}^* \end{cases} \tag{7.27}$$

where M is the total number of individuals; $n_{-k,c}$ denotes the number of individuals in class c, not including individual k; $|\mathcal{C}^*|$ is the number of classes in $\mathcal{C}^*$; and b is the appropriate normalising constant. The model parameter α_G controls the mixing rate, i.e. the probability of creating new classes; and the m parameter widens the search space of new classes in each iteration. If $c_k \in \mathcal{C}^*$ (i.e. individual k was assigned to a new class) then c_k is added to $\mathcal{C}$. The basic sampling algorithm is summarised in algorithm 3.2 on page 60.

For evaluating the probability of group allocation with equation (7.27), the joint density $p(\mathbf{y}_k, \phi_c \,|\, \mathbf{Z}_k)$ is the full joint in equation (7.24) for a single individual, k, marginalised over the $\boldsymbol{\tau}^2$ variable, defined as

$$p(\mathbf{y}_k, \phi_c \,|\, \mathbf{Z}_k) = \int_{-\infty}^{\infty} p(\mathbf{y}_k, \boldsymbol{\beta}, \boldsymbol{\tau}^2, \beta_0, \sigma^2 \,|\, \mathbf{Z}_k)\; d\boldsymbol{\tau}^2 \tag{7.28}$$

$$\begin{aligned} &= \frac{1}{(2\pi\sigma^2)^{N_k/2}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top (\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta}) \right\} \\ &\quad \times \frac{1}{(2\pi\sigma^2_{\beta_0})^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2_{\beta_0}} (\beta_0 - \mu_{\beta_0})^2 \right\} \\ &\quad \times \frac{\gamma^\alpha}{\Gamma(\alpha)} \left(\sigma^2\right)^{-\alpha-1} \exp\left\{ -\frac{\gamma}{\sigma^2} \right\} \cdot \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left\{ -\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}} \right\}. \end{aligned} \tag{7.29}$$

A full derivation is given in section E.3 of appendix E. Constant normalising terms can also be removed, leaving

$$\begin{aligned} p(\mathbf{y}_k, \phi_c \,|\, \mathbf{Z}_k) \propto\; &\frac{1}{(\sigma^2)^{N_k/2}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top (\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta}) \right\} \\ &\times \exp\left\{ -\frac{1}{2\sigma^2_{\beta_0}} (\beta_0 - \mu_{\beta_0})^2 \right\} \cdot \left(\sigma^2\right)^{-\alpha-1} \exp\left\{ -\frac{\gamma}{\sigma^2} \right\} \\ &\times \prod_{j=1}^{p} \frac{\lambda}{\sqrt{\sigma^2}} \exp\left\{ -\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}} \right\}. \end{aligned} \tag{7.30}$$

7.7.3.3 Sampler variation to improve mixing of groups

In preliminary testing of the grouping model using Neal's Algorithm 8 as outlined in sections 3.3.1 and 7.7.3.2, a limitation was identified that effects mixing of groups in this model.

Demonstration of the limitation The under-mixing of groups can be seen by considering the same test data used for testing the standard Bayesian Lasso model in section 7.7.1.1 (see equation (7.14) for the test data generation model). The three main non-zero $\hat{\beta}_i$ coefficients from figure 7.9 on page 200 are repeated in figure 7.13 below. The prior probability of β_i is also shown, numerically integrated over σ^2. It is clear that random draws from the prior would be unlikely to provide the values of the three coefficients.

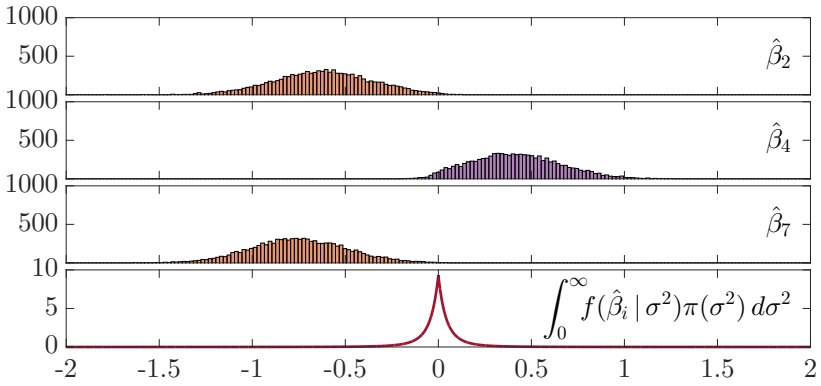


Figure 7.13: Three main non-zero $\hat{\beta}_i$ coefficients from figure 7.9 with the prior density of $\hat{\beta}_i$, numerically marginalised over σ^2. The probability of drawing the mode of any of the three $\hat{\beta}_i$ coefficients from the prior is ≈ 0, and the probability of drawing all three is even smaller. This demonstrates the under-mixing of groups, as described in section 7.7.3.3, because it means that drawing samples near to the "correct" coefficients directly from the prior for a given individual is extremely unlikely.

The effect of this prior model can also be seen by considering the joint density of the Bayesian Lasso, $p(\mathbf{y}, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2 \mid \mathbf{Z})$, calculated using equation (7.12), for sampled variable values using the test data points $\mathbf{y}$ and $\mathbf{Z}$ generated as described in section 7.7.1.1. This is demonstrated in figure 7.14 on page 214. Figure 7.14(a) shows the log joint probability for variables sampled from 10 000 iterations of Gibbs sampling $\hat{\beta}_1, \ldots, \hat{\beta}_p$ and $\hat{\sigma}^2$ from the Bayesian Lasso Gibbs sampler conditioned on $\mathbf{y}$ and $\mathbf{Z}$. These are the same samples shown in figures 7.9 and 7.13. The log probabilities of the Gibbs sampled variables are mainly spread between -150 and -138. Figure 7.14(b), however, shows the log joint probability of 10 000 random draws of $\hat{\beta}_1, \ldots, \hat{\beta}_p$ and $\hat{\sigma}^2$ from the prior model defined in equations (7.11b) to (7.11e). The range of log probabilities now stretches from a maximum value of around -150 down to -4000. This is not surprising given the narrow prior distribution of the $\hat{\beta}_i$ coefficients in the model seen in figure 7.13. Finally, Figure 7.14(c) shows the combined distributions of the sampled variables and the draws from the prior model, restricted to log probabilities between -200 and -130. This shows more clearly the level of overlap between samples drawn from the prior and samples from the Gibbs sampler. Clearly samples from the prior almost never provide joint probabilities near the mode of the Gibbs sampled values, and certainly not towards the higher tails of the distribution.

Solution The implemented solution to this limitation, described in detail in algorithm 7.1, is to allow new samples to be drawn both from the prior directly, but also as Metropolis-Hastings samples based on the current group variables (see section 3.2.4.1 on page 50 for background to the Metropolis-Hastings sampling algorithm).

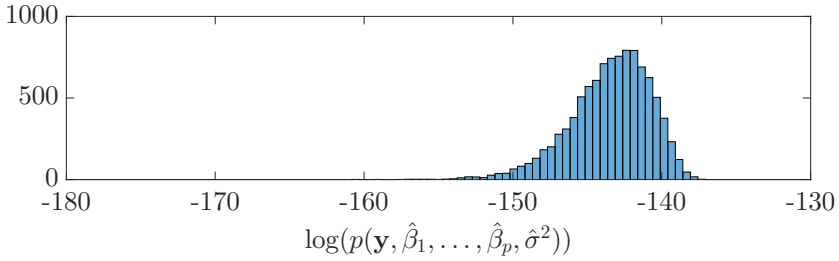
Referring to Radford Neal's Algorithm 8 (Neal, 2000) outlined in algorithm 3.2 on page 60, this modifies the creation of $\mathcal{C}^*$, the set of m new groups. Ordinarily, $\mathcal{C}^*$ is created using m random draws from G_0, the prior model, in this case equations (7.26d) to (7.26g). In the solution presented here, when sampling the group for individual k, proposed new group variable values (denoted *) σ^{2^*}, $\boldsymbol{\beta}^*$, and β_0^* are drawn from Gaussian distributions centred on the current sampled values for the

Algorithm 7.1 Drawing of new samples in the Dirichlet process grouping.

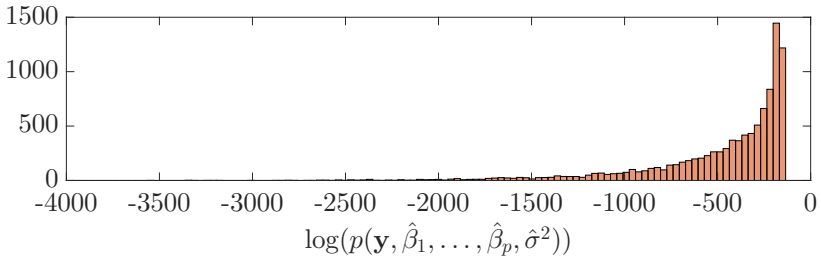
data: $\mathbf{y}_k, \mathbf{Z}_k$ Observed data for participant k.
$\hat{\sigma}^2, \hat{\boldsymbol{\beta}}, \hat{\beta}_0$ Current sampled variables for group.
result: $\mathcal{C}^*$ Set of new groups to sample from.

function DRAWNEWSAMPLES($\mathbf{y}_k, \mathbf{Z}_k, \hat{\sigma}^2, \hat{\boldsymbol{\beta}}, \hat{\beta}_0$)
    Let $\mathcal{C}^*$ be a set of m new groups, each with undefined parameters ϕ_c.
    for $c \in \mathcal{C}^*$ **do** ▷ Set initial parameter values for new groups.
        Draw proposed parameters conditioned on $\hat{\sigma}^2$; $\hat{\boldsymbol{\beta}}$; and $\hat{\beta}_0$.
        $\sigma^{2^*} \sim \mathcal{N}(\hat{\sigma}^2, \hat{\sigma}^2)$
        $\boldsymbol{\beta}^* \sim \mathcal{N}_p(\hat{\boldsymbol{\beta}}, \mathbf{I}_p)$
        $\beta_0^* \sim \mathcal{N}(\hat{\beta}_0, \sigma^{2^*})$
        Calculate Metropolis-Hastings acceptance for proposed parameters, using joint in equation (7.29).
        $r \leftarrow \min\left(1, \dfrac{p(\mathbf{y}_k, \hat{\sigma}^2, \hat{\boldsymbol{\beta}}, \hat{\beta}_0 \mid \mathbf{Z}_k)}{p(\mathbf{y}_k, \sigma^{2^*}, \boldsymbol{\beta}^*, \beta_0^* \mid \mathbf{Z}_k)}\right)$
        $r^* \leftarrow r \times \nu$ ▷ Control mixing rate using ν, $0 \leq \nu \leq 1$.
        if $\mathcal{U}(0,1) < r^*$ **then**
            $\phi_c \leftarrow \{\sigma^{2^*}, \boldsymbol{\beta}^*, \beta_0^*\}$ ▷ Assign proposed parameters to group
        else
            $\phi_c \sim G_0$ ▷ Assign parameters of group from DP prior
        end if
    end for
    return $\mathcal{C}^*$
end function

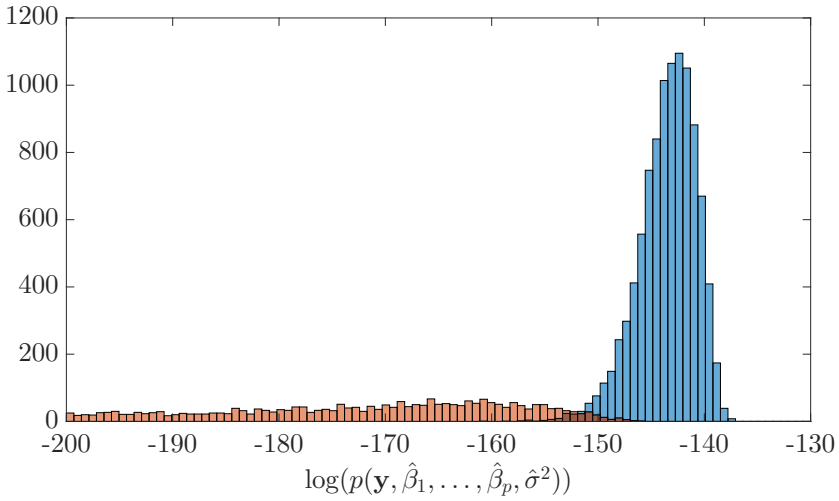
group that the individual currently belongs to. The σ^{2^*} variable is drawn first from a Gaussian distribution with variance of $\hat{\sigma}^2$; the β_i^* variables are drawn from Gaussian distributions with variance of 1; and finally the β_0^* variable is drawn from a Gaussian distribution with variance of σ^{2^*}. The acceptance probability, r, is calculated as per equation (3.39) using the observed data for the individual, and the current and proposed variables. A new model parameter, ν, is introduced to scale the acceptance rate, i.e. $r' = r \times \nu$, which provides a method to control how much the model is allowed to explore the area surrounding the current samples against the whole prior. The proposed samples are now accepted with probability r'. If they are rejected, then samples are drawn from the prior instead. This process is repeated until m new classes have been created.



(a) Histogram of log joint probability of variables sampled conditioned on simulated $\mathbf{y}$ and $\mathbf{Z}$.



(b) Histogram of log joint probability of variables drawn from the prior model.



(c) Histogram of log joint probability of sampled variables and variables drawn from the prior model.

Figure 7.14: Histograms of log joint probability, $p(\mathbf{y}, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2 \mid \mathbf{Z})$, of test data points $\mathbf{y}$ and $\mathbf{Z}$ generated as described in section 7.7.1.1 with variables $\hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2$ (a) sampled from 10 000 iterations of Gibbs sampling in the Bayesian Lasso conditioned on $\mathbf{y}$ and $\mathbf{Z}$; (b) drawn directly from the Bayesian Lasso prior in equations (7.11b) to (7.11e); and (c) both sets of variables, restricted to the range -200 to -130. In both cases, the joint probability of the test data a set of samples is calculated using equation (7.12) integrated over β_0 with $\mathbf{y}$ and $\mathbf{Z}$ drawn from the test model specified in equation (7.14).

Of the 10 000 variables drawn from the prior, none match the maximum (or even the mode) log likelihood of the sampled variables conditioned on the test data. This shows how unlikely it is to draw suitable model variables directly from the prior.

7.7.4 1st pass grouping algorithm tests

This section presents results from the 1st pass grouping applied to simulated data. Results on real data are presented in the next section.

The simulated data for this section were generated using the following process, similar to the one used in equation (7.14) to test the Bayesian Lasso:

$$p = 10 \qquad \text{(number of predictors)} \tag{7.31a}$$

$$M = 30 \qquad \text{(number of individuals)} \tag{7.31b}$$

$$N = 50 \qquad \text{(number of data points per individual)} \tag{7.31c}$$

Three sets of model coefficients were randomly chosen for individuals 1, $M/2$ and M

$$\tilde{\boldsymbol{\beta}}_1 \,;\, p \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p) \qquad \text{(“true” model coefficients for individual 1)} \tag{7.31d}$$

$$\tilde{\boldsymbol{\beta}}_{M/2} \,;\, p \sim \mathcal{N}_p(\boldsymbol{\beta}_1, 0.5\mathbf{I}_p) \qquad \text{(model coefficients for individual } M/2) \tag{7.31e}$$

$$\tilde{\boldsymbol{\beta}}_M \,;\, p \sim \mathcal{N}_p(\boldsymbol{\beta}_1, 0.5\mathbf{I}_p) \qquad \text{(model coefficients for individual } M) \tag{7.31f}$$

from which the other model coefficients are linearly interpolated.

The remainder of the model is set up as follows:

$$\boldsymbol{\beta}_i \,|\, \tilde{\boldsymbol{\beta}}_i \sim \mathcal{N}_p(\tilde{\boldsymbol{\beta}}_i, 0.01\mathbf{I}_p) \;\; \forall \;\; 1 \leq i \leq M \;\; \text{(coefficients for individual } i) \tag{7.31g}$$

$$\mathbf{z}_{i,j} \,;\, p \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p) \qquad \forall \;\; 1 \leq i \leq M, \;\; 1 \leq j \leq N$$
$$\text{(predictors for data point } j \text{ of individual } i) \tag{7.31h}$$

$$y_{i,j} \,|\, \mathbf{z}_{i,j}, \boldsymbol{\beta}_i \sim \mathcal{N}(\mathbf{z}_{i,j}{}^\top \boldsymbol{\beta}_i, 0.2^2) \;\; \forall \;\; 1 \leq i \leq M, \;\; 1 \leq j \leq N$$
$$\text{(data labels for data point } j \text{ of individual } i) \tag{7.31i}$$

Thus M individuals are created, where each individual i has a “true” set of model coefficients, $\tilde{\boldsymbol{\beta}}_i$. The model coefficients for the first individual are randomly allocated; the coefficients for individuals $M/2$ and M are allocated from $\tilde{\boldsymbol{\beta}}_1$ with added noise; and the coefficients for the remaining individuals are linearly interpolated between the three previously allocated coefficients. A small amount of noise is then added to each of the “true” coefficients to form the final coefficients $\boldsymbol{\beta}_i$ for each individual.

The 10 coefficients for each of the 30 individuals are shown in figure 7.15 with each coefficient in a different colour. The general trend of the coefficients between the three originally allocated individuals (individuals 1; 15; and 30) can be seen, but although there are clear differences between the different individuals, no groups are obvious.

The results from running the 1st pass grouping on the data generated using the coefficients in figure 7.15 are shown in figure 7.16(a). Each individual is shown as a column, and the rows are the iterations of Gibbs sampling. The colour in each iteration indicates the allocated group number. All individuals start in group 1. Individual 30 is then allocated to a new group, which quickly expands to include individuals 22 and 24–29. A similar splitting of group 1 is seen with individuals 1–9 on the left of the figure. After about 100 iterations, group 3 on the left is further split into two separate groups. Finally, after about 220 iterations, the individuals on the border between groups 1 and 2 form a further new group. This demonstrates how the model splits the total set of individuals into groups that appear most similar.

Figure 7.16(b) shows the model without the variation described in section 7.7.3.3, where no groups are found.

The form of the test data generated through equations (7.31a) to (7.31i) does not elucidate the "correct" number of groups to find. The three main coefficient values generated in equations (7.31d) to (7.31e) indicates that three groups would be reasonable, but linear interpolation of the remaining coefficients between these points means that the boundaries between the groups is fuzzy. The five groups found in figure 7.16(a) should therefore not be seen as a correct grouping, rather a demonstration of the functionality of the model, dependent on the parameters used, in this case $\alpha_G = 1$. Indeed by setting $\alpha_G = 0.001$, three groups are found, and by setting $\alpha_G = 100$, seven groups are found. The question of determining the "optimal" number of groups will be explored in greater detail when tuning the model parameters to the geolocation data in section 7.10.2.

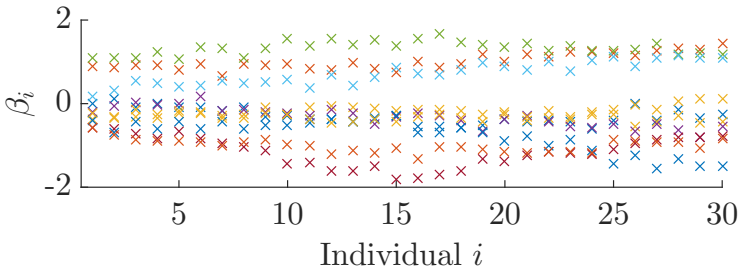
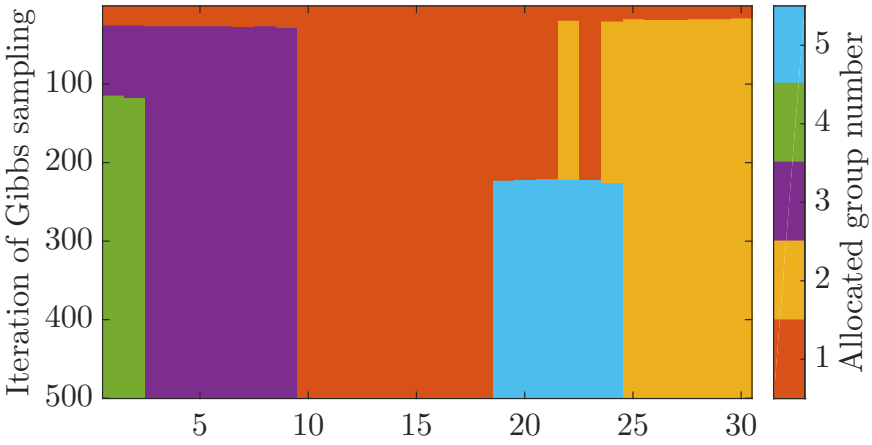
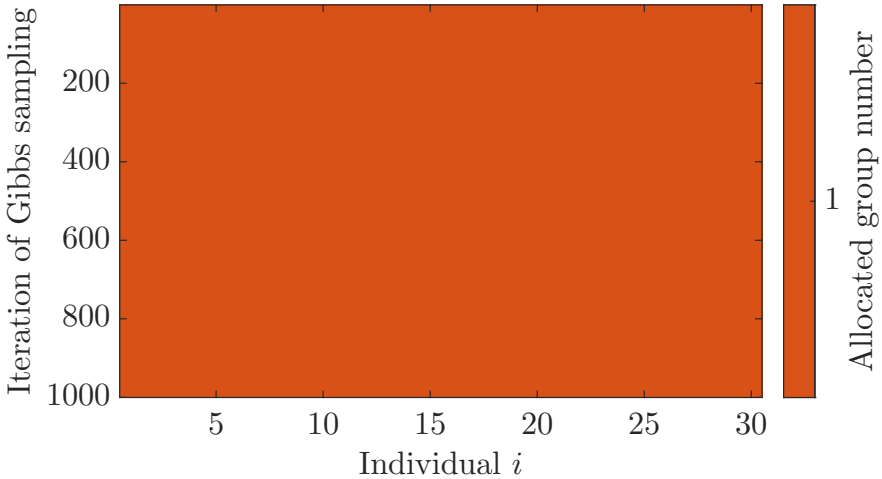


Figure 7.15: Coefficients of test individuals in simulated data used to generate results in figure 7.16. Each of the 30 individuals has 10 simulated model coefficients.



(a) Test results with the sampler variation in section 7.7.3.3 to improve mixing of groups.



(b) Test results using standard sampler (only one group was sampled after 1000 iterations).

Figure 7.16: Results from running the grouped Bayesian Lasso on simulated data using coefficients shown in figure 7.15. Results are shown using (a) the sampler with the variation described in section 7.7.3.3; and (b) using the standard sampling algorithm. In both cases, individuals are shown in columns with the colour in each iteration (on the rows) indicating the group that the individual is allocated to. With the sampler variation, the individuals are gradually split into distinct groups.

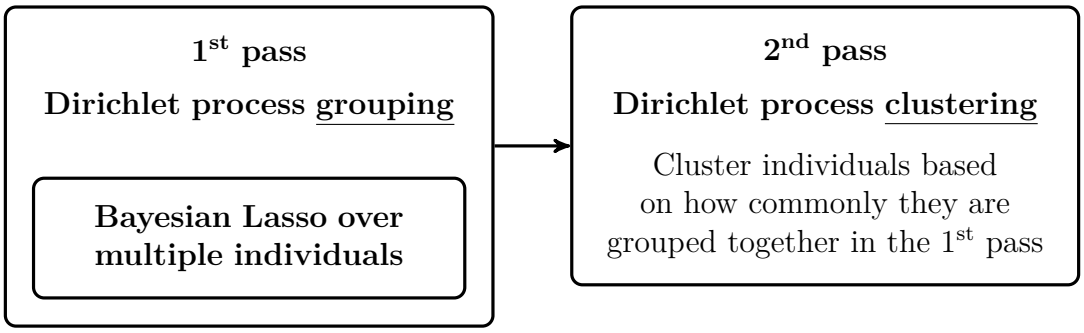


Figure 7.17: Simplified overview of the grouped Bayesian Lasso, adapted from figure 7.6 on page 191. Results from both passes are shown in figure 7.18 on pages 220 and 221. The results in figures 7.18(a) and (b) on page 220 show the raw unsorted output after running the 1st pass grouping of individuals. The 1st pass does not converge into static groups, but there appears to be structure to the results nonetheless. Figures 7.18(c) and (d) on page 221 show the individuals sorted into the clusters found by the 2nd pass clustering algorithm. This reveals the structure in the grouping results and demonstrates the need for the two-pass process.

7.8 Results from 1st pass grouping of individuals

The previous section demonstrated results from running the 1st pass grouping algorithm on simulated data. This section presents results from running the 1st pass grouping algorithm on the geolocation-derived feature data. The aim is to group individuals based on their self-reported QIDS scores and features extracted from the recorded geolocation data. It also demonstrates the need for the 2nd pass clustering algorithm to find structure in the groupings found in the 1st pass algorithm. The 2nd pass clustering algorithm will be fully described in the next section. As a recap, figure 7.17 above shows a simplified version of the two-pass architecture shown in figure 7.6 and described fully in section 7.6. All the parts of the 1st pass grouping of individuals based on their raw feature data have been described in the previous section.

The 1st pass grouping was run using only the 10 main features calculated on full calendar weeks of recorded data. The features were filtered and smoothed as described in sections 6.1.2 and 6.1.3, and the 59 participants who had more than 5 weeks of filtered data were included in the analysis. All participant cohorts (HC; BD; and BPD) were included in this analysis.

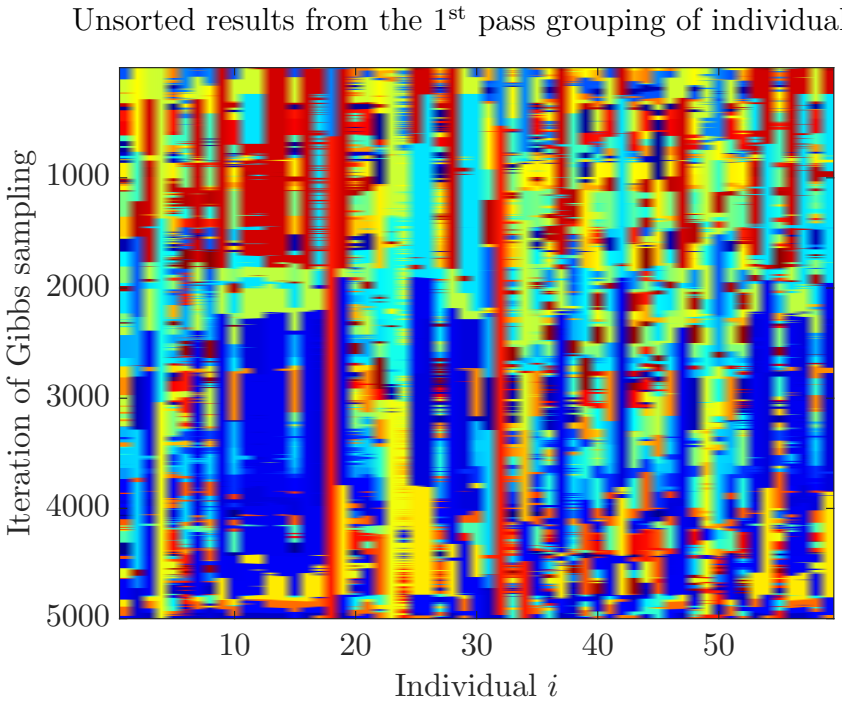
For the purposes of this section, the 1st pass grouping was run with Lasso regularisation parameter $\lambda = 1$ and Dirichlet process concentration parameter $\alpha_G = 0.0001$. Parameter tuning will be discussed in section 7.10. The model was sampled through 5000 iterations of Gibbs sampling using the sampler described in sections 7.7.3.2 and 7.7.3.3, and the first 100 samples were discarded as burn-in.

Figure 7.18 across the next two pages shows an overview of the groups sampled from the model. Figures 7.18(a) and (b) on the next page show the unsorted results from the 1st pass grouping, while figures 7.18(c) and (d) on page 221 show the results after running the 2nd pass clustering algorithm.

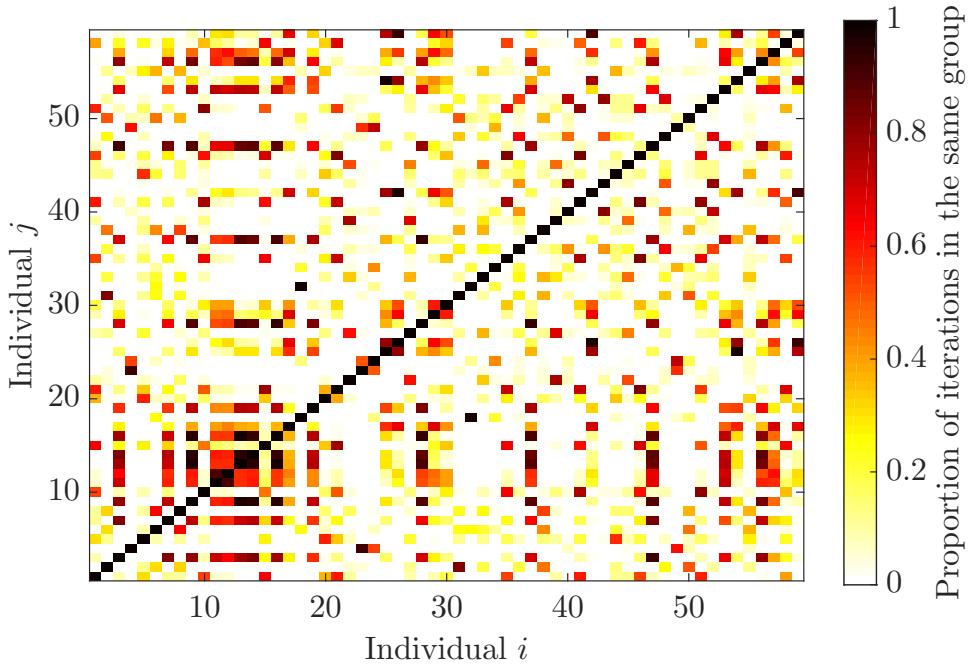
Figure 7.18(a) shows the group allocations for all individuals through all 5000 iterations in the same format used to demonstrate the results on the simulated data in figure 7.16. The group that an individual is allocated to in each iteration is identified by colour, and colours repeat as new groups are created through the sampling of the Dirichlet process.

Ideally the group allocations of individuals would converge to a stable set in the same way that the model did on simulated data in figure 7.16. In practice, with the noisy geolocation-derived features it can be seen that the model does not appear to converge to a stable set of groups. However, there are pairs of individuals that appear to transition through similar sequences of groups, for example individuals 18 and 32. To explore this is greater detail, figure 7.18(b) shows the proportion of iterations when each pair of individuals i and j are sampled in the same group, with darker colours indicating higher proportions of time together. In this plot it is clearer that there are certain individual pairs that are commonly in the same groups. It is also clear that there are many individuals that are very rarely in the same group. Figures 7.18(a) and (b) indicate that there also appear to be larger blocks of individuals that commonly occur together. These individuals will tend to transition through similar colours in figure 7.18(a), and will have darker colours at their intersection in figure 7.18(b).

If the individuals in the columns of figures 7.18(a) and (b) could be rearranged so that they were in clusters of individuals that commonly occur together, the clusters should show similar transitions through the different group allocations indicated by



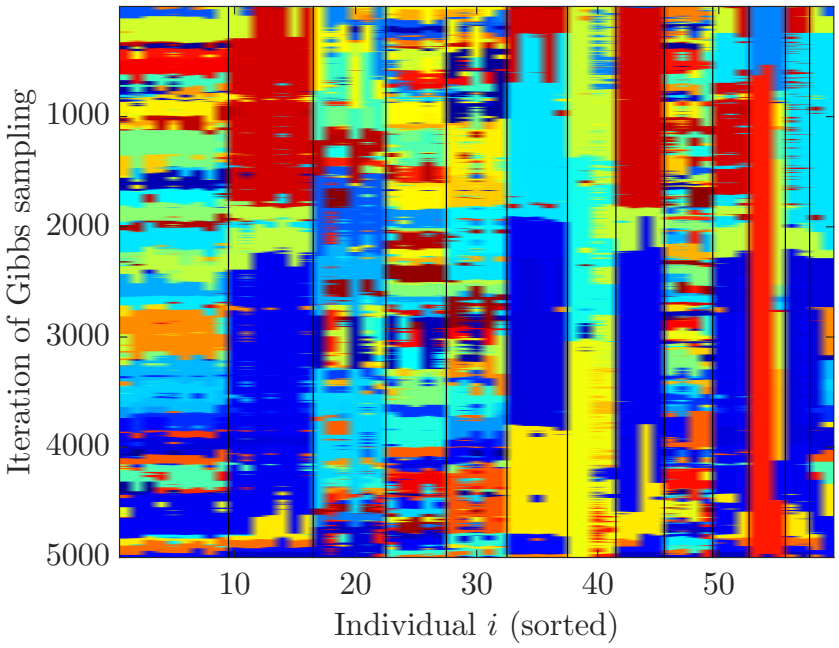
(a) Group allocations of individuals sampled from the grouped Bayesian Lasso.



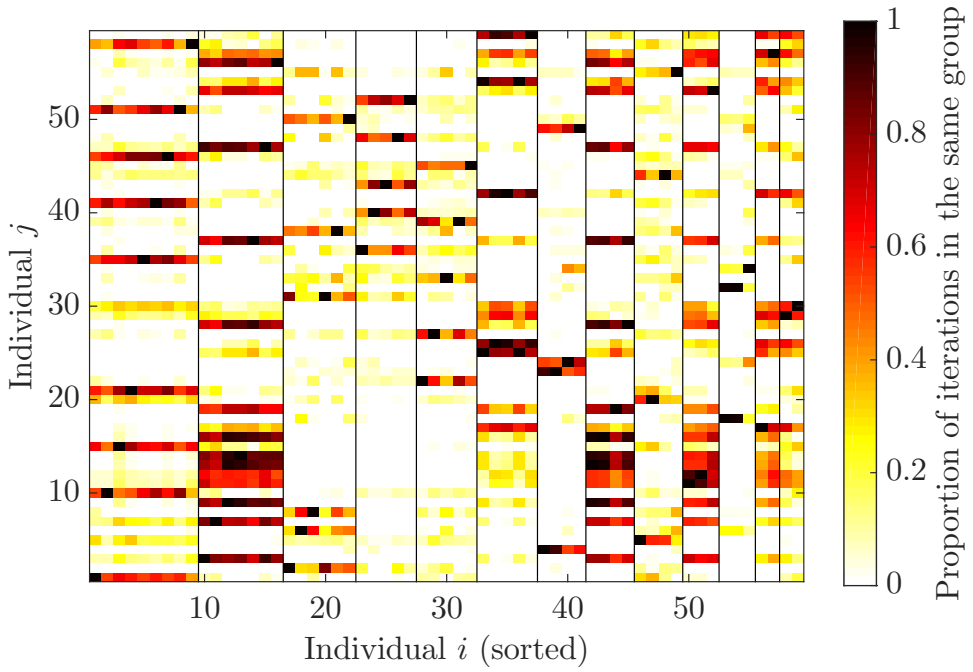
(b) Proportion of iterations where two individuals i and j are in the same group.

Figure 7.18: Results from running the grouped Bayesian Lasso for 5000 iterations of Gibbs sampling (described in sections 7.7.3.2 and 7.7.3.3) on the 10 main features extracted from the geolocation data recorded from 59 individuals. Figures (a) and (b) on this page show the unsorted output after running the 1st pass grouping of individuals. Figures (c) and (d) on the next page show the individuals sorted into the clusters found by the 2nd pass clustering algorithm. (continued on the next page)

Individuals sorted into the clusters found in the 2nd pass clustering algorithm



(c) Group allocations of individuals sampled from the grouped Bayesian Lasso.



(d) Proportion of iterations where two individuals i and j are in the same group.

Figure 7.18 (cont.): The top plots on both pages, (a) and (c), show the group allocations in each iteration for the 59 individuals included in the test. The colour represents the group, although colours repeat as new groups are created. The lower plots, (b) and (d), show the proportion of iterations in which any pair of individuals i and j were grouped together. The results on this page show that the 2nd pass clustering is required to expose the structure in the results from the 1st pass grouping.

the colours. This uncovers the clusters shown in figures 7.18(c) and (d), which show the same data as figures 7.18(a) and (b) respectively, but with individuals on the horizontal axis sorted into clusters of individuals that are commonly grouped together, separated by vertical lines.

From the clustering of individuals shown in figures 7.18(c) and (d) it is clear that there are sets of individuals that commonly occur in the same groups in the 1st pass grouping, even if they do not converge to stable groups. It can therefore be hypothesised that these clusters of individuals have similar relationships between their geolocation features and depression levels.

Formal convergence-based stopping criteria were not defined in the inference process of the 1st pass grouping algorithm. However, the clear patterns in the group allocations shown in figure 7.18(d), and the strength of the blocks within the groups in the similarity matrix in figure 7.18(d) indicate that the 5000 iterations used was sufficient for convergence. In practice, it is likely that fewer iterations would be required, but analysis of convergence is left as future work.

The next section details the 2nd pass clustering method that finds the clusters in figures 7.18(c) and (d)based the proportions shown in figure 7.18(b).. The subsequent section covers optimisation of parameters in the 1st pass grouping and performance of depression regression based on clustered individuals.

7.9 2nd pass clustering of similar individuals

This section describes the 2nd pass clustering of individuals based on the proportion of time that they are allocated to the same groups in the 1st pass grouping algorithm. The 2nd pass clustering process uses the proportion of iterations where individuals are in the same group in the 1st pass grouping (shown in figure 7.18(b) and referred to as the *similarity matrix*) to determine the clusterings shown in figures 7.18(c) and (d). An overview of the 2nd pass clustering process is shown in figure 7.19.

The 2nd pass clustering is performed through three main steps, described in the following subsections. The core part of the clustering is the Dirichlet process model

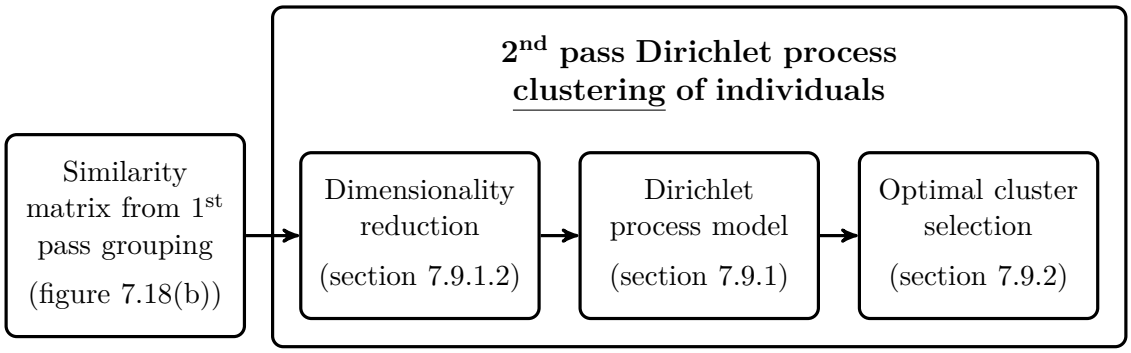


Figure 7.19: Overview of the 2nd pass clustering in the grouped Bayesian Lasso.

defined in section 7.9.1 and optimised in section 7.9.1.4. To reduce the computational complexity of the Dirichlet process model, dimensionality reduction is first performed on the similarity matrix using principal component analysis as described in section 7.9.1.2. Finally, the Dirichlet process samples many potential clusterings, and a method is needed to extract the "best" clusterings. This is discussed in greater detail in section 7.9.2.

7.9.1 Dirichlet process clustering of individuals

The main purpose of the 2nd pass clustering of individuals is to find clusters of individuals that commonly occur together in the 1st pass grouping. Due to the unknown number of clusters in the data, a Dirichlet process model that clusters individuals based on the similarity matrix forms the core element of this pass.

The Dirichlet process model used for the 2nd pass clustering is described in this section. Section 7.9.1.1 describes the core model definition used in the clustering process, and inference of the model is described in section 7.9.1.3. Due to the computational complexity of running the model on the full similarity matrix, dimensionality reduction is first performed as described in section 7.9.1.2. Finally, optimisation of the model parameters is presented in section 7.9.1.4.

7.9.1.1 Model definition

The Dirichlet process model used for the 2nd pass clustering is a simple hierarchical model with Gaussian observations, defined as follows:

$$\mathbf{y}_i \,\Big|\, \boldsymbol{\mu}^{(g)}\,;\, \sigma_c^2 \;\sim\; \mathcal{N}_k(\boldsymbol{\mu}^{(g)}, \boldsymbol{\Sigma}_C), \quad \boldsymbol{\Sigma}_C = \sigma_c^2 \mathbf{I}_k \tag{7.32a}$$

$$\boldsymbol{\mu}^{(g)} \,\Big|\, G \;\sim\; G \tag{7.32b}$$

$$G \,\Big|\, G_0\,;\, \alpha \;\sim\; \mathrm{DP}(\alpha, G_0) \tag{7.32c}$$

where G_0 is defined as a distribution over $\boldsymbol{\mu}$ such that

$$G_0^{\boldsymbol{\mu}}\,;\, \boldsymbol{\Sigma}_D \;=\; \mathcal{N}_k(\mathbf{0}_k, \boldsymbol{\Sigma}_D), \tag{7.32d}$$

and $\mathbf{y}_i$ is a vector of length k defining the similarity of individual i to the other individuals. The full similarity matrix is therefore given by $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M]^\top$. Where the model is run on the full similarity matrix, k is the total number of individuals M. To improve the computational performance of the model, the dimensionality of the similarity matrix can be reduced using principal component analysis as discussed in the next section. In this case, k is the number of principle components included in the model.

The model cluster centres are given in $\boldsymbol{\mu}^{(g)}$, which are drawn from a zero-centred multivariate normal distribution with covariance $\boldsymbol{\Sigma}_D$. For simplicity, $\boldsymbol{\Sigma}_D$ was set to the covariance of the similarity matrix $\mathbf{Y}$,

$$\boldsymbol{\Sigma}_D = \mathrm{cov}(\mathbf{Y}). \tag{7.33}$$

When dimensionality reduction is used on the similarity matrix, $\boldsymbol{\Sigma}_D$ is, from the definition of PCA, a diagonal matrix with the variance of each principle component on the diagonal.

This leaves two free parameters: α controlling the concentration of the Dirichlet process; and $\boldsymbol{\Sigma}_C = \sigma_c^2 \mathbf{I}_k$ controlling the size of each cluster. Tuning of these parameters will be discussed in section 7.9.1.4.

7.9.1.2 Dimensionality reduction of the similarity matrix

If there are M individuals included in the model, the similarity matrix will clearly be $M \times M$. For the Dirichlet process-based clustering method, described in the previous section, it is computationally beneficial to reduce this dimensionality to $N \times k$, where $k < M$. Reducing the dimensionality may also improve performance by reducing the amount of noise in the data.

One commonly used method of dimensionality reduction is principal component analysis (PCA), developed independently by Pearson (1901) and Hotelling (1933). PCA uses an orthogonal linear transform of the M-dimensional feature space to find orthogonal components ranked by the highest variance where the first (principal) component contains the highest variance. PCA can be thought of as an $(M-1)$-dimensional rotation about the mean of the feature space so that the dimension with the highest variance ends up along the x-axis, the second highest along the y-axis and so on. The resulting components (dimensions) may contain elements from all input features, so PCA transforms the input features into a new coordinate space where the highest ranking principal components contain the highest variance. Often the majority of the variance of the feature space is contained in the first k components where $k \ll M$.

More formally, the aim of PCA is to de-correlate the feature matrix $\mathbf{Z}$ (an $M \times p$ matrix where p is the number of features and M is the number of samples) to form a de-correlated feature matrix $\mathbf{Y}$ (also an $M \times p$ matrix) by multiplying it with a $p \times p$ de-correlation matrix $\mathbf{W}$ such that

$$\mathbf{ZW} = \mathbf{Y} \tag{7.34}$$

constrained such that

$$\operatorname{cov}(\mathbf{ZW}) = \operatorname{diag}(\mathbf{w}) \tag{7.35}$$

where $\operatorname{cov}(\mathbf{ZW})$ is the covariance matrix of $\mathbf{ZW}$; and $\mathbf{w}$ is a scalar vector of (principal component) variances.

There are two main methods used to solve the PCA equations: using the eigendecomposition of cov($\mathbf{Z}$); or using the singular value decomposition (SVD) of $\mathbf{Z}$. Both of these methods are derived and described in detail in section E.4 of appendix E (see also Shlens, 2014, for an overview of PCA).

To cluster similar individuals, PCA was performed on the similarity matrix in figure 7.18(b) and the first principal components that together made up 95 % of the variance in the data were retained for clustering. This resulted in the full 59-dimensional similarity matrix being reduced to 9 principal components for the data shown in figure 7.18(b). The 9-component latent similarity matrix computed from the original similarity matrix in figure 7.18(d) is shown in figure 7.20(a). Additionally, the recovered similarity matrix after dimensionality reduction is shown in figure 7.20(b), which clearly shows how the noise in the original matrix is reduced, but the main characteristics are sill present. This indicates that the latent similarity matrix contains the required similarity information for grouping.

Other methods of dimensionality reduction are also available, including traditional methods such as linear discriminant analysis as well as more modern methods such as non-negative matrix factorisation, discussed in the context of recommender systems in section 7.2. For the purpose used here, PCA provides a simple method of linear dimensionality reduction that is demonstrated in figure 7.20 and later to provide the expected results. It is unclear if more advanced methods would provide improved performance.

$\longrightarrow$

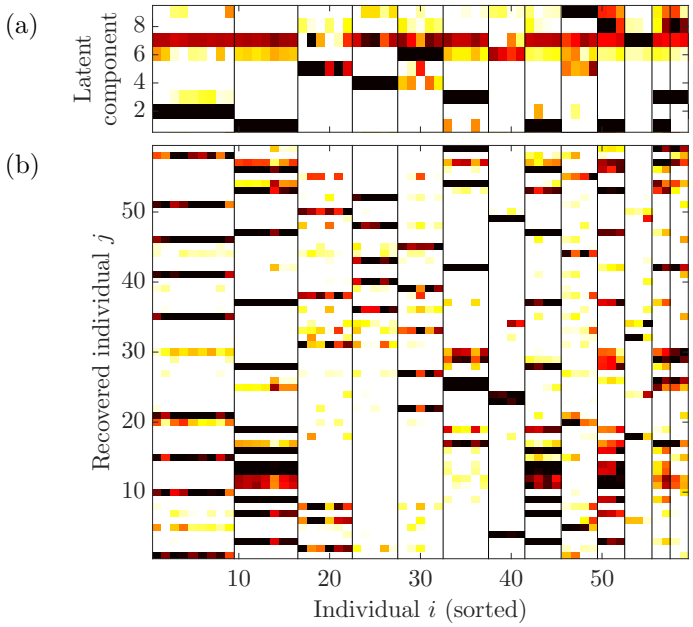


Figure 7.20: Similarity matrix in figure 7.18(d) after dimensionality reduction, showing (a) the 9-dimensional latent representation; and (b) the recovered similarity matrix after dimensionality reduction. The recovered matrix has reduced the level of noise while retaining the main elements of the original matrix, demonstrating that the latent representation contains the required similarity information for grouping of individuals.

7.9.1.3 Clustering model sampling

The model defined in equation (7.32) above can be sampled in a standard implementation of Radford Neal's Algorithm 8, described in section 3.3.1. The joint distribution for the n individuals in each cluster is defined as

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_n, \boldsymbol{\mu}\,;\, \sigma_c^2, \boldsymbol{\Sigma}_D) = p(\mathbf{y}_1 \,|\, \boldsymbol{\mu}\,;\, \sigma_c^2) \times \quad \ldots \quad \times p(\mathbf{y}_n \,|\, \boldsymbol{\mu}\,;\, \sigma_c^2) \times p(\boldsymbol{\mu}\,;\, \boldsymbol{\Sigma}_D) \quad (7.36)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}_C|}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}_C^{-1} (\mathbf{y}_i - \boldsymbol{\mu})\right\}$$

$$\times \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}_D|}} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}_D^{-1} \boldsymbol{\mu}\right\}, \quad \boldsymbol{\Sigma}_C = \sigma_c^2 \mathbf{I}_k, \quad (7.37)$$

which is conjugate, meaning that $\boldsymbol{\mu}$ can be sampled from

$$\boldsymbol{\mu} \,\Big|\, \mathbf{y}_1, \ldots, \mathbf{y}_n\,;\, \sigma_c^2, \boldsymbol{\Sigma}_D \;\sim\; \mathcal{N}_k\left(n\boldsymbol{\Sigma}_C^{-1}\bar{\mathbf{y}}(\boldsymbol{\Sigma}_D^{-1} + n\boldsymbol{\Sigma}_C^{-1})^{-1}, (\boldsymbol{\Sigma}_D^{-1} + n\boldsymbol{\Sigma}_C^{-1})^{-1}\right) \quad (7.38)$$

where $\bar{\mathbf{y}}$ is the sample mean of the data currently assigned to the cluster.

In order to increase the number of combinations of cluster allocations sampled from the posterior, a small modification was made to the sampling of parameters to add a small, and decreasing, amount of multivariate Gaussian noise. Instead of sampling directly from equation (7.38), samples for $\boldsymbol{\mu}$ were instead sampled from

$$\boldsymbol{\mu}^* \,\Big|\, \boldsymbol{\mu}\,;\, \epsilon, t \;\sim\; \mathcal{N}_k\left(\boldsymbol{\mu}, \frac{\epsilon}{t}\boldsymbol{I}_k\right) \quad (7.39)$$

where ϵ is a parameter controlling the level of noise and t is the number of the current iteration. This adaptation of the basic sampling algorithm allows exploration of a larger range of the posterior distribution in the initial sweeps, eventually converging in the same way as the regular implementation of Algorithm 8. It is similar in principle to optimisation by simulated annealing, first described by Kirkpatrick et al. (1983), in that it initially allows transitions to worse states, eventually converging towards the optimum.

The effect of varying ϵ is shown in figure 7.21 where the 2$^{\text{nd}}$ pass clustering algorithm was run for 200 iterations, 20 different times (hence a total of 4000 clusters were sampled). As ϵ is increased, the total number of unique clusters increases as

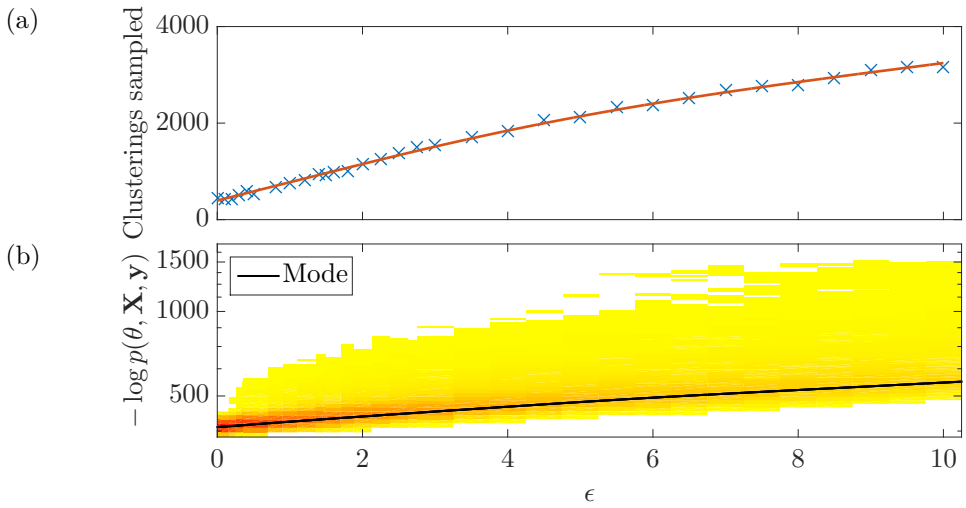


Figure 7.21: Effect on (a) the total number of unique clusterings sampled; and (b) the distribution of negative log likelihoods of the sampled clusters from varying ϵ in equation (7.39).

expected (shown in figure 7.21(a)), as does the range of likelihoods of the sampled clusters (shown in figure 7.21(b)). Figure 7.21(b) also shows that the mode likelihood of the sampled clusters decreases. This indicates that the adapted algorithm is indeed sampling a wider range of the posterior as expected, albeit with reduced mode likelihood. As discussed in greater detail in the next sections, sampling a wider range of groupings may be more important than sampling the groupings with the absolute optimal likelihood. For this reason, the results that follow use a value of $\epsilon = 2$, which provides a reasonable trade-off between increasing the number of clusters sampled without excessively affecting the mode log likelihood of the sampled clusters.

Formal convergence-based stopping criteria were not defined in the inference process of the 2$^{\text{nd}}$ pass clustering algorithm. However, the number of unique clusterings sampled across the 4000 iterations described above when testing the performance of the algorithm with $\epsilon = 2$ in figure 7.21 are shown in figure 7.22 on the following page. It can be seen that within each of the 20 runs, shown as the grey lines, many unique clusterings are samples in the first 50 iterations (almost one unique clustering per iteration initially) before starting to level off. The overall number of unique clusterings sampled, shown as the red line, has a similar profile but much more clearly levels off

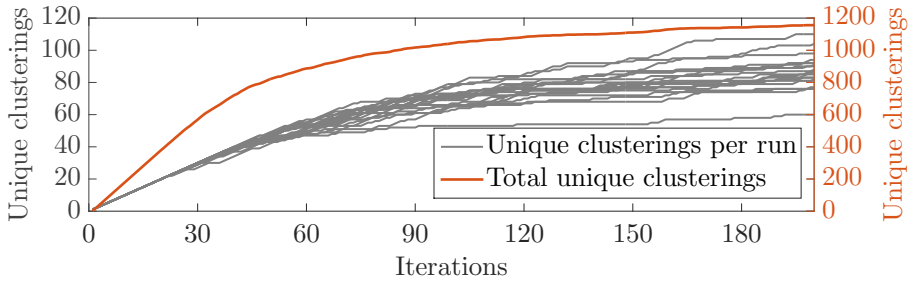


Figure 7.22: Number of unique clusterings sampled during inference in the 2nd pass clustering algorithm when testing model performance with $\epsilon = 2$ in figure 7.21. As described in the text, the clustering algorithm was run for 200 iterations, 20 different times. The grey lines indicate the number of unique clusterings sampled in each iteration of each run, and the red line indicates the total number of unique clusterings across all runs.

after all 4000 iterations (200 iterations in each of the 20 runs) as the same clusterings are sampled in multiple runs. It is therefore unlikely that more runs are required for optimal performance. This is further supported by figure 7.24 on page 235, described in section 7.9.2, which shows how the log likelihoods of the sampled clusterings has also levelled off after 200 iterations in each run.

7.9.1.4 Clustering model parameter optimisation

As discussed above, there are two free parameters in the clustering model: α controlling the concentration of the Dirichlet process; and $\mathbf{\Sigma}_C = \sigma_c^2 \mathbf{I}_k$ controlling the size of each cluster. This section describes the optimisation of these two parameters based on the results from the 1st pass grouping of individuals based on their geolocation-derived features shown in figures 7.18(a) and (b).

A grid search was performed over σ_c^2 between 10^{-3} and 10^1, and α between 10^{-6} and 10^3, the results of which are shown in figure 7.23 across pages 232 and 233. The clustering algorithm described above was run 20 times for each parameter combination with 200 iterations in each run, giving a total of 4000 groups sampled for each parameter combination (although not necessarily 4000 unique groups). Figure 7.23(a) shows the mean negative log likelihood of all the clusterings sampled for each parameter combination, interpolated over the whole parameter space. Actual test results

are shown as black dots and the mesh is interpolated using biharmonic spline interpolation as described by Sandwell (1987). The mean negative log likelihood was chosen as an appropriate metric to optimise as it is data agnostic and well defined for the Dirichlet process model.

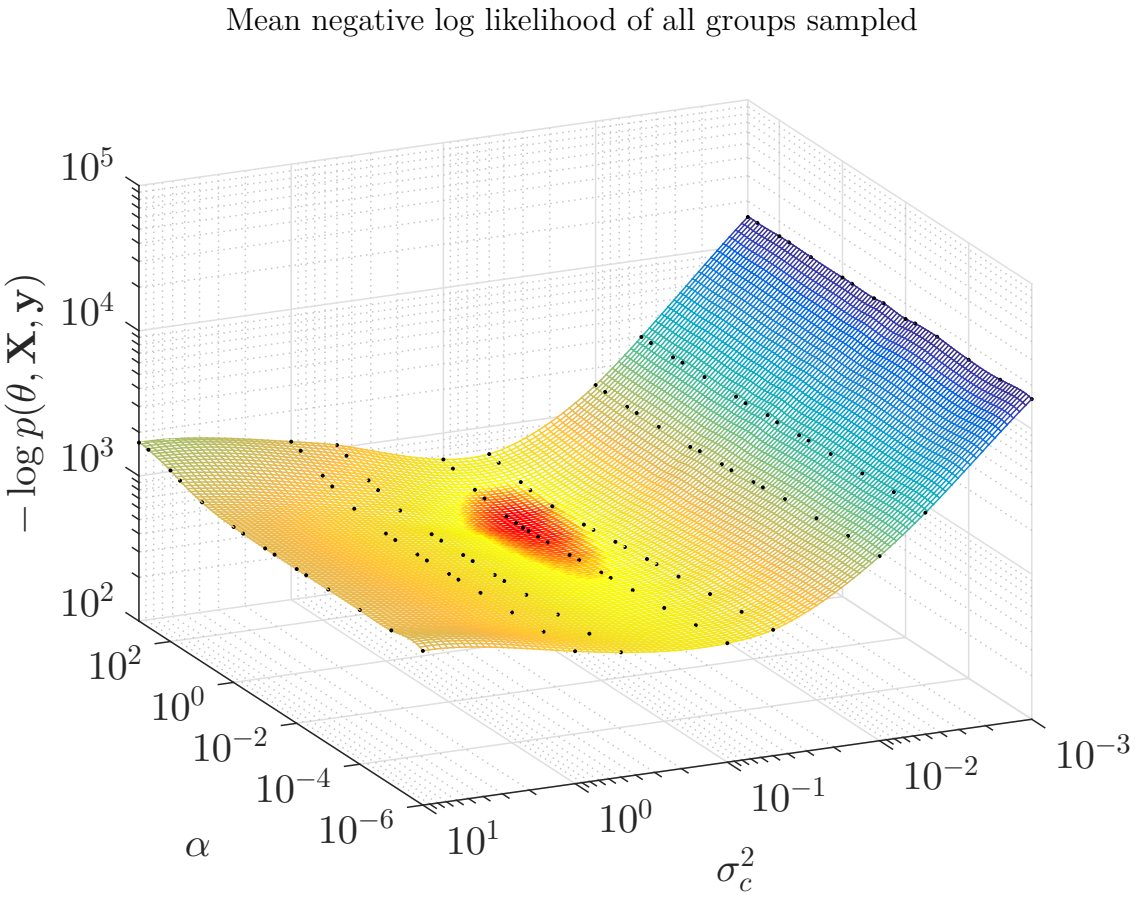
The global minimum mean negative log likelihood is in the dark red area shown in figure 7.23(a), which lies on the plane where $\sigma_c^2 = 10^{-1}$. A slice through the surface at $\sigma_c^2 = 10^{-1}$ is shown in figure 7.23(b), and the optimal point sampled is shown as the red cross, at $\alpha = 3$. Figure 7.23(b) also shows the mode negative log likelihood of the sampled clusters. Although there is more noise in the mode it is broadly the same as the mean, but shifted towards zeros, indicating the asymmetry of the distribution of log likelihoods.

The mean number of clusters in each sample is shown in figures 7.23(c) and (d), with figure 7.23(c) showing a surface over the same parameter space as figure 7.23(a) and figure 7.23(d) showing the same slice as figure 7.23(b). Figure 7.23(d) additionally shows the mode number of clusters in each sample. Similar to the mode of the negative log likelihoods of the sampled clusters, the mode number of clusters is also slightly lower than the mean in all cases, again indicating that the distribution of the number of clusters is skewed towards zero. This shows that at $\alpha = 3$, there are a mean of 14.21 clusters in each sample, and a mode of 13. This matches the number of clusters shown in figures 7.18(c) and (d), which appears to be reasonable.

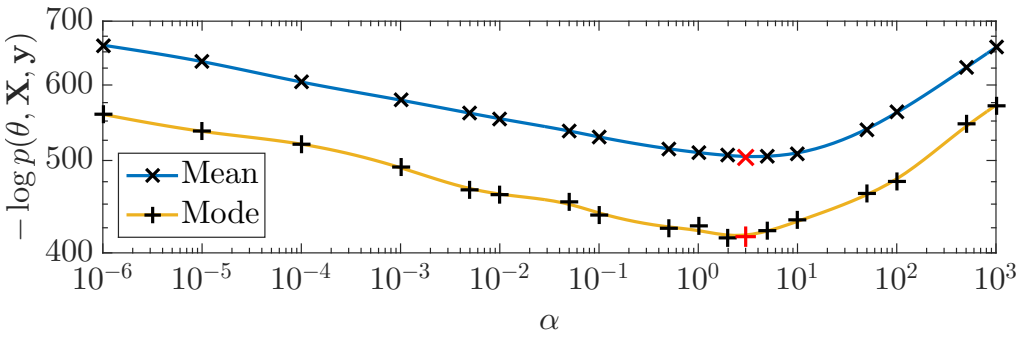
The optimal parameters used for the clustering of individuals in the 2^{nd} pass clustering are therefore chosen as $\sigma_c^2 = 10^{-1}$, and $\alpha = 3$.

7.9.2 Optimal cluster selection

When running the Dirichlet process clustering described in the previous section on the geolocation-derived features, the model would ideally converge to a static clustering of individuals. Additionally, running the model multiple times from different initialisations would ideally converge to the same static clusterings. In practice, neither of these properties hold. While the model does converge to relatively static clusterings, small changes and switches between clusters continue to be made, and running the



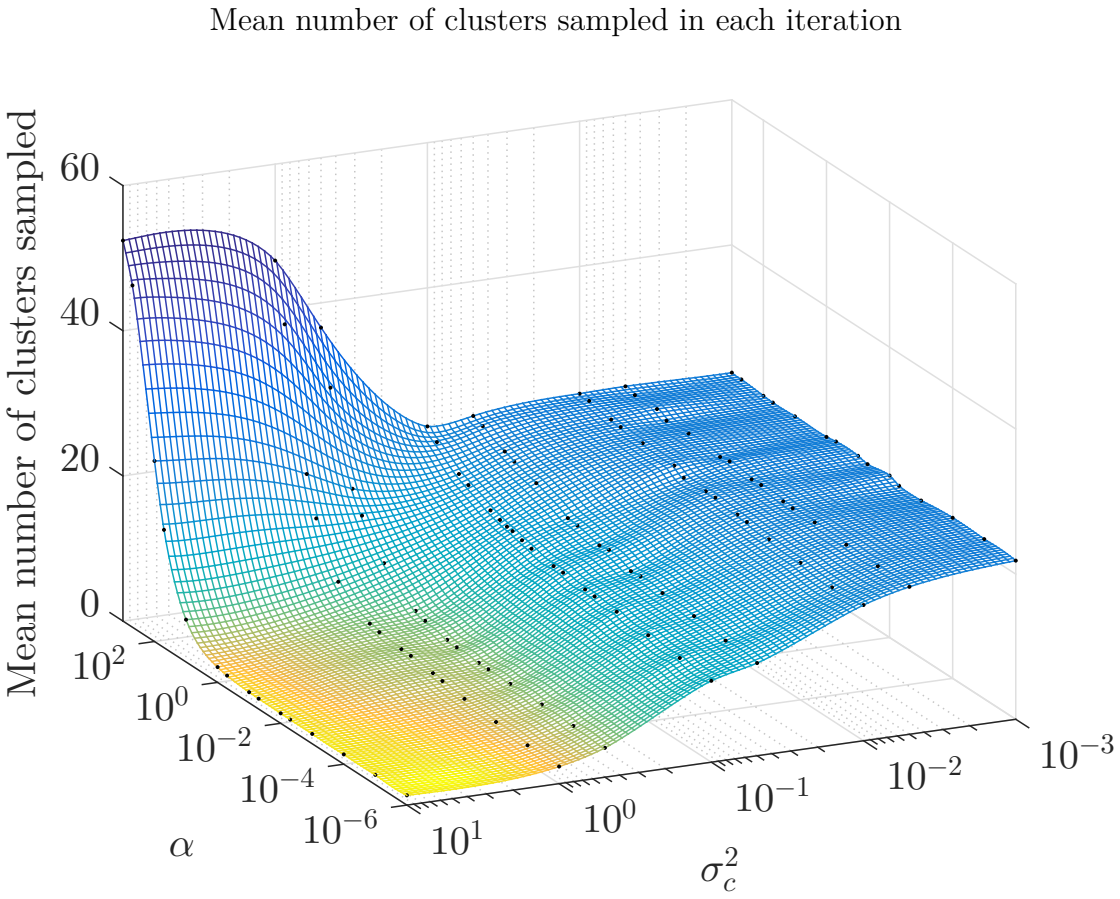
(a) The mean negative log likelihood of sampled clusterings over the entire parameter space.



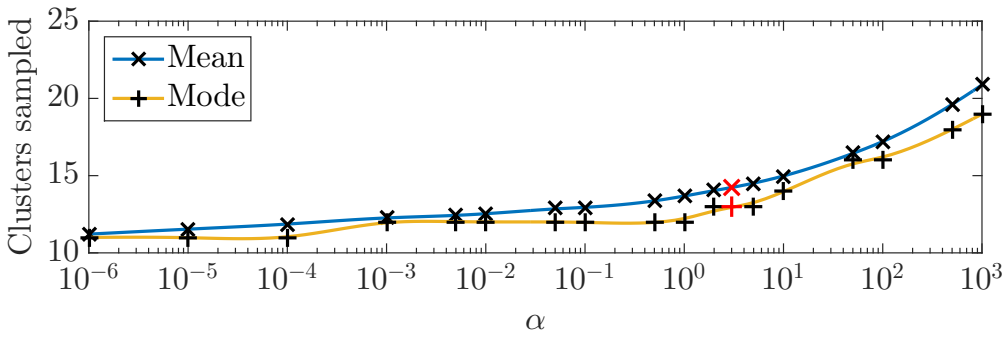
(b) Slice through surface in figure 7.23(a) at $\sigma_c^2 = 10^{-1}$, also showing the mode negative log likelihood.

Figure 7.23: Results of parameter optimisation in the clustering of similar individuals based on by the similarity of the groupings shown in figure 7.18(a) generated by running the 1st pass grouping algorithm on the geolocation-derived features. Results were generated by running the Dirichlet process-based clustering algorithm described in section 7.9.1.3 20 times with 200 iterations each (hence a total of 4000 samples were taken). Figures (a) and (b) on this page show the mean negative log likelihood of the sampled groupings, while figures (c) and (d) on the next page show the mean number of groups sampled in each of the sampled groupings.

(continued on the next page)



(c) Mean number of clusters in the sampled clusterings over the entire parameter space.



(d) Slice through surface in figure 7.23(c) at $\sigma_c^2 = 10^{-1}$, also showing the mode number of clusters.

Figure 7.23 (cont.): The α parameter was varied between 10^{-6} and 10^{3}, and the σ_c^2 parameter was varied between 10^{-3} and 10^{1}. The top figures, (a) and (c) show a mesh over the whole parameter space explored. The black dots show the actual tests performed, while the mesh is interpolated using biharmonic spline interpolation (see Sandwell, 1987). The lower figures, (b) and (d) show slices through the mesh at $\sigma_c = 10^{-1}$, which was found in figure 7.23(a) to generally have the lowest mean negative log likelihood for any given α. The red crosses in (b) and (d) show the minimum mean negative log likelihood, where a mean of 14.21 groups are sampled.

model several times tends to converge to different, although often similar, clusterings. Given the number of possible clusterings of the 59 individuals[3] and the noise in the similarity matrix in figure 7.18(b), this behaviour is not surprising.

The Dirichlet process clustering model does, however, provide a likely subset of the total cluster space given the available data. Choosing a single *best* clustering from this subset to apply to the data is still non-trivial, especially given that runs do not tend to converge and that different runs converge to different clusterings. This section describes a method to extract a single representative clustering, such as the one used in figures 7.18(c) and (d), from all the clusterings sampled.

The sampling algorithm described and optimised above (with parameters $\epsilon = 2$; $\sigma_c^2 = 10^{-1}$; and $\alpha = 3$) run 20 times with 200 iterations in each run, samples 101 ± 8.9 (mean $\pm$ SD) unique clusterings per run. Over all 4000 iterations, 1360 unique clusterings are sampled. The log likelihoods of the individual clusterings (calculated using equation (7.48)) sampled in each iteration are shown in figure 7.24, with the distribution over log likelihoods shown in the histogram on the right of the figure. It is clear that the sampler initially explores the less likely clusterings before converging towards more optimal clusterings at the mode of the posterior. In the histogram of likelihoods, the peak log likelihood is around -415, with a small tail of samples with greater likelihood and a large tail of samples with lower likelihood. This form implies a noisy posterior distribution. In a smooth posterior distribution, such as in a conjugate Gaussian model, samples from the posterior have a peak at the maximum *a posteriori* (MAP) estimate, with only a tail of samples with lower likelihood – in other words, in a smooth posterior, selecting the sample with the maximum posterior likelihood is likely to be very characteristic of the bulk of the posterior distribution.

[3] The number of potential clusterings grows exponentially as the number of data points to cluster increases. With 3 data points there are 5 possible clusterings; with 5 data points there are 52; with 9 data points there are 21 147; and with 10 data points there are already 115 975 potential clusterings. Linear extrapolation on a log scale provides a lower-bound of 10^{40} possible clusterings with the 59 data points included in this analysis.

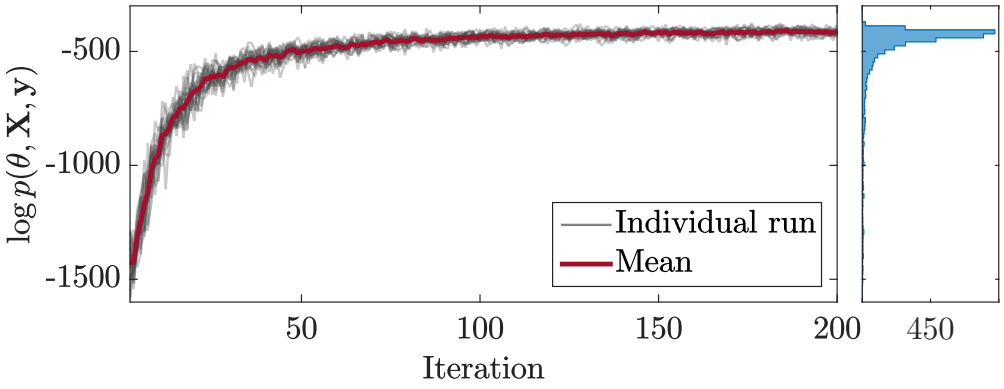


Figure 7.24: Log likelihoods of sampled clusterings from 20 runs of the Dirichlet process model described in the previous section with 200 iterations in each run. Each run is shown as a grey line and the mean likelihood is shown in red. The likelihood for each clustering is calculated using equation (7.48). The sampler starts off exploring less likely clusterings before increasing towards more likely splits. However, the sampler is still continuously making small changes in the group allocations with the mean still increasing, albeit slowly, after 200 iterations. The histogram of sampled likelihoods on the right shows that there is a strong peak around -415, but there is also a small tail of more likely clusterings found. Higher likelihood in the posterior does not necessarily mean a "better" clustering, and this section describes a method to extract a single clustering that is most representative of the posterior distribution.

For a distribution where few points are near the maximum likelihood, such as the samples of clusterings shown in figure 7.24, it is possible that the MAP estimate may not be characteristic of the overall distribution. It is likely that the bulk of the posterior is around the peak in the distribution of log likelihoods, and the few points with higher log likelihood may be outliers. This may imply a noisy distribution with many local modes.

Additionally, although the sampler clearly converges towards the bulk of the posterior as expected, the last 20 samples from each of the 20 runs (i.e. 400 samples from the posterior) contain 95 unique clusterings, demonstrating the lack of convergence. The MAP estimate from each run is different, and may not be similar.

Ultimately it is desirable to pick a single clustering that best characterises the posterior. Using the MAP estimate is a natural choice, but as discussed above it may not be characteristic of the posterior. Additionally, since the sampler does not converge towards the MAP estimate, it may be highly unpredictable in different runs

of the sampler. This section describes a method to select a characteristic clustering.

In the case of a continuous distribution, or a numerical discrete distribution, a suitable statistic value characterising the distribution may be obtained by calculating the expected value of the random variable, defined as a weighed sum over all possible values that the random variable can take, formally

$$\mathbb{E}(X) = \begin{cases} \int_{-\infty}^{\infty} xp(x)\,dx & \text{if } X \text{ is continuous} \\ \sum_{i=1}^{\infty} x_i p(x_i) & \text{if } X \text{ is discrete.} \end{cases} \tag{7.40}$$

In the case of having n samples from a numeric distribution (continuous or discrete), the expectation is defined as

$$\hat{\mathbb{E}}(X) = c\sum_{i=1}^{n} x_i p(x_i), \tag{7.41}$$

where $c = 1/\sum_{i=1}^{n} p(x_i)$ is a normalisation constant to ensure that $\sum_{i=1}^{n} p(x_i) = 1$. While the expectation of a random variable is the mean of a distribution, not a mode, it is nevertheless a useful summary statistic.

For an arbitrary distribution (such as a distribution over discrete clusterings), the expectation can be calculated using

$$\hat{\mathbb{E}}(X) = \underset{x_j}{\operatorname{argmin}} \left| \sum_{i=1}^{n} \Delta(x_i, x_j) p(x_i) \right| \tag{7.42}$$

where the function $\Delta(x_i, x_j)$ is a distance metric between pairs of points x_i and x_j. This is derived from equation (7.41) in section E.5 of appendix E. With a suitable distance metric $\Delta(x_i, x_j)$, equation (7.42) can be used to calculate an expected value of the sampled clusters. This expected cluster is characteristic of the bulk of the posterior distribution.

There are a number of information theoretic-based distance measures available that calculate a distance between two clusterings and which would be suitable for this purpose (Vinh et al., 2010, provide a review). Changing notation to match Vinh et al. (2010), if S is a set of N data items, then S can be partitioned into two set of non-overlapping subsets $\mathbf{U} = \{U_1, U_2, \ldots, U_R\}$ and $\mathbf{V} = \{V_1, V_2, \ldots, V_C\}$. In the

Table 7.3: Contingency table summarising the similarity of clusterings $\mathbf{U}$ and $\mathbf{V}$.

	V_1	V_2	$\cdots$	V_C	$a_i = \sum_j n_{ij}$
U_1	n_{11}	n_{12}	$\cdots$	n_{1C}	a_1
U_2	n_{21}	n_{22}	$\cdots$	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	a_R
$b_j = \sum_i n_{ij}$	b_1	b_2	$\cdots$	b_C	

$$n_{ij} = \left|U_i \bigcap V_j\right|, \quad \sum_{ij} n_{ij} = \sum_i a_i = \sum_j b_j = N$$

current applications, all the individuals are split into R clusters in clustering $\mathbf{U}$; and C clusters in clustering $\mathbf{V}$ – the aim is to find a metric comparing the similarity of the two clusterings.

The information shared between the two clusterings, $\mathbf{U}$ and $\mathbf{V}$, can be summarised in an $R \times C$ contingency table $M = \left[n_{ij}\right]_{j=1,\ldots,C}^{i=1,\ldots,R}$ describing the number of the same individuals assigned to each pair of clusters, U_i and V_j, where $n_{ij} = \left|U_i \bigcap V_j\right|$. The contingency table is shown in table 7.3 above.

Using the values in the contingency table shown in table 7.3, the following information-theoretic properties can be calculated:

$$H(\mathbf{U}, \mathbf{V}) = -\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}}{N} \tag{7.43}$$

$$I(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{a_i b_j / N^2}, \tag{7.44}$$

where $H(\mathbf{U}, \mathbf{V})$ is a standard joint entropy calculated using equation (5.4) (Shannon, 1948; Cover & Thomas, 2006; Meilă, 2007; Yao, 2003), and $I(\mathbf{U}, \mathbf{V})$ is a measure of the mutual information shared between $\mathbf{U}$ and $\mathbf{V}$ (Cover & Thomas, 2006; Meilă, 2007; Yao, 2003).

Several distance metrics based on entropy and mutual information have been proposed (for a review, see Vinh et al., 2010). One such metric, originally proposed by Kraskov et al. (2005) and closely related to the *variation of information* by Meilă

(2007), is defined as

$$\Delta(\mathbf{U}, \mathbf{V}) = 1 - \frac{I(\mathbf{U}, \mathbf{V})}{H(\mathbf{U}, \mathbf{V})}. \tag{7.45}$$

This is a normalised distance metric[4] that can be used in equation (7.42) to find an expected characteristic clustering out of all the sampled clusterings $\mathcal{G}$ as

$$\hat{\mathbb{E}}(\mathcal{G}) = \underset{\mathbf{U} \in \mathcal{G}}{\operatorname{argmin}} \left| \sum_{\mathbf{V} \in \mathcal{G}} \Delta(\mathbf{U}, \mathbf{V}) p(\mathbf{V}) \right| \tag{7.46}$$

where $\Delta(\mathbf{U}, \mathbf{V})$ is the distance between clusterings $\mathbf{U}$ and $\mathbf{V}$ as defined by equation (7.45); and $p(\mathbf{V})$ is the probability of clustering $\mathbf{V}$. In the present application, the probability of a given clustering $\mathbf{V} = \{V_1, V_2, \ldots, V_C\}$ is given by

$$p(\mathbf{V}) = p(\mathbf{V}, \boldsymbol{\mu} \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_n \,;\, \alpha, \sigma_c^2, \boldsymbol{\Sigma}_D) \tag{7.47}$$

$$= \underbrace{\alpha^C \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \prod_{c=1}^{C} \left[\Gamma(n_c)}_{\substack{\text{probability of cluster allocation from DP} \\ p(\mathbf{V}\,;\alpha) \text{ – see equation (3.54)}}} \cdot \underbrace{\mathcal{N}_k(\boldsymbol{\mu}^{(c)} \,;\, \mathbf{0}_k, \boldsymbol{\Sigma}_D) \cdot \prod_{j=1}^{n_c} \mathcal{N}_k(\boldsymbol{y}_j^{(c)} \,;\, \boldsymbol{\mu}^{(c)}, \sigma_c^2 \mathbf{I}_k) \right]}_{\substack{\text{joint probability of cluster variables} \\ p(\mathbf{y}_1^{(c)}, \ldots, \mathbf{y}_{n_c}^{(c)}, \boldsymbol{\mu}\,; \sigma_c^2, \boldsymbol{\Sigma}_D) \text{ – see equation (7.36)}}}. \tag{7.48}$$

Here n denotes the total number of individuals; n_c denotes the number of individuals allocated to group c; $\boldsymbol{\mu}^{(c)}$ is the cluster centre of group c in the similarity matrix; and $\boldsymbol{y}_j^{(c)}$ denotes the similarity matrix entries for the jth individual in group c.

Using equations (7.45) and (7.48) in equation (7.46), the expectation $\hat{\mathbb{E}}(\mathcal{G})$ can be calculated from the sampled clusterings. This returns a single clustering that is characteristic of the posterior distribution. This is more robust than using other metrics such as the MAP estimate, which can vary between runs. While the expectation can also vary depending on the available samples, due to the way that it is calculated it will always be representative of the bulk of the posterior, which the MAP estimate may not be.

[4] Equation (7.45) fills the essential criteria for a distance metric of (a) non-negativity: $\Delta(\mathbf{U}, \mathbf{V}) \geq 0$; (b) reflexivity: $\Delta(\mathbf{U}, \mathbf{V}) = \Delta(\mathbf{V}, \mathbf{U})$; (c) the identity of indiscernibles: $\Delta(\mathbf{U}, \mathbf{V}) = 0$ i.i.f $\mathbf{U} = \mathbf{V}$; and (d) the triangle inequality: $\Delta(\mathbf{U}, \mathbf{V}) \leq \Delta(\mathbf{U}, \mathbf{W}) + \Delta(\mathbf{W}, \mathbf{V})$ (see Deza & Deza, 2016, chapter 1 for details).

7.9.3 Results from 2nd pass clustering of individuals

Applying the 2nd pass clustering of individuals, using the sampled clusterings depicted in figure 7.24, returns the clustering shown previously in figures 7.18(c) and 7.18(d) as the expected, 'mean', clustering.

7.10 Personalised regression model parameter tuning

The previous sections have presented the two-pass grouped Bayesian Lasso framework for performing personalised regression. Section 7.7 described the 1st pass grouping of individuals, adapting the Bayesian Lasso model by Park & Casella (2008) to a hierarchical model over multiple individuals, grouped by the similarity of their regression model coefficients. Section 7.9 detailed the 2nd pass clustering of the output from the 1st pass grouping based on how often individuals were allocated to the same group, to find the final clusters of individuals. This section will demonstrate tuning of the model parameters on the geolocation data to produce the regression results presented previously in figure 7.2 on pages 178 and 179.

7.10.1 Test methodology

Section 7.11 describes the final results of running the grouping and clustering algorithms on the geolocation-derived features. Central to this is testing the performance of the model under different parameter combinations and choosing the optimal model. Section 7.10.1.1 describes the test methodology used to run the models on the test participants, and section 7.10.1.2 describes the main metrics used to evaluate performance and how each one is interpreted.

7.10.1.1 Grouping and clustering

The individuals in the full data set were first grouped by running the 1st pass grouping described in section 7.7 for 5000 iterations. The 1st pass grouping was run with a selection of parameter combinations, as described in the following sections. This results

in a set of raw groupings in each iteration, similar to that shown in figure 7.18(b) on page 220. These raw groupings were clustered by similarity using the 2nd pass clustering techniques described in section 7.9.

For this section, data from all participants in the AMoSS study (HC; BD; and BPD) who provided more than 5 calendar weeks of labelled data were included. Regression was performed only on the 10 main features derived from the geolocation data (detailed in section 5.10) calculated on whole calendar weeks of geolocation data.

7.10.1.2 Performance evaluation

Once the individuals had been grouped and clustered, performance of the regression was measured using a number of metrics.

To assess performance, the Bayesian Lasso model over multiple individuals (described in section 7.7.2) was run using the appropriate parameter values with left out data to assess performance. For each individual, the model was trained on labelled training data from (a) the first half of the available labelled data for that individual, up to a maximum of 8 data points; and (b) all the labelled data from the other individuals in the same cluster. If the individual was in a singleton cluster (i.e. the only individual in the cluster) then the model was trained only on the first half of the available labelled data for that individual, up to a maximum of 8 data points. Each model was trained for 1000 iterations on the training data, with the mean of all the sampled $|\hat{\beta}_i|$ values used to estimate the value of the QIDS score for the left out test data. The following metrics are presented from the sampled variables.

Mean MAE First the MAE is calculated from the estimated QIDS score and the labelled test data for the left out data from each individual in turn, and the mean MAE across all individuals is presented.

The mean MAE is the primary metric to optimise, where lower mean MAE is desirable. The ultimate aim is to improve performance of regression on the left out weeks of data, and MAE is a suitable metric for this.

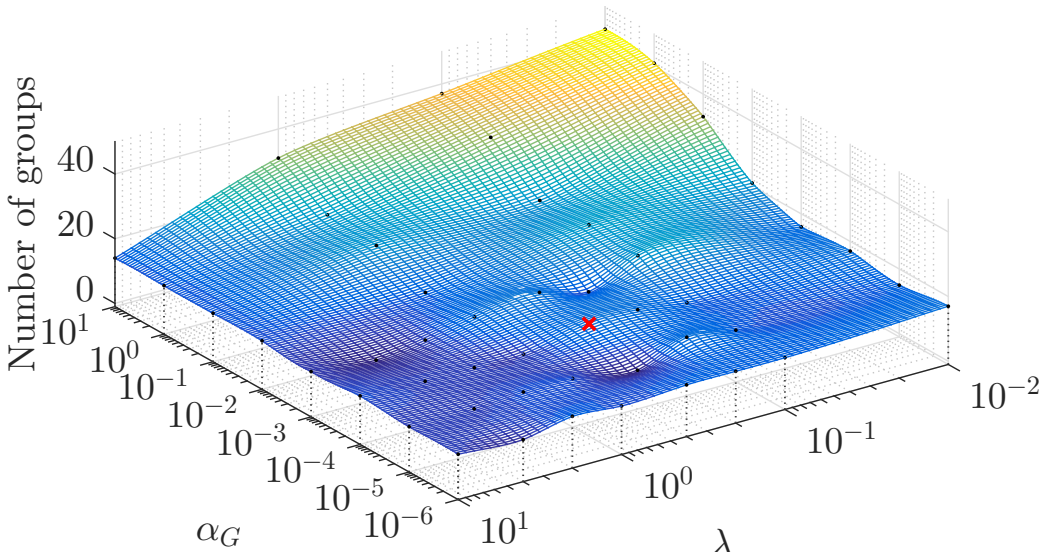
Mean absolute $\hat{\beta}_i$ values The mean absolute $\hat{\beta}_i$ values sampled, denoted $\overline{|\hat{\beta}_i|}$, is calculated as the mean of all sampled $|\hat{\beta}_i|$ values $\forall\ i$ for each partially left out individual in turn. The mean across all individuals is presented.

The absolute $\hat{\beta}_i$ values sampled is a measure of how closely the models sampled are fitted to the training data. Generally, with smaller λ, the level of regularisation is less, which should lead to larger absolute $\hat{\beta}_i$ values. This is also related to the number of clusters sampled, discussed next.

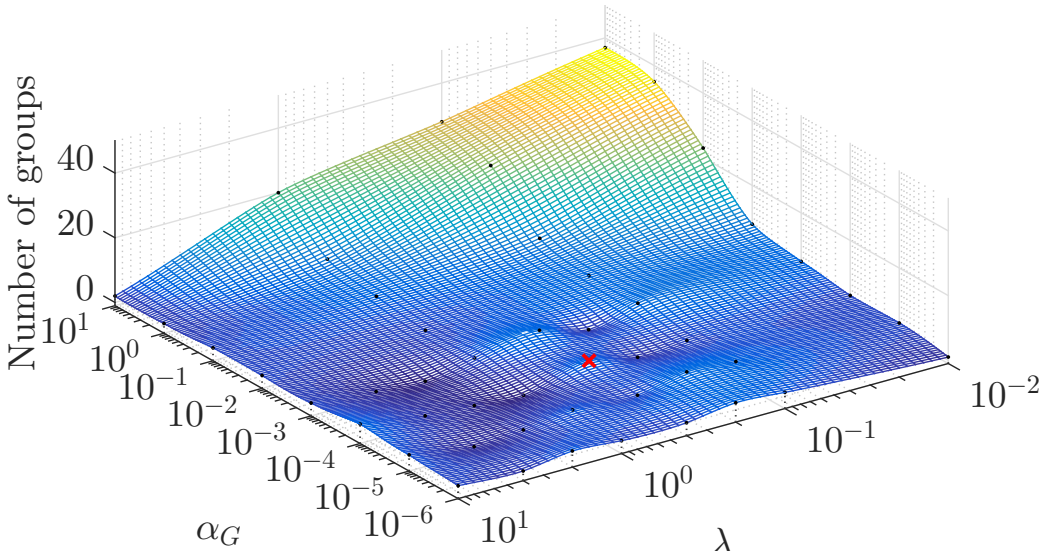
Number of clusters The number of distinct clusters sampled is a useful indicator of how the model is functioning. Although not a direct indicator of performance, and not directly useful for model optimisation, generally sampling fewer clusters is desirable because it means that the clusters will contain more individuals. This in turn should lead to more generalisable models in each cluster. Clusters with fewer individuals will tend to result in regression models fitting very closely to the individuals which may reduce performance on unseen test data. In the extreme case, singleton clusters containing only one individual indicate that the model is fitting very closely to the training data, and performance will be similar to the "fully personalised" models discussed previously.

7.10.2 Parameter tuning results

There are two main parameters of the 1st pass grouping model that need to be tuned. These are α_G, the concentration parameter of the Dirichlet process prior, and λ, the regularisation parameter of the Bayesian Lasso model. These parameters are intrinsically linked and must be optimised together. A grid search was therefore run over both parameters, varying α_G on a log scale between 10^{-6} and 10^{1}; and varying λ on a log scale between 10^{-2} and 10^{1}. The results of the parameter optimisation are shown in the four graphs in figure 7.25 across the next two pages.

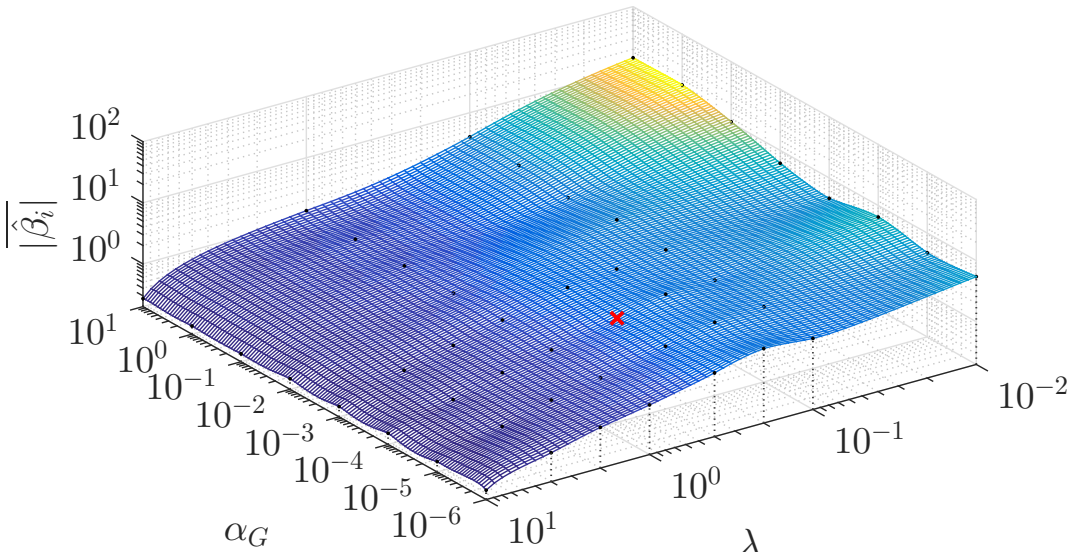


(a) The total number of clusters identified.

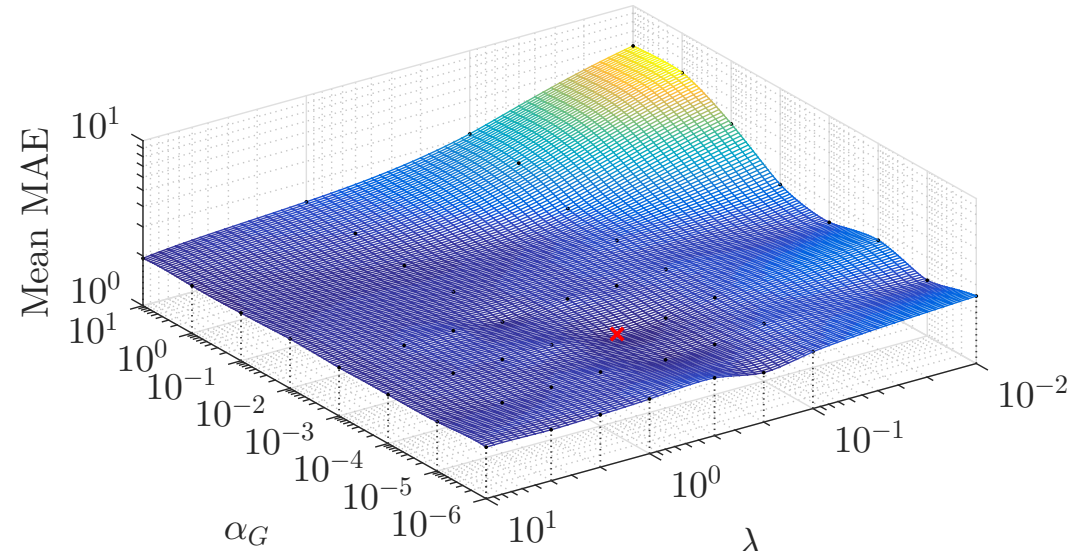


(b) The number of clusters containing only one individual.

Figure 7.25: Parameter tuning of the λ and α_G parameters of the 1st pass grouping model on geolocation data. The full two-pass grouped Bayesian Lasso was run on the 10 main features of the geolocation data for 5000 iterations for each pair of parameter values shown as black dots. Four different metrics are presented on this and the next page demonstrating different aspects of performance as discussed in detail in the text. The total number of clusters (a) shows that high values of λ or lower values of α_G result in fewer clusters being created. The total number of clusters is also highly correlated with the number of singleton clusters with only one individual (b) implying that when λ is small and α_G is large the model for each cluster closely fits the individuals allocated to that group, and any remaining individuals are allocated to their own clusters rather than fitting into the closest matching cluster. The red cross indicates the 'optimal' parameter combination ($\alpha_G = 10^{-4}$; $\lambda = 0.4$) with the minimum mean MAE in (d). (continued on the next page)



(c) Mean absolute $\hat{\beta}_i$ values sampled from a Bayesian Lasso model.



(d) Mean absolute error of regression on left out individuals.

Figure 7.25 (cont.): Performance of regression in clusters found by the grouped Bayesian Lasso on geolocation data. Results are from the Bayesian Lasso over multiple individuals trained for 1000 iterations on the first half of the data for an individual, up to a maximum of 8 data points, and all the labelled data from other individuals assigned to the same cluster (if any). This is repeated for all individuals and results are calculated on the labelled data points for each individual not included in the training of the model. The mean absolute $\hat{\beta}_i$ values sampled (c) demonstrates the effect of the regularisation under the different parameter configurations. This shows that when λ is large, the model fits to the average of the training data with small $\hat{\beta}_i$ values, while when λ is smaller the model closely fits the data with large $\hat{\beta}_i$ values. The mean absolute error of regression (d) demonstrates the predictive performance of the model. The mean test value is used to calculate the MAE for that individual. The error blows up when λ is small due to over-fitting in the singleton clusters.

Number of clusters Firstly, figure 7.25(a) shows the number of clusters found in the final clustering. While there is an element of noise due to the clustering method used, there is a clear trend that with large α_G and small λ more distinct clusters are sampled. The relationship with α_G is expected from the generative definition of the Dirichlet process in section 3.3, in particular equation (3.53), which states that as α_G gets larger, it is more likely that new groups are created. The relationship with λ is more complex, but also makes intuitive sense from the definition of the Lasso given in equation (3.69) on page 67. When λ is large, the β_i coefficients in the model are highly regularised, meaning that as $\lambda \to \infty$, $\beta_i \to 0 \ \ \forall \ \ i > 0$, and therefore $\beta_0 \to \bar{\mathbf{x}}$. In other words, the model simply finds the mean value of the data when λ is large. Because the variance in mean QIDS scores for all individuals is relatively small, fewer groups are created, even when α_G is large. This is particularly clear when $\lambda = 10^1$. At the other extreme, when λ is small, the amount of regularisation on the β_i coefficients is small, to the point where there is effectively no regularisation. This will encourage more groups to be created, even when α_G is small. This is because the regression model in any clusters created will fit the individuals in that cluster so closely that many individuals do not fit well into any of the existing clusters, and must therefore form new clusters. At the peak when $\alpha_G = 10^1$ and $\lambda = 10^{-2}$, almost every individual was allocated to a unique cluster.

Number of singleton clusters The number of singleton clusters found, i.e. individuals that did not commonly occur with any other individuals in the 1$^{\text{st}}$ pass grouping of individuals, is shown in figure 7.25(b). This is highly correlated with the total number of clusters shown in figure 7.25(a) and is useful to understand the effect of the parameters. Singleton clusters tend to occur when the regression model in each cluster is closely fitting to individuals. When λ is small and α_G is large, the number of singleton clusters is also large. This indicates that the model is clustering the most similar individuals and leaving all remaining individuals in singleton clusters.

Mean absolute $\hat{\beta}_i$ values To further explore the effect of the regularisation under the different parameter configurations, it is useful to consider the mean absolute $\hat{\beta}_i$ values sampled in each group, denoted $\overline{|\hat{\beta}_i|}$, shown in figure 7.25(c). This shows clearly how large values of λ, i.e. $\lambda = 10^1$ result in small mean absolute $\hat{\beta}_i$ values, again indicating that the model for each cluster is simply finding the mean value of the data for the left out individuals. As λ is decreased, the mean $|\hat{\beta}_i|$ values increase relatively evenly across all values of α_G showing that the model is including more features and fitting the data more closely. There is however a noticeable peak when λ is small and α_G is large, which is also where most clusters are created, and many singleton clusters are created. This indicates that the model is over-fitting the training data when many singleton clusters are created.

Mean MAE Finally, the performance of the model is demonstrated in figure 7.25(d), presented as mean MAE for all individuals. As discussed above, parameter combinations that result in models that fit closely to the data in the cluster tend to result in many singleton clusters being formed. For individuals in singleton clusters, the Bayesian Lasso model over multiple individuals was only trained using the first half of the data, up to a maximum of 8 data points, and tested on the remaining data. As seen from the fully personalised models shown in figure 7.2, this tends to over-fit the training data resulting in poor performance. This can be seen by the peak in mean MAE in figure 7.25(d) when λ is small and α_G is large, matching the peak in the number of singleton clusters in figure 7.25(b). Although not shown in figure 7.25, the mean MAE blows up if λ is decreased further.

7.10.2.1 Optimal parameters

As discussed in section 7.10.1.2, the most suitable metric to optimise on is the mean MAE. The optimal performance, indicated by the red cross in figure 7.25, was found to be when $\alpha_G = 10^{-4}$ and $\lambda = 0.4$. The clusterings found at this point are shown in figure 7.26 on page 247. The similarity in groupings found in the different clusters

is clear. The grouped regression results shown previously in figure 7.2 on pages 178 and 179 were generated from these clusterings.

From the results in figure 7.25, it can be seen that this is in a region of the mean MAE results with optimal performance, and the mean absolute $\hat{\beta}_i$ values are not at either extreme of over-fitting or under-fitting the training data. The number of clusters is also towards the lower end, although there are singleton clusters found. This optimal clustering thus appears to match the expectations given in section 7.10.1.2.

7.10.3 Varying the λ parameter

Although the clusterings shown in figure 7.26 were generated with parameter $\lambda = 0.4$, having this clustering enables testing the effect of varying the parameter in greater detail without allowing the clusterings to change. This is shown in figure 7.27 for the same two metrics given in figures 7.25(c) and (d) where the λ parameter is varied between 10^{-3} and 10^{1}. The results were generated using the same methodology as described in section 7.10.1.2 above.

The two metrics show the same trends as discussed above but in greater detail and clarity. As λ is decreased, the regularisation is decreased which means that the mean $|\hat{\beta}_i|$ value increases, but the test error (mean MAE) also increases due to the model over-fitting the training data. Although not shown in this scale, if λ is decreased further to 10^{-4}, the mean MAE blows up to $\approx 10^{150}$. On the other end of the scale, as λ is increased, the mean $|\hat{\beta}_i|$ value decreases steadily towards zero. The mean MAE initially improves as the over-fitting is reduced, and then worsens again as the model simply finds the mean of the training data in each cluster.

The results where $\lambda = 0.4$ (the parameter value used to generate this particular clustering) is indicated by the red cross. A non-significant improvement in mean MAE is seen by increasing the value of λ from 0.4 to 0.6. This change in λ reduces the mean MAE from 1.9001 to 1.8865 ($p = 0.52$).

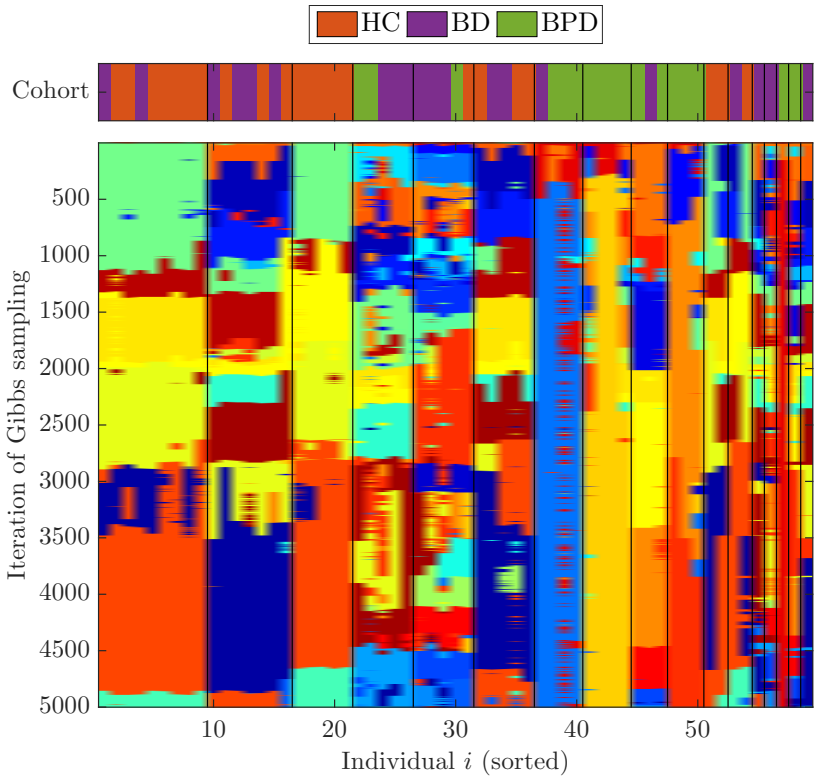


Figure 7.26: Clusterings of individuals found under the optimal parameters from figure 7.25 when $\alpha_G = 10^{-4}$ and $\lambda = 0.4$. The coloured bar at the top of the figure shows the cohort of each individual.

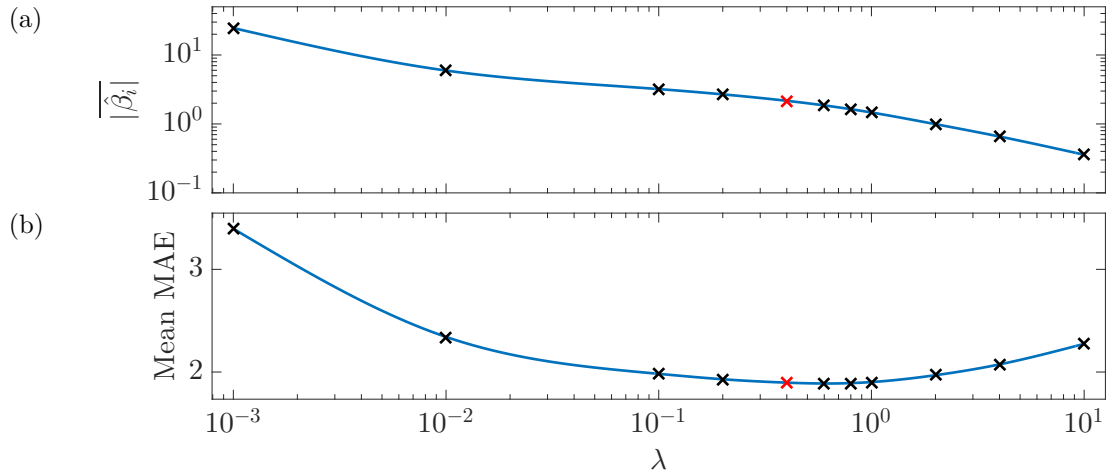


Figure 7.27: Effect of varying the λ parameter on the fixed clustering shown in figure 7.26 using the same two metrics given in figures 7.25(c) and (d). The red cross shows the point where $\lambda = 0.4$, the value used to generate the grouping originally.

7.10.4 Varying other parameters

The model definition of the 1st pass grouping in equation (7.26) shows that there are four prior parameters for two variables that have not yet been considered. These are the β_0 variable, which has two parameters (μ_{β_0} and $\sigma^2_{\beta_0}$); and the σ^2 variable, which also has two parameters (α and γ).

A grid search over these parameters is shown in figure 7.28 where the four parameters have been varied while running the grouped Bayesian Lasso over multiple individuals using the same methodology as described in section 7.10.1.2 above. The parameters for each of the two variables were tuned separately. The clusterings shown in figure 7.26 were used for all tests. In both figures, the parameters used in generating the results shown in figure 7.25 ($\mu_{\beta_0} = 0$, $\sigma^2_{\beta_0} = 10$, $\alpha = 1$ and $\gamma = 1$) are shown as a red cross. The baseline performance at this point is a mean MAE of 1.9001.

Varying the parameters of the β_0 variable in figure 7.28(a), especially the mean μ_{β_0}, has a small but noticeable effect on mean MAE with optimal performance at $\mu_{\beta_0} = 6$ and $\sigma^2_{\beta_0} = 15$ where the mean MAE is 1.8702 ($p = 0.30$).

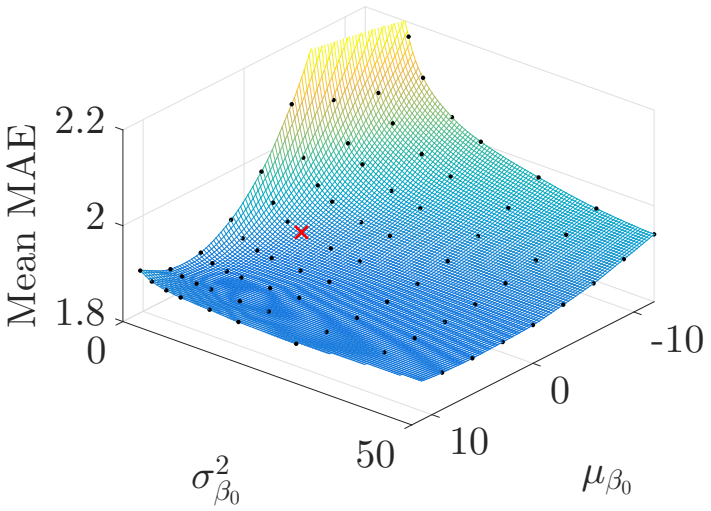
Varying the parameters of the σ^2 variable in figure 7.28(b) has a much smaller effect on mean MAE with optimal performance at $\alpha = 20$ and $\gamma = 10$ where the mean MAE is 1.8884 ($p = 0.23$).

This demonstrates that the model is robust to changes in these parameter values.

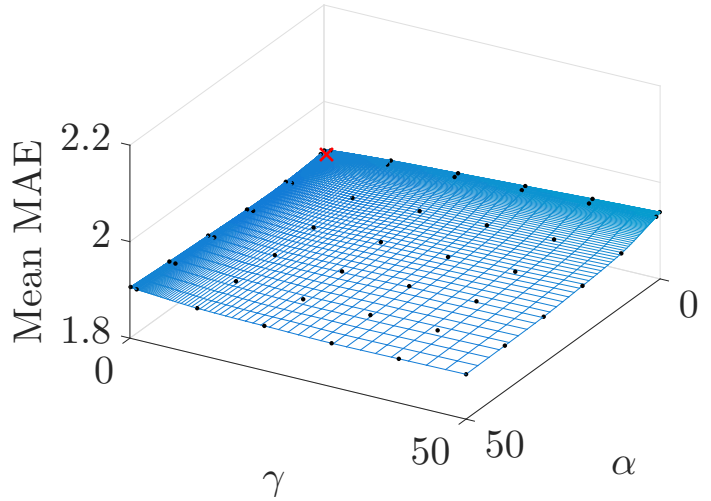
7.10.5 Combined model with optimised parameters

Combining the optimal parameter values for μ_{β_0}, $\sigma^2_{\beta_0}$, α and γ found in the previous section with the optimal value for λ found in section 7.10.3, a final "optimal" model can be created from the tuned parameters ($\lambda = 0.6$, $\mu_{\beta_0} = 6$, $\sigma^2_{\beta_0} = 15$, $\alpha = 20$ and $\gamma = 10$). This model, run with the same methodology described in section 7.10.1.2, also gives non-significant improvement in mean MAE of 1.8642 ($p = 0.38$).

This again demonstrates that the model is robust and that further parameter optimisation is unlikely to significantly improve the results.



(a) Parameters of β_0 variable, μ_{β_0} and $\sigma^2_{\beta_0}$.



(b) Parameters of σ^2 variable, α and γ.

Figure 7.28: Effect of varying other model parameters on fixed clustering shown in figure 7.26. The β_0 variable (a) has two parameters (μ_{β_0} and $\sigma^2_{\beta_0}$); and the σ^2 variable (b) has two parameters (α and γ). The red cross shows the parameter values used to generate the results in figures 7.25 and 7.26. Although minor improvements are possible for both variables with small changes to the parameters, they are not significant.

7.11 Results on geolocation data

This section presents results of QIDS scores estimation from the geolocation-derived features using the grouped Bayesian Lasso model presented in sections 7.6 to 7.9 and tuned to the geolocation data in section 7.10. The performance of individuals in the grouped Bayesian Lasso model is presented with

a) individuals allocated to the groups found to be optimal in figure 7.26; and
b) individuals allocated to groups using limited calibration data as if they were new individuals as described in section 7.3.

Additionally, comparative results are presented showing estimation using

a) a 'global' model trained across all other individuals;
b) a 'fully personalised' model trained using cross-validation over all data from the test individual;
c) a 'fully personalised' model trained only on the calibration data from the test individual;
d) an adapted version of the feature-based clustering model presented by Lane et al. (2011) / Abdullah et al. (2012).

The different models tested, and the evaluation framework, are described in the subsection that follows.

7.11.1 Evaluation framework

Most of the models tested, described in detail in the following subsections, were evaluated in a leave-one-participant-out framework (see section 3.5) where each participant is left out and the relevant model is trained on the data from the other relevant participants (as described for each model in turn).

In practice, 'fully personalised' and 'grouped personalised' models generally require a certain level of calibration data (in this case objective features and QIDS scores) as described in section 7.3. To simulate this, in most of the models evaluated, a subset of data from the start of the recording for the left-out test individual was

included in the training of the model (but was excluded when evaluating test performance). The number of weeks of calibration data included in the model for training is defined as the first half of the available data for the test individual up to a maximum of 8 weeks, i.e.

$$n_{\text{cal}} = \min(\lceil N/2 \rceil, 8) \tag{7.49}$$

where N is the total number of weeks of data available for the individual. For example, for an individual who provided 10 weeks of data, the first 5 weeks are included in the training, and the last 5 weeks are used for evaluation. For an individual who provided 16 or more weeks of data, the first 8 weeks are included in the training, and all the remaining data are used for evaluation.

In all cases (except the feature-based clustering model described in section 7.11.1.6), standard Bayesian Lasso models by Park & Casella (2008), described in detail in section 7.7.1, were trained on the calibration data for the left out individual and all data from the other relevant individuals. The Bayesian Lasso model was trained with regularisation parameter $\lambda = 0.4$ (the same value found to be optimal for the 1$^{\text{st}}$ pass grouping model in section 7.10.2) for 1000 iterations. Because the Bayesian Lasso is implemented using a Gibbs sampler, this leads to 1000 samples of the model coefficients. The performance of the model was evaluated in each iteration by applying the Bayesian Lasso with the sampled coefficients on the evaluation data for the left out individual. Performance was evaluated as the mean MAE of the estimated QIDS scores and the reported values. Improvements in performance were compared using a single-tailed paired t-test of the mean MAE for each individual.

7.11.1.1 Global model

The 'global' model was tested by training a Bayesian Lasso model trained on the calibration data from the test individual and all the data available from all the other individuals.

7.11.1.2 Grouped model with optimal cluster allocations

The grouped personalised model was run using Neal's Algorithm 8 for 5000 iterations with all the data available from all individuals as described in section 7.10.1.1. This provided the optimal groupings shown in figure 7.26.

To evaluate the performance of these optimal groupings, a Bayesian Lasso model was trained on the calibration data from the test individual and all other individuals found to be in the same group. Note that if an individual ends up in a group containing just itself (i.e. the five individuals on the right in figure 7.26), then the grouped personalised model model will just be trained on the calibration data from that individual, which reduces to the fully personalised model described below.

7.11.1.3 Grouped model with clusters allocated using calibration data

The grouped personalised model tested retrospectively as described above is useful to demonstrate the principles of the operation of the model and the clinical relevance of the groups found.

To apply the model prospectively, the calibration data from new individuals must be used to allocate them into one of the set of groups. To test this, each individual was left out in turn and separate Bayesian Lasso models were trained for each of the groups. For the group that the left-out individual was originally assigned to, the Bayesian Lasso model was trained on the remaining individuals in that group (if any).

QIDS scores were estimated from the calibration data for the left-out individual using the model trained on each of the groups. The mean MAE of the estimated QIDS scores using the model for each group was evaluated, and the individual was allocated to the group that provided optimal performance.

To provide the final prediction, a new Bayesian Lasso model was trained using the calibration data from the left-out individual and all the data from the other individuals assigned to the allocated group.

7.11.1.4 Fully personalised model using all available data

It is common in the literature (e.g. Gruenerbl et al., 2014; Grünerbl et al., 2015; Abdullah et al., 2016) to use cross-validation over all available data points from an individual to demonstrate a 'personalised' model.

This was implemented using random sub-sampling of all the data from each individual, where the Bayesian Lasso model was trained using 80 % of the data randomly selected from the test individual, with performance evaluated on estimation of QIDS scores using the remaining 20 %. This split was repeated 10 times.

7.11.1.5 Fully personalised model using calibration data only

While the fully personalised model using all the available data is commonly presented in the literature, it may not provide a fair representation of the accuracy of the model in practice because it does not demonstrate how well the model will generalise when trained with limited calibration data.

As an alternative, the fully personalised model using calibration data only was tested by training a Bayesian Lasso model on just the calibration data from the test individual.

7.11.1.6 Feature-based clustering model

A comparative model was tested using the feature-based clustering method described by Lane et al. (2011) and Abdullah et al. (2012). To provide a comparative result, only the part of the method that clusters individuals by the similarity of their feature values was included. The clustering method by Lane et al. / Abdullah et al. used a locality-sensitive hashing method known as random projection (Andoni & Indyk, 2008) as a similarity measure of the feature values between all pairs of individuals. The random projection method samples random values in the feature space, and calculates the distance from these random values to the extracted feature values for each individual. By repeating this process multiple times, features from individuals that are similar will commonly be closest to the same randomly sampled values. Lane et al. / Abdullah et al. use the resulting similarity matrix to condition an online

Boosting classification algorithm. Because the QIDS estimation model tested here concerns regression rather that classification, this has been replaced with a regression equivalent described by Drucker (1997).

Because the clustering model does not use the relationship between the objective features and the output, all the available data from all individuals, including the test individual, were included when assessing similarity. In practice, only data up to and including the week being tested would be available.

7.11.2 Performance results overview

The performance of the different models tested across all individuals are summarised in table 7.4. Results are presented as MAE (mean ± std) for all models. Overall estimation accuracy is increased from the 'global' model by using both the fully personalised and the grouped personalised models.

The optimal results are achieved with the fully personalised model trained using sub-samples of all the available data. This is not surprising since the full range of values are included in the training of the model. This means that future labelled data may be used to make past predictions, which is not a realistic use-case. Consecutive data points also have high levels of short-term correlation between both behaviour and mental state. This means that if the 'test' set consists of points randomly removed from the full data set then there may be high levels of leakage of information, leading to misleading results.

The grouped personalised model with individuals allocated to their optimal groups improves over the 'global' model and is similar to the performance of the fully personalised model trained using sub-samples of all the available data. A single-tailed paired t-test confirms that the improvement in performance by using the grouped personalised model with optimised clusters from the 'global' model is highly significant ($p < 10^{-12}$).

The grouped personalised model with individuals assigned to groups using only their calibration data also performs significantly better than the 'global' model. It

Table 7.4: QIDS score estimation results using personalised and comparative models.

Model type	HC [a]	BD [a]	BPD [a]	Overall [a,b]
'Global' model*	4.86 ± 2.54	4.74 ± 2.07	6.43 ± 2.58	5.27 ± 2.48
Fully personalised model using cross validation over all data points	0.80 ± 0.76	2.27 ± 1.44	3.05 ± 1.67	1.94 ± 1.60 $^{*}p = 3.8 \times 10^{-12}$ $^{**}p = 1.5 \times 10^{-4}$
Fully personalised model trained on calibration data**	1.06 ± 0.75	3.67 ± 2.60	4.38 ± 2.43	2.90 ± 2.49 $^{*}p = 5.0 \times 10^{-7}$
Feature-based clustering based on Lane et al. (2011) / Abdullah et al. (2012)	5.33 ± 3.11	4.50 ± 2.30	6.15 ± 2.88	5.29 ± 2.82
Grouped personalised model with optimal cluster allocations	0.83 ± 0.52	2.30 ± 1.96	2.82 ± 1.28	1.90 ± 1.60 $^{*}p = 2.1 \times 10^{-13}$ $^{**}p = 8.1 \times 10^{-7}$
Grouped personalised model with clusters allocated using calibration data	0.86 ± 0.46	3.30 ± 3.09	3.75 ± 2.07	2.52 ± 2.47 $^{*}p = 2.6 \times 10^{-8}$ $^{**}p = 0.0208$

[a] Results presented as mean $\pm$ std of MAE of QIDS score estimation.
[b] Significance of improvement in performance from:
* 'global' model; and
** fully personalised model trained on calibration data
shown after the overall results indicated by asterisks.

also improves on the fully personalised model trained only on the same calibration data with mild significance.

The performance of the fully personalised model trained using only the calibration data indicates the difficulty of generalisability of a model trained using limited data. The feature-based clustering method by Lane et al. (2011) / Abdullah et al. (2012) improves on the 'global' model for BD and BPD participants but overall does not perform better. This can be explained by the method being based entirely on similarity of the input features, and not on the similarity of the relationship between the input features and the QIDS score.

7.11.3 Grouped personalised detailed results

The distributions of regression errors of the global; fully personalised; and grouped personalised models in table 7.4 are presented in figure 7.29 for all 59 individuals included in the analysis. This shows the MAE of estimations made using (a) the 'global' model; (b) the fully personalised model trained on calibration data only; (c) the fully personalised model trained using cross validation over all available data; (d) the grouped personalised model with individuals allocated to their optimal clusters; and (e) the grouped personalised model with individuals allocated to clusters using their calibration data.

In all graphs the bars are shaded in proportions corresponding to the cohort of the individuals in that bar. It can be seen from this that the BPD individuals tend to perform worst under the 'global' model. The best performance under all the personalised models are from the HC participants, who showed very variable performance under the 'global' model. However, this should be viewed with caution because the HC participants also tend to show very low QIDS scores with very little variation, as demonstrated in figure 5.7, meaning that a model may just estimate a constant low value. Similarly, some of the BD participants also perform very well under the grouped personalised model, but as will be discussed in the next section, some of the BD individuals were allocated with HC individuals indicating that they may also have relatively constant low QIDS scores. The remaining BD and BPD individuals clearly show a large improvement under the grouped personalised model.

→

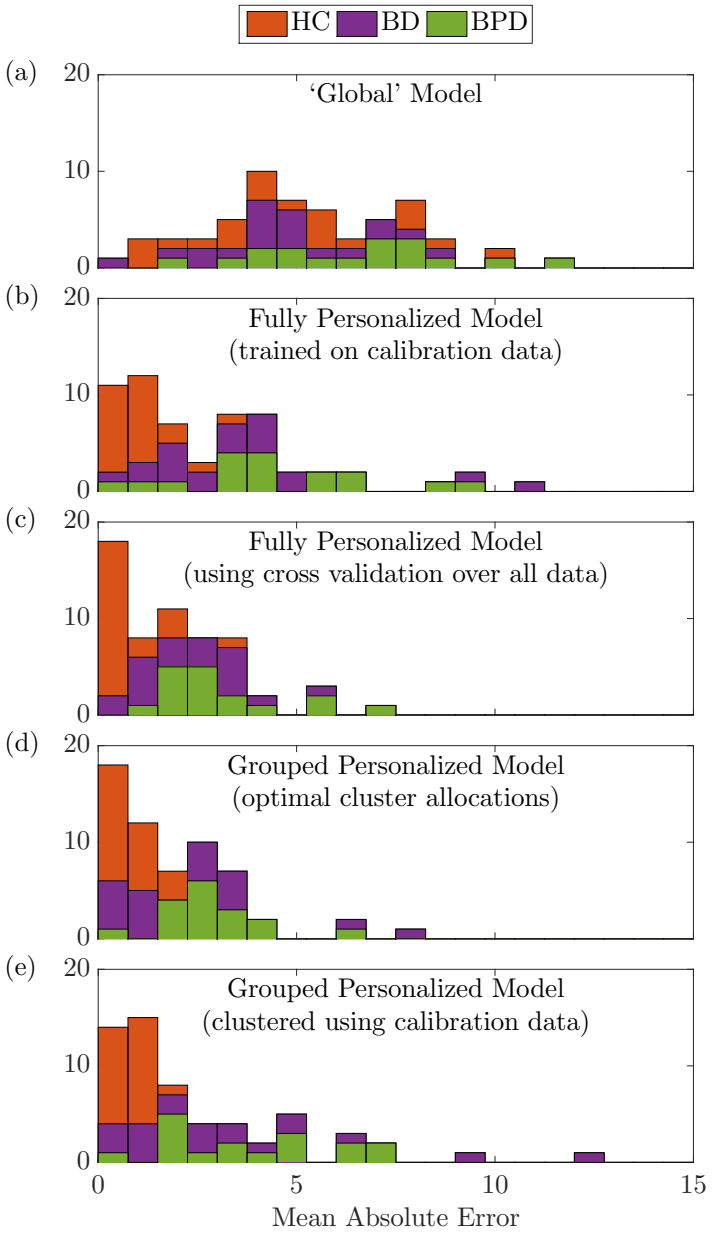


Figure 7.29: Distributions of regression errors in (a) the 'global' model shown in figure 7.2; (b) the fully personalised model trained on calibration data only; (c) the fully personalised model trained using cross validation over all data points; (d) the grouped personalised model with optimal cluster allocations shown in figure 7.2 and tuned in section 7.10.2; (e) the grouped personalised model with clusters allocated using calibration data. Bars are shaded in proportions corresponding to the cohort of the individuals in that bar.

7.11.4 Cluster analysis

The coloured bar at the top of figure 7.26 indicates the cohort of individuals assigned to each cluster. These groupings are shown in greater detail in figure 7.30. In total 17 groups were found, each containing between 1 and 9 individuals. Groups 1 to 4 are predominantly HC individuals with BD participants who tend to exhibit low variability in their QIDS scores. This is to be expected from HC or euthymic BD participants. The remaining groups containing BD and/or BPD tend to have greater variability in QIDS scores as well as higher baseline values. Groups 5 and 6 are predominantly BD individuals, and groups 7 to 10 are predominantly BPD individuals. Nine individuals were assigned to groups 11 to 17 with only one or two individuals in each group. These individuals have been shown as unassigned because they did not fit well into any of the other groups, and therefore solid conclusions cannot be drawn from them. As more individuals are available for analysis, these individuals would likely be assigned to larger groups.

These findings imply that there is an element of overlap between HC and BD individuals. This fits with the informal observation that euthymic BD individuals exhibit relatively normal mood. By contrast, there is very little overlap between HC and BPD group membership. This finding broadly aligns with other subjective measures of mental state, sampled much more frequently than weekly, which show a gradient of abnormal mood where HC $<$ BD $<$ BPD (Carr et al., 2018a, 2018b).

The number of groups found (17 for the 59 participants) indicates the high level of variability in the relationships between the behavioural data and mental state for the individuals in the study. The number of groups may also be affected by the properties of the DP prior on smaller sample sizes. As discussed in section 3.3, asymptotically the DP exhibits a 'rich-gets-richer' property, where larger clusters tend to get larger (Wallach et al., 2010). This also has the side effect that with more data points, the number of clusters relative to the number of samples decreases (more precisely, the expected number of clusters grows logarithmically with the number of data points

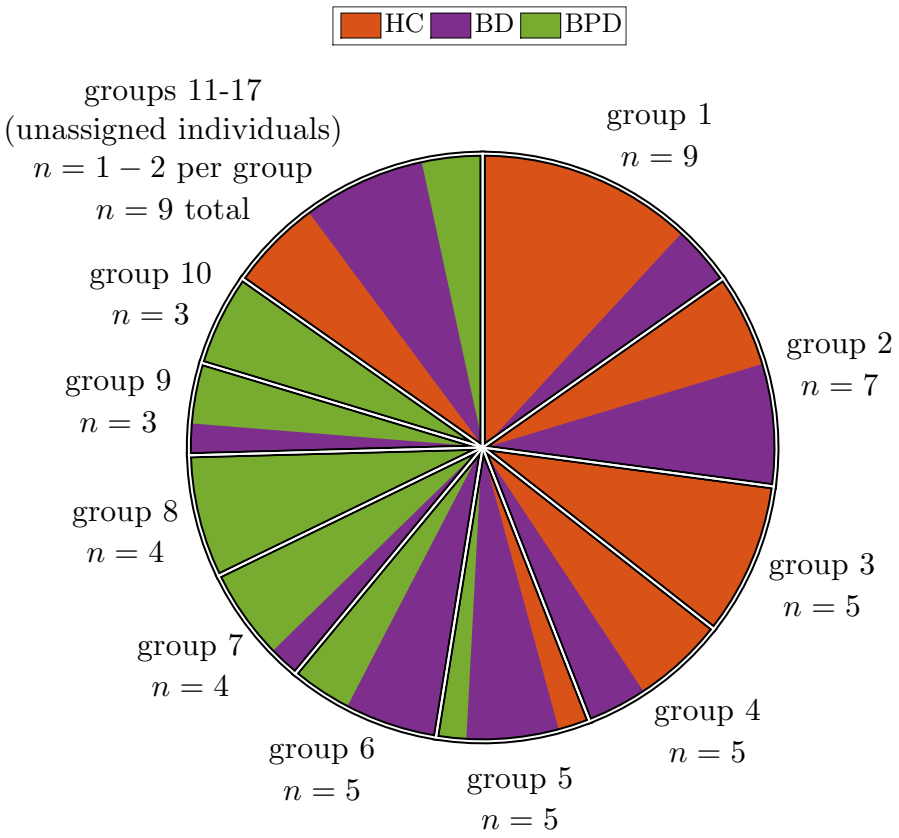


Figure 7.30: Allocations of individuals to different groups showing the cohort of each individual. Each group is shown as an outlined slice, with the shading indicating how many individuals are from each cohort in that group.

proportional to the concentration parameter in the DP prior[5]). With fewer data points, none of the clusters are yet large enough to sufficiently attract other data points. This means that as more data are available to train the model, the relative number of clusters should decrease, thus improving the stability of the model.

[5] This follows by summing the probability of creating new clusters in the DP prior given in equations (3.52) and (3.53) over N data points, and taking the limit as $N \to \infty$ (Wallach et al., 2010).

7.11.5 Group characteristics

One of the key advantages of finding similar groups of individual within the population is that the models for each of the groups found may indicate different characteristics of subgroups of patients. This can be explored by inspecting the coefficient values sampled for each group. For this, a Bayesian Lasso model was run on all individuals allocated to each group with regularisation parameter $\lambda = 0.4$ for 1000 iterations. This results in the sampled coefficient values shown in figure 7.31 for groups 7 and 9 in figure 7.30.

The group in figure 7.31(a) is characterised mainly by the DM, TT, and TD features; while the group in figure 7.31(b) is characterised by the number of clusters visited. In both cases, the other parameters are effectively removed from the model, as is expected from a Lasso-based model. Statistically significant Pearson correlation coefficients ($p < 0.01$) of each feature with the reported QIDS scores for individuals in the group are shown on the right the feature plots. This shows that the grouped personalised model tends to pick out the most correlated features for each group, although this may not necessarily be the case as two features with low correlations may be predictive when combined in a regression model. Inspection of the model coefficients in each of the clusters found indicate that they all model different characteristics of the range of relationships between the behavioural data and the QIDS scores. Most groups are also characterised by three or fewer main features, indicating the interpretability of the models found.

→

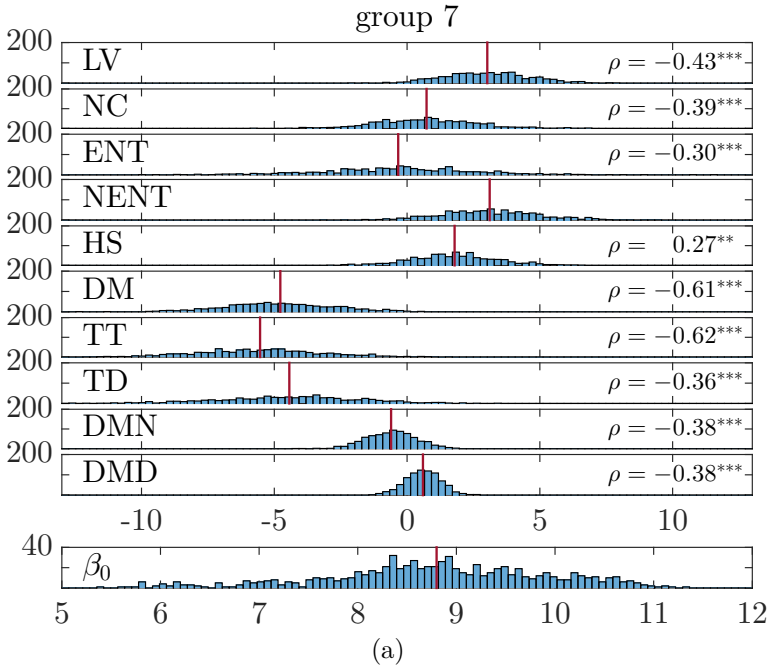


Figure 7.31: Sampled coefficient values for (a) group 7; and (b) group 9 in figure 7.30 found by the grouped personalised model. Each row shows the parameter coefficients for the 10 features included in the model, and the mean parameter β_0. Group 7 (a) is characterised by the DM, TT, and TD features; while group 9 (b) is strongly characterised by the NC feature only. Significant correlations ($p < 0.01$) between features for individuals in the group and the reported QIDS scores are shown on the right of each feature (significance of the correlation is indicated by asterisks: $^{*}p < 0.01$; $^{**}p < 0.005$; $^{***}p < 0.001$).

7.12 Discussion

The results in this chapter have demonstrated a new model to find groups of individuals that exhibit similar regression relationships between geolocation-derived features and self-reported depression as measured through the QIDS questionnaire. The previous chapter demonstrated the utility of geolocation-derived features as predictors of mental state in BD patients, and this chapter extended this analysis to cover HC; BD; and BPD participants in the AMoSS study.

Using 'global' models trained on all individuals was demonstrated to be suboptimal as the inter-individual variability in behavioural patterns was too high, as the data shown in figure 7.1 suggested would be the case. While fully personalising a model for each individual intuitively makes sense, it requires too much data to be practicable in a prospective fashion and provides little clinical insight into how behaviour changes with varying levels of depression as every subject follows a different model. The significant improvement obtained by using fully personalised models trained using cross-validation over all available data rather than just calibration data from the start of the recording is important. In addition to the correlation between successive data points, it is clearly not generally useful in practice to include future data to make past predictions. Yet this method is commonly reported in the literature as an example of a fully personalised model. These results should therefore be interpreted with extreme caution.

The grouped personalised model is presented as an alternative method for personalisation where sub-groups of individuals that exhibit similar relationships between their objective features and mental state are automatically identified. It was shown that this leads to plausible groupings, as most HC, BD and BPD individuals have been assigned to different groups, reflecting the relationship between different disease categories and different behavioural patterns. However, further validation with more data is needed to assess whether these are the 'optimal' groupings.

While the groupings provide insights, the key challenge for applying grouped personalised models in practice remains the determination of which group to allocate new

individuals to. In the results presented, up to 8 data points (pairs of objective features and actual QIDS scores) were used to allocate each individual to their group based on the performance of QIDS score estimation using the model for each group. This provided a significant improvement over the 'global' model and a smaller, but still mildly significant, improvement over the fully personalised model trained on the same data used for calibration. The main advantage of the use of the groups in practice might be to enable better understanding of the characteristics of the patient. While the fully personalised model may take any form, the grouped personalised model is restricted to a known set of behavioural phenotypes. With enough individuals available to train the models, an exhaustive set of behavioural phenotypes can be obtained, and perhaps specific therapies tailored to each one. Matching individuals to groups purely based on the predictive performance of the extracted groups on the calibration data, as done here, is a naïve approach, and using other indicators of similarity, such as demographics or similarity of sensor data, may improve group identification when only limited calibration data are available.

The fact that a high number of groups relative to the number of individuals are found suggests that there is indeed a high level of inter-individual variation between subjects. Having access to larger datasets should increase the number of subjects belonging to each group due to the properties of the DP prior employed. With the current dataset, groups with clearly different behavioural patterns could be identified, as shown in figure 7.31. Significant improvements in performance in estimating levels of depression from objective, geolocation-derived, features were shown, demonstrating further the appropriateness of the groupings, but also the utility of using geolocation as an objective marker for mental health.

Similar to the 'global' model results in the previous chapter, the personalised results here have been presented as independent estimates of depression levels for weeks of geolocation data and ignore any time-series dynamics that appear to exist in the data examples shown in figure 7.2. Future work should explore incorporating these dynamics into the personalised models, as discussed in greater detail in section 8.3.2.

Chapter 8

Conclusions and Future Work

This thesis has thoroughly demonstrated the potential of the geolocation data collected as part of the AMoSS study. Analysis was performed in two parts. First the data were analysed using population-level 'global' models, trained on all individuals and tested using cross-validation. With 'global' models, classification of depression in the BD participants achieved pooled accuracy of over 75 % from geolocation-derived features alone. Regression of the level of depression achieved MAE rates of 3.593 in the same participants.

The results from the global models demonstrated the utility of geolocation-derived features as objective markers of mental state. However, inspection of the feature distributions indicated that there were high levels of inter-individual variability in the data distributions. This indicated that different individuals follow different models linking their objective features to levels of depression. There was therefore a need to personalise the models applied for each individual, rather than using the 'global' models trained over all other individuals. A grouped model was developed to find groups of individuals that have similar regularised linear regression models relating their objective features and levels of depression. These models were found to perform significantly better than the 'global' models. When individuals were assigned to groups based on calibration data from the start of their recordings, the grouped model performed significantly better than fully personalised models trained only on the same calibration data.

This thesis has therefore demonstrated the limitations of 'global' models, and the importance of personalisation when processing these types of data. Grouped personalisation in particular is demonstrated as an appropriate way augment the limited training data from an individual with data from similar individuals in the population.

8.1 Key methodology

The data analysed in this thesis were collected from participants in the AMoSS study. While data were collected from over 130 participants, initial problems with data collection meant that suitable data for analysis were only available from 78 participants, 59 of which provided more than 5 weeks of usable data for analysis. The recorded data were also found to be very noisy with variable levels of data coverage. While this was likely partially due to technological and data collection issues that will be resolved as technology improves, for the purpose of analysis, significant preprocessing techniques had to be developed to handle the data. This included filtering noise and filling missing data where possible. These were developed using heuristics, and while they worked effectively for the purposes required, there is likely scope for improvement as outlined in section 8.3.1 below. A new method for reliably clustering stationary places was developed, which ensured the reliability of the derived features. This was shown to provide suitable clusters when applied to geolocation data, and is likely to be applicable to a range of geolocation data clustering tasks.

Features were extracted from the preprocessed geographic location coordinates and extracted location clusters based on previous work, with new features developed to better quantify the daily regularity of individuals. The new features developed were subtle variations on existing features, but provided improved individual performance. Overall, the extracted features performed as expected. They are relatively simple features, but with their simplicity comes interpretability and understandability.

To reduce noise between features calculated on consecutive weeks, allowing for missing weeks, a Kalman filter-based technique was developed. Filtering the extracted

features was shown to improve the discriminative performance of the extracted features, demonstrating the importance of considering noise in the feature space. While filtering noise in the input data is commonly applied in the literature, filtering in the feature space is not so common, but when working with longitudinal data this appears to be a potentially important step to optimise performance. The method developed uses an adaptive Kalman filter, trained on window-filtered feature data. The filter therefore emulates window filtering without requiring future data, and in the present application showed generally superior performance compared to the window filter. This is likely to be applicable to many other applications with noisy time-series features.

Analysis was first performed on the BD participants, where statistical differences were found between weeks where participants self-reported clinical-level depression and those where they did not. This matched previous results on mild depression in community samples by Saeb et al. (2015a, 2015b, 2016), but this is the first time that this length of data has been analysed in patients with bipolar disorder. Classification of depression from geolocation was demonstrated, achieving an accuracy of over 75 %, which is comparable to previous results on larger data sets.

Regression performance when estimating the level of depression from the geolocation-derived features for the same BD participants achieved an MAE of 3.6 using a quadratic logistic GLM model. It was shown that one of the factors limiting performance was that individuals naturally have different baseline values of features, but also fundamentally different behavioural traits under different levels of depression. Personalisation of models is key to optimising performance.

A new technique was therefore developed to find groups of individuals that have similar regression models linking their geolocation-derived features with their self-reported depression levels. This found that there were indeed several groups of individuals that appeared to behave similarly, and regression models trained on these grouped individuals improved on the regression results from the 'global' models. Also significant is that when individuals were assigned to groups based on limited calibration data from the start of the recording, the regression performance in the group

was significantly better than training fully personalised models only on the limited calibration data. This shows that finding groups of models within the population that relate behavioural markers to mental state may be a promising approach. With more data, the quality of the extracted groups should improve.

The existence of groups of individuals should not be surprising given the results presented by Servia-Rodríguez et al. (2017) (see section 2.2.4.9 for a description of their study). Servia-Rodríguez et al. found that there were significant correlations between most of the demographic features for the 18 000 participants analysed and the objective data recorded. For example, physical activity levels measured in the evening were found to be correlated with gender, age, and occupational status, as well as personality type. This means that in their analysis, there were at least different baselines values of activity for each of the demographic groups. While this is valuable, it most likely hides an even more complex reality of correlations of objective features with combinations of demographics. Analysis such as the one performed by Servia-Rodríguez et al. is also fundamentally limited by the demographic data available. Non-parametric clustering techniques have the ability to reveal additional structure that may not be otherwise accessible, as demonstrated in this thesis.

The grouped personalised model was developed and tested for mental state prediction. The core of the model is regularised regression, and while the model presented is based on first-order linear regression, this can potentially be extended to more complex underlying models if needed (see section 8.3.3.1 below for details). Being based on linear regression means that the grouping model likely has further applications outside mental state prediction. The model is particularly suited to longitudinal applications with multiple data points per individual, and mental state prediction is highly suitable in that regard. However, as discussed in section 8.3.2 below, it does not incorporate time-series dynamics so can be applied to any situation where there are a range of predictors and it is hypothesised that different subsets of predictors are relevant to the model output for different data points. Having multiple data points per individual/item is helpful because it enables the model to better handle noise in

the dependent and independent variables, but it may not be necessary if the noise level is low and only the groupings are of interest.

A final important result is the significant improvement in performance from testing fully personalised models on all data from an individual using cross-validation by leaving out random weeks of data compared to the model trained on just calibration data from the start of the recording. In practice, data will likely only be available from before the time period being estimated. Therefore, future data should not be used to make predictions about past events. Yet this is commonly seen in the literature, so these results should be treated with caution. It also suggests that using a grouped model to augment the limited calibration data available from an individual is a promising approach.

8.2 Limitations

Although the data recorded in the AMoSS study forms one of the largest datasets of longitudinal objective data collected from psychiatric patients in the community, one of the main limitations of the data is the range of QIDS scores available for each participant. Of the 59 individuals available who provided more than 5 weeks of data, the majority (41 participants) had a range of QIDS scores less than 10, and only 7 had a range of QIDS scores greater than 15. The value of the models trained on individuals with a low range of QIDS scores is limited because the models tend to find minor fluctuations from the mean value. A key limitation therefore is the amount of data with a wide range of QIDS scores available for analysis. One of the reasons why this is important is because one of the main areas of interest in using objective markers to monitor mental state is to identify the transitions between mood states. Detailed investigation of state transitions compared to stable states may provide useful data about which variations in behaviour are normal for an individual in remission (as in BD) and which variations are significantly correlated with the onset of illness episodes. Separating the two will always be crucial for predictive models to perform adequately.

The limitations of using QIDS scores as indicators of mental state have been discussed in detail in sections 2.2.1 and 4.4.3. At best they can be considered subjective proxies of mental state. This is an accepted limitation that affects much of the work in this field.

Some of the methods described above also have their limitations. The Kalman-filter feature smoothing method effectively filters the features and improved discriminative performance is demonstrated in the results. However, it ignores any underlying dynamics in the features that could be included in the filtering model. This is discussed in greater detail in section 8.3.2 below.

A limitation of the grouped personalised model is that it assumes that individuals always remain in the same group. This limitation is explored in greater detail in section 8.3.3.3 below.

8.3 Future work

Given the limitations with the data outlined above, in order to develop these techniques further, it is important to collect more data from individuals with a range of levels of depression. This will enable further exploration of the different types of episodes that individuals may suffer, and greater individualisation of models. This will also enable investigation of other aspects of behavioural response to mental illness such as model switching in individuals, or the effects of different pharmaceutical treatments.

Additionally, there are several specific areas in this thesis that could be developed, as detailed in the following sections.

8.3.1 Improved data preprocessing

Methods for pre-processing of the raw recorded geolocation coordinates, in particular data filtering and imputation, were presented in sections 5.3 and 5.5 respectively. They were important methods to apply to the input data to enable the extracted

features to accurately represent the behaviour of the individual, as was demonstrated in section 5.5.3.

The methods developed and presented were based on heuristics found to be suitable given the data available for analysis. However, there are several ways in which these methods could potentially be improved as detailed in the following sections.

8.3.1.1 Data filtering

The data filtering method presented in section 5.3 used thresholds on the speed travelled between points to identify location coordinates that were likely to be noise. This method was based on specific features found in the data and proved relatively successful.

When providing location coordinates, Android also provides an estimate of the standard deviation of the location point. This accuracy estimate of each point was not recorded on the custom app used for data collection on the AMoSS study, but recording it in future studies may enable a range of other techniques to be applied for filtering.

One such method might be Kalman filtering of the recorded coordinates, similar to the technique presented in section 6.1.2.2 for filtering of the extracted feature values. There are several potential ways to implement a Kalman filter for geolocation data filtering in this context. The simplest implementation would be to assume that the next location is the same as the previous location with an appropriate level of noise. In a more advanced implementation, the current trajectory of travel can be used to predict the next location, again with an appropriate level of noise. In either implementation, the noise in the measured coordinate would be the accuracy provided by Android.

An alternative option is to fit a smoothing spline to the recorded data points, with the spline being fit to the measured coordinates weighted by the estimated accuracy level.

8.3.1.2 Data imputation

The data imputation method presented in section 5.5 used a number of heuristic rules developed from inspection of the missing data, as described in detail in section 5.5.3. This imputation was based only on the values of the geolocation data surrounding the missing segment. However, the battery and activity data shown in figures 5.2 and 5.3 indicate a correlation between these data and travelling state. For example, stationary periods tend to be associated with lower activity levels. Additionally, when the phone is plugged into the charger, activity levels are usually low and the phone probably also stationary.

Increased accuracy of imputation could therefore possibly be obtained by utilising the data from the other sources collected, in a similar way to Yue et al. (2017). This additional information could either be developed into additional heuristic rules, or combined in a dynamic model such as a Kalman filter.

8.3.2 Incorporation of time-series dynamics

The results in both chapters 6 and 7 were presented as independent estimates of depression for calendar weeks of geolocation data. However, the data examples shown in figure 7.2 indicate that levels of depression follow a longitudinal course and are not independent week-to-week. This fits with the wider literature, which indicates that depression levels do not tend to fluctuate wildly (Solomon et al., 2010). Even if they do, depression is only considered clinically relevant if symptoms persist for two or more weeks (see section 2.1.1.1 for an overview of the clinical definition of depression).

Time-series dynamics exist in several parts of the analysis pathway presented in this thesis. In the raw geolocation data, time-series dynamics exist in the coordinates recorded from the phone. The filtering of noise in the preprocessing of the geolocation data described in section 5.3 uses heuristic rules, but time-series methods such as Kalman filtering or functional methods such as spline smoothing may provide alternative methods that incorporate likely dynamics as described in section 8.3.1.

The features extracted from the preprocessed geolocation data were Kalman filtered as described in section 6.1.2.2. This was shown in section 6.1.4 to improve the discriminative performance of the extracted features. However, the form of the Kalman filter model applied in equations (6.5a) and (6.5b) ignores time-series dynamics that are likely to exist in the features and simply smooths their values. As was discussed in section 6.1.2.2, it is unclear whether incorporating a model of feature dynamics would improve results or impose unrealistic assumptions on the feature space. Although it would most likely require more data than is currently available, dynamics in the feature values should be investigated further. The technique developed in section 6.1.2.2 used a window filtered signal to train a Kalman filter. Considering the feature values instead as a functional system, and fitting a smoothing spline to the calculated values, may reveal functional correlations between the feature values and mental state. This can in turn be used to develop refined dynamical models of feature values, which can be used to improve model performance, but may also have clinical applications by predicting future changes in feature values and mental state, perhaps enabling more appropriate treatment plans.

Finally, there are several ways in which time-series dynamics in depression levels can be incorporated into both the 'global' and personalised models presented in this thesis. At the simplest level, the current estimate could be used to place a prior on the possible values for the next time point, similar to the Kalman filter described in section 6.1.2.2. This would reduce the likelihood of unrealistically large swings in depression-level estimates. More advanced methods may include using Gaussian processes (see section 7.2) to model the week-to-week dynamics.

An end-to-end functional system, incorporating functional filtering of raw data and feature values, and dynamical modelling of mental state (while still allowing personalisation) would likely provide improved performance over the results presented here.

8.3.3 Improvements to the grouped personalised model

The grouped personalised model was demonstrated to provide significantly improved performance over 'global' models and fully personalised models trained only on limited calibration data used to allocate individuals into groups. However, the model was presented as a proof-of-concept, and there are several areas that could potentially be developed as detailed in the following sections.

8.3.3.1 Generalisation of the model

The core model is currently based on first-order linear regression. While this is a reasonable initial model to test, optimal performance using global models for regression was obtained using a quadratic logistic GLM, and it is possible that generalisation of the grouped model could improve results. Generalisations of the Bayesian Lasso model that forms the core of the grouped model have been demonstrated. Leng et al. (2014) presented a GLM generalisation of the Bayesian Lasso using a least squares approximation originally developed by Wang & Leng (2007) in a non-Bayesian context. This could be applied to the grouped personalised model to enable, for example, logistic outputs.

8.3.3.2 Improved allocation of individuals

Allocation of new individuals into groups remains a challenge in the grouped model. The method presented used the predictive performance of the grouped models on calibration data for an individual to determine the optimal group. However, this is a very simple approach and incorporating other metrics of similarity, for example sensor values and demographics, may improve results.

8.3.3.3 Handling switching of models

Related to group allocation, one key limitation of the grouped personalised model presented here is that it assumes that individuals always remain in the same group. In reality, individuals may exhibit temporal variability in their response to illness. The model is designed to handle inter-individual variability in regression models, but

not intra-individual variability. This may arise in many ways. For example, in some individuals improvements in mental state may follow different models to deterioration, or the same individual may have different relationships with objective features under different episodes.

This is a challenging problem and will likely require significantly more data to first converge towards an exhaustive set of groupings. This in itself should improve group allocation and performance, but also the characteristics of individuals in each of the groups can be explored in greater detail.

8.3.4 End-to-end Bayesian model

The current modelling framework uses point-estimates of the features as observed values. The point-estimates of the feature values rely on the availability of the underlying data, which necessitated the development of the data filtering and imputation methods presented in sections 5.3 and 5.5 respectively. Section 8.3.1 has described potential enhancements to these methods, but an alternative approach is to develop an 'end-to-end' Bayesian model that incorporates uncertainty in the feature values. This could be applied naturally into either a 'global' or personalised Bayesian modelling framework.

In this end-to-end model, the location coordinates and features would all be modelled as random variables, which would therefore have distributions and an element of uncertainty associated with them. This could eliminate the need to impute raw location coordinates because the missing data could be incorporated into the distributions of the features.

There are several ways in which such a model could be realised. One way is demonstrated in the graphical model in figure 8.1 on page 277, where $\mathbf{x}_t$, $\mathbf{y}_t$, and $\mathbf{z}_t$ represent data sources (i.e. location coordinates, accelerometer levels, battery levels, etc.) at time t; u_t represents the location (cluster) at time t; and f_1, f_2, and f_3 represent three feature values calculated from the location clusters (features could also be calculated directly from the location coordinates, but the model design would be slightly different).

In this model, the location cluster assignment at each time point, t, denoted by u_t, is modelled using a non-parametric approach similar to the location clustering presented in section 5.9. At any time point, there would be a probability of the individual being in any of the previously identified location clusters, or transitioning between locations. Where data are available for a given time point, u_t would be set from the received data. New location clusters can also be created if required.

Where there are no data for a time point, the individual could be either in the last-recorded location; in the next-recorded location; transitioning between locations; or at another (perhaps new) location. These options can all be modelled as probabilities of the state of u_t, possibly informed by the other sensor values as outlined in section 8.3.1 and indicated by the conditions on $\mathbf{x}_t$, $\mathbf{y}_t$, and $\mathbf{z}_t$ in figure 8.1. Note that the model in figure 8.1 is presented as Markovian, but in practice this might not be sufficient.

A maximum likelihood estimate of the state sequence $u_1, u_2, \ldots, u_T$ can be obtained, in a similar way to how inference is performed in a hidden Markov model, from which feature values can be calculated directly. Alternatively, the posterior distribution of state sequences could be marginalised over $u_1, u_2, \ldots, u_T$ to obtain full distributions for the features with confidence intervals.

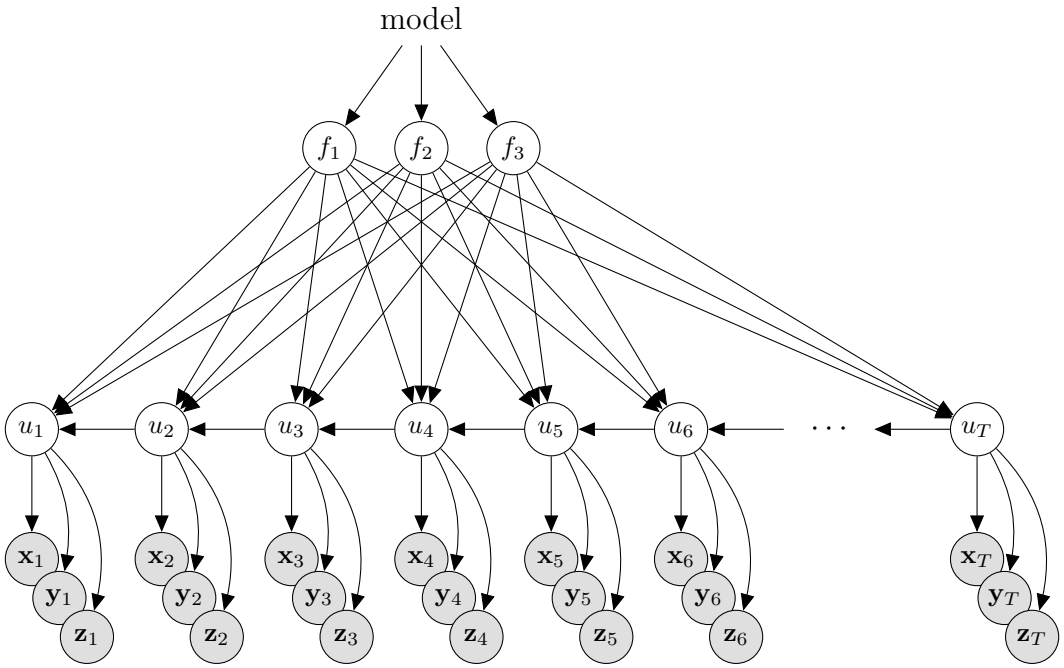


Figure 8.1: Proposed end-to-end Bayesian model. Model variables $\mathbf{x}_t$, $\mathbf{y}_t$, and $\mathbf{z}_t$ represent data sources (i.e. location coordinates, accelerometer levels, battery levels, etc.) at time t; u_t represents the location cluster at time t; and f_1, f_2, and f_3 represent three features.

Appendix A

Data Imputation Algorithms

This appendix gives details of the data imputation methods described in section 5.5. Algorithm A.1 describes the rules used to select the appropriate data imputation method, as given in table 5.1, repeated in table A.1 below. The three data imputation methods given in section 5.5.2 are described in algorithms A.2 to A.4.

Table A.1: Missing data imputation method rules (repeated from table 5.1).

Distance criteria	Time criteria	Imputation method
$\mathbf{d}^{\Delta} < 1\,\text{km}$	$t^{\Delta} \leq 6\,\text{h}$ **or** $t^{-} \geq 9\,\text{P.M.}$ **and** $t^{\Delta} \leq 12\,\text{h}$	Midpoint imputation (algorithm A.2)
$\|\|\mathbf{d}^{-}\|\| > 750\,\text{m}$ **and** $\|\|\mathbf{d}^{+}\|\| > 750\,\text{m}$	$t^{\Delta} \leq 6\,\text{h}$ **or** $t^{-} \geq 9\,\text{P.M.}$ **and** $t^{\Delta} \leq 12\,\text{h}$	
$\|\|\mathbf{d}^{-}\|\| < 250\,\text{m}$ **and** $\|\|\mathbf{d}^{+}\|\| < 250\,\text{m}$	$t^{-} \geq 9\,\text{P.M.}$ **and** $t^{\Delta} \leq 18\,\text{h}$	
$\|\|\mathbf{d}^{-}\|\| > 750\,\text{m}$ **and** $\|\|\mathbf{d}^{+}\|\| < 750\,\text{m}$	$t^{\Delta} \leq 6\,\text{h}$ **or** $t^{-} \geq 8\,\text{P.M.}$ **and** $t^{\Delta} \leq 18\,\text{h}$	Travelling to home (algorithm A.3)
$\|\|\mathbf{d}^{-}\|\| < 750\,\text{m}$ **and** $\|\|\mathbf{d}^{+}\|\| > 750\,\text{m}$	$t^{\Delta} \leq 6\,\text{h}$ **or** $t^{-} \geq 8\,\text{P.M.}$ **and** $t^{\Delta} \leq 18\,\text{h}$	Travelling from home (algorithm A.4)

$t^{-}\,/\,t^{+}$	Time of data point immediately before/after missing section.
$\mathbf{d}^{-}\,/\,\mathbf{d}^{+}$	Data coordinates immediately before/after missing section.
$\|\|\mathbf{d}^{-\,/\,+}\|\|$	Distance from home to coordinates immediately before/after missing section.
$t^{\Delta} := t^{+} - t^{-}$	Length of missing section.
$\mathbf{d}^{\Delta} := \|\|\mathbf{d}^{+} - \mathbf{d}^{-}\|\|$	Euclidean distance between locations before and after section.

Algorithm A.1 Data imputation method selection rules.

data: $t^- \,/\, t^+$ Time of data point immediately before/after missing section.
$\mathbf{d}^- \,/\, \mathbf{d}^+$ Data coordinates immediately before/after missing section.

begin
  $t^\Delta \leftarrow t^+ - t^-$ ▷ Length of missing section.
  $\mathbf{d}^\Delta \leftarrow ||\mathbf{d}^+ - \mathbf{d}^-||$ ▷ Distance between locations before and after section.
  if ($\mathbf{d}^\Delta < 1\,\text{km}$) **and** ($t^\Delta \leq 6\,\text{h}$ **or** ($t^- \geq 9\,\text{P.M.}$ **and** $t^\Delta \leq 12\,\text{h}$)) **then**
    ImputeMidpoint($t^-, t^+, \mathbf{d}^-, \mathbf{d}^+$) ▷ Algorithm A.2.
  else if ($||\mathbf{d}^-|| > 0.75\,\text{km}$) **and** ($||\mathbf{d}^+|| > 0.75\,\text{km}$) **and**
      ($t^\Delta \leq 6\,\text{h}$ **or** ($t^- \geq 9\,\text{P.M.}$ **and** $t^\Delta \leq 12\,\text{h}$)) **then**
    ImputeMidpoint($t^-, t^+, \mathbf{d}^-, \mathbf{d}^+$) ▷ Algorithm A.2.
  else if ($||\mathbf{d}^-|| < 250\,\text{m}$) **and** ($||\mathbf{d}^+|| < 250\,\text{m}$) **and**
      ($t^- \geq 9\,\text{P.M.}$ **and** $t^\Delta \leq 18\,\text{h}$) **then**
    ImputeMidpoint($t^-, t^+, \mathbf{d}^-, \mathbf{d}^+$) ▷ Algorithm A.2.
  else if ($||\mathbf{d}^-|| > 0.75\,\text{km}$) **and** ($||\mathbf{d}^+|| < 0.75\,\text{km}$) **and**
      ($t^\Delta \leq 6\,\text{h}$ **or** ($t^- \geq 8\,\text{P.M.}$ **and** $t^\Delta \leq 18\,\text{h}$)) **then**
    ImputeTravellingToHome($t^-, t^+, \mathbf{d}^-, \mathbf{d}^+$) ▷ Algorithm A.3.
  else if ($||\mathbf{d}^-|| < 0.75\,\text{km}$) **and** ($||\mathbf{d}^+|| > 0.75\,\text{km}$) **and**
      ($t^\Delta \leq 6\,\text{h}$ **or** ($t^- \geq 8\,\text{P.M.}$ **and** $t^\Delta \leq 18\,\text{h}$)) **then**
    ImputeTravellingFromHome($t^-, t^+, \mathbf{d}^-, \mathbf{d}^+$) ▷ Algorithm A.4.
  end if
end

Algorithm A.2 Data imputation method – midpoint imputation.

data: $t^- \,/\, t^+$ Time of data point immediately before/after missing section.
$\mathbf{d}^- \,/\, \mathbf{d}^+$ Data coordinates immediately before/after missing section.

function ImputeMidpoint($t^-, t^+, \mathbf{d}^-, \mathbf{d}^+$)
  $t^\Delta \leftarrow t^+ - t^-$ ▷ Length of missing section.
  $\bar{\mathbf{d}} \leftarrow (\mathbf{d}^+ + \mathbf{d}^-) \,/\, 2$ ▷ Midpoint between $\mathbf{d}^-$ and $\mathbf{d}^+$.
  $t^* \leftarrow ||\bar{\mathbf{d}} - \mathbf{d}^-|| \,/\, 80\,\mathrm{km}$ ▷ Time to travel from $\mathbf{d}^-$ to $\bar{\mathbf{d}}$ at $80\,\mathrm{km\,h^{-1}}$.
  if $t^\Delta \leq 2t^*$ **then**
    Fill t^- to t^+ with evenly spaced points between $\mathbf{d}^-$ and $\mathbf{d}^+$.
  else
    Fill t^- to $t^- + t^*$ with evenly spaced points between $\mathbf{d}^-$ and $\bar{\mathbf{d}}$.
    Fill $t^- + t^*$ to $t^+ - t^*$ with $\bar{\mathbf{d}}$.
    Fill $t^+ - t^*$ to t^+ with evenly spaced points between $\bar{\mathbf{d}}$ and $\mathbf{d}^+$.
  end if
end function

Algorithm A.3 Data imputation method – travelling to home.

data: $t^- \,/\, t^+$ Time of data point immediately before/after missing section.
$\mathbf{d}^- \,/\, \mathbf{d}^+$ Data coordinates immediately before/after missing section.

function ImputeTravellingToHome($t^-, t^+, \mathbf{d}^-, \mathbf{d}^+$)
  $t^\Delta \leftarrow t^+ - t^-$ ▷ Length of missing section.
  $t^* \leftarrow ||\mathbf{d}^-|| \,/\, 80\,\mathrm{km}$ ▷ Time to travel from $\mathbf{d}^-$ to home at $80\,\mathrm{km\,h^{-1}}$.
  if $t^\Delta \leq t^*$ **then**
    Fill t^- to t^+ with evenly spaced points between $\mathbf{d}^-$ and $\mathbf{d}^+$.
  else if $t^+ - t^- - t^* \leq 2\,\mathrm{h}$ **then**
    Fill t^- to $t^+ - t^*$ with $\mathbf{d}^-$.
    Fill $t^+ - t^*$ to t^+ with evenly spaced points between $\mathbf{d}^-$ and $\mathbf{d}^+$.
  else
    Fill t^- to $t^- + 2\,\mathrm{h}$ with $\mathbf{d}^-$.
    $t^- \leftarrow t^- + 2\,\mathrm{h}$
    if $t^+ - t^- - t^* - 2\,\mathrm{h} > 30\,\mathrm{min}$ **then**
      $t^\dagger \leftarrow \min(t^+ - t^- - t^* - 2\,\mathrm{h}, 2\,\mathrm{h})$
      Fill $t^+ - t^\dagger$ to t^+ with $\mathbf{d}^+$.
      $t^+ \leftarrow t^\dagger$
    end if
    $t^\dagger \leftarrow (t^+ - t^-) \,/\, 2$
    Fill $t^\dagger - t^*/2$ to $t^\dagger + t^*/2$ with evenly spaced points between $\mathbf{d}^-$ and $\mathbf{d}^+$.
  end if
end function

Algorithm A.4 Data imputation method – travelling from home.

data: $t^- \,/\, t^+$ Time of data point immediately before/after missing section.
$\mathbf{d}^- \,/\, \mathbf{d}^+$ Data coordinates immediately before/after missing section.

function ImputeTravellingFromHome($t^-, t^+, \mathbf{d}^-, \mathbf{d}^+$)
    $t^\Delta \leftarrow t^+ - t^-$ ▷ Length of missing section.
    $t^* \leftarrow ||\mathbf{d}^+|| \,/\, 80\,\mathrm{km}$ ▷ Time to travel from home to $\mathbf{d}^+$ at $80\,\mathrm{km\,h^{-1}}$.
    if $t^\Delta \leq t^*$ **then**
        Fill t^- to t^+ with evenly spaced points between $\mathbf{d}^-$ and $\mathbf{d}^+$.
    else if $t^+ - t^- - t^* \leq 2\,\mathrm{h}$ **then**
        Fill t^- to $t^- + t^*$ with evenly spaced points between $\mathbf{d}^-$ and $\mathbf{d}^+$.
        Fill $t^- + t^*$ to t^+ with $\mathbf{d}^+$.
    else
        Fill $t^+ - 2\,\mathrm{h}$ to t^+ with $\mathbf{d}^+$.
        $t^+ \leftarrow t^+ - 2\,\mathrm{h}$
        if $t^+ - t^- - t^* - 2\,\mathrm{h} > 30\,\mathrm{min}$ **then**
            $t^\dagger \leftarrow \min(t^+ - t^- - t^* - 2\,\mathrm{h}, 2\,\mathrm{h})$
            Fill t^- to $t^- + t^\dagger$ with $\mathbf{d}^-$.
            $t^- \leftarrow t^\dagger$
        end if
        $t^\dagger \leftarrow (t^+ - t^-) \,/\, 2$
        Fill $t^\dagger - t^*/2$ to $t^\dagger + t^*/2$ with evenly spaced points between $\mathbf{d}^-$ and $\mathbf{d}^+$.
    end if
end function

Appendix B

Additional Feature Statistics

This appendix gives additional feature statistics for the features described in chapter 5, in particular sections 5.10 and 5.11.

B.1 Feature correlation coefficients

Table B.1 across the following two pages gives correlation coefficients for all features described in section 5.10 calculated on the different data subsets described in section 5.11 for BD participants. Only statistically significant correlations ($p < 0.001$) are shown and the absolute strength of the correlation is indicated by the shading.

Table B.1: Feature correlation coefficients for BD participants.

		Base										Weekday										Weekend									
		ENT	NENT	LV	HS	TT	TD	NC	DM	DMN	DMD	ENT	NENT	LV	HS	TT	TD	NC	DM	DMN	DMD	ENT	NENT	LV	HS	TT	TD	NC	DM	DMN	DMD
Base	ENT		0.88	0.51	-0.93	0.59	0.49	0.70	0.58	0.45	0.50	0.96	0.77	0.55	-0.87	0.57	0.47	0.67	0.63	0.48	0.55	0.79	0.68	0.46	-0.73	0.52	0.31	0.68	0.48	0.33	0.37
	NENT			0.41	-0.94	0.36	0.32	0.36	0.55	0.59	0.57	0.85	0.93	0.46	-0.92	0.36	0.33	0.34	0.60	0.61	0.61	0.65	0.67	0.32	-0.62	0.29	0.14	0.43	0.35	0.33	0.35
	LV				-0.44	0.49	0.76	0.56	0.94	0.21	0.22	0.55	0.40	0.92	-0.47	0.49	0.72	0.55	0.88	0.26	0.27	0.23	0.17	0.55	-0.22	0.38	0.49	0.36	0.51	0.22	0.30
	HS					-0.39	-0.38	-0.47	-0.56	-0.55	-0.54	-0.88	-0.84	-0.49	0.95	-0.37	-0.38	-0.43	-0.61	-0.58	-0.59	-0.73	-0.70	-0.39	0.73	-0.35	-0.21	-0.52	-0.42	-0.31	-0.34
	TT						0.61	0.81	0.42		0.21	0.57	0.28	0.48	-0.32	0.98	0.53	0.79	0.41		0.24	0.54	0.34	0.51	-0.45	0.83	0.52	0.70	0.48	0.21	0.29
	TD							0.67	0.70		0.22	0.56	0.32	0.70	-0.40	0.64	0.96	0.69	0.66	0.17	0.27	0.20		0.41	-0.22	0.41	0.66	0.38	0.37	0.15	0.22
	NC								0.49		0.21	0.69	0.27	0.55	-0.41	0.80	0.60	0.97	0.50		0.28	0.56	0.32	0.52	-0.45	0.66	0.54	0.80	0.49	0.24	0.31
	DM									0.52	0.50	0.64	0.55	0.89	-0.61	0.42	0.68	0.48	0.94	0.53	0.50	0.27	0.25	0.55	-0.25	0.32	0.43	0.34	0.54	0.29	0.35
	DMN										0.83	0.49	0.64	0.25	-0.62		0.15		0.52	0.91	0.75	0.20	0.30	0.18	-0.18				0.23	0.29	0.29
	DMD											0.52	0.60	0.28	-0.58	0.22	0.21	0.18	0.51	0.78	0.88	0.31	0.34	0.34	-0.25	0.13		0.24	0.39	0.35	0.38
Weekday	ENT												0.82	0.61	-0.90	0.59	0.56	0.69	0.69	0.51	0.57	0.60	0.51	0.38	-0.54	0.40	0.31	0.57	0.39	0.28	0.33
	NENT													0.47	-0.90	0.31	0.34	0.25	0.61	0.65	0.62	0.43	0.46	0.22	-0.39	0.15		0.32	0.26	0.28	0.32
	LV														-0.53	0.49	0.72	0.56	0.95	0.27	0.27	0.26	0.19	0.41	-0.23	0.36	0.32	0.37	0.37	0.16	0.23
	HS															-0.33	-0.42	-0.39	-0.67	-0.64	-0.62	-0.53	-0.53	-0.29	0.50	-0.20	-0.18	-0.41	-0.32	-0.26	-0.30
	TT																0.57	0.81	0.42		0.25	0.44	0.27	0.44	-0.35	0.71	0.49	0.62	0.42	0.18	0.26
	TD																	0.64	0.68	0.18	0.24	0.14		0.27	-0.17	0.29	0.43	0.32	0.24		0.16
	NC																		0.50		0.26	0.47	0.26	0.44	-0.38	0.59	0.53	0.67	0.40	0.19	0.26
	DM																			0.54	0.51	0.30	0.26	0.43	-0.26	0.30	0.31	0.36	0.41	0.23	0.30
	DMN																				0.86	0.23	0.30	0.23	-0.21				0.27	0.33	0.36
	DMD																					0.34	0.35	0.38	-0.29	0.15	0.24	0.27	0.41	0.40	0.46
Weekend	ENT																						0.87	0.52	-0.92	0.68	0.23	0.77	0.55	0.37	0.37
	NENT																							0.50	-0.85	0.46	0.13	0.48	0.54	0.40	0.37
	LV																								-0.47	0.57	0.56	0.53	0.99	0.64	0.67
	HS																									-0.61	-0.24	-0.61	-0.49	-0.32	-0.30
	TT																										0.49	0.75	0.55	0.24	0.30
	TD																											0.35	0.54	0.20	0.25
	NC																												0.52	0.31	0.36
	DM																													0.66	0.66
	DMN																														0.95
	DMD																														
Median	ENT																														
	NENT																														
	LV																														
	HS																														
	TT																														
	TD																														
	NC																														
	DM																														
	DMN																														
	DMD																														
Optimised	ENT																														
	NENT																														
	LV																														
	HS																														
	TT																														
	TD																														
	NC																														
	DM																														
	DMN																														
	DMD																														

(continued on the next page)

Table B.1 (cont.): Feature correlation coefficients for BD participants.

		Median										Optimised										Questionnaires		
		ENT	NENT	LV	HS	TT	TD	NC	DM	DMN	DMD	ENT	NENT	LV	HS	TT	TD	NC	DM	DMN	DMD	ASRM	GAD-7	QIDS
Base	ENT	1.00	0.87	0.49	-0.93	0.59	0.47	0.72	0.57	0.44	0.50	0.97	0.83	0.62	-0.89	0.59	0.53	0.73	0.70	0.54	0.59		-0.40	-0.33
	NENT	0.88	1.00	0.40	-0.93	0.35	0.30	0.37	0.53	0.59	0.57	0.88	0.95	0.55	-0.91	0.38	0.42	0.42	0.68	0.65	0.63		-0.35	-0.28
	LV	0.51	0.41	1.00	-0.45	0.50	0.75	0.56	0.95	0.22	0.23	0.41	0.34	0.75	-0.35	0.39	0.56	0.53	0.69	0.26	0.32	-0.17	-0.26	-0.30
	HS	-0.93	-0.93	-0.43	1.00	-0.39	-0.36	-0.49	-0.54	-0.55	-0.54	-0.92	-0.90	-0.56	0.97	-0.41	-0.44	-0.50	-0.68	-0.66	-0.63	0.14	0.35	0.31
	TT	0.59	0.35	0.48	-0.39	1.00	0.57	0.81	0.42		0.21	0.55	0.34	0.54	-0.34	0.97	0.69	0.81	0.47		0.23		-0.39	-0.38
	TD	0.49	0.31	0.76	-0.38	0.62	1.00	0.67	0.71	0.13	0.23	0.43	0.28	0.57	-0.32	0.53	0.66	0.65	0.53	0.18	0.32		-0.32	-0.27
	NC	0.70	0.34	0.56	-0.46	0.81	0.65	1.00	0.50		0.22	0.65	0.34	0.58	-0.42	0.77	0.56	0.97	0.53		0.31		-0.38	-0.37
	DM	0.58	0.54	0.93	-0.56	0.42	0.69	0.50	1.00	0.53	0.51	0.52	0.49	0.79	-0.50	0.34	0.55	0.50	0.81	0.51	0.51	-0.17	-0.29	-0.33
	DMN	0.44	0.59	0.20	-0.54				0.50	1.00	0.84	0.52	0.61	0.39	-0.62		0.16		0.60	0.87	0.67		-0.15	
	DMD	0.50	0.57	0.20	-0.53	0.20	0.19	0.23	0.47	0.83	1.00	0.59	0.62	0.53	-0.62	0.24	0.42	0.31	0.69	0.72	0.76		-0.26	-0.22
Weekday	ENT	0.96	0.85	0.54	-0.88	0.57	0.54	0.70	0.62	0.48	0.53	0.93	0.81	0.64	-0.85	0.57	0.55	0.73	0.75	0.58	0.60	-0.13	-0.40	-0.33
	NENT	0.77	0.93	0.39	-0.84	0.28	0.30	0.29	0.54	0.63	0.61	0.78	0.93	0.52	-0.84	0.31	0.40	0.33	0.69	0.69	0.63		-0.34	-0.27
	LV	0.55	0.46	0.92	-0.50	0.49	0.69	0.56	0.89	0.26	0.29	0.46	0.40	0.79	-0.41	0.40	0.58	0.54	0.72	0.26	0.33	-0.18	-0.28	-0.31
	HS	-0.87	-0.91	-0.46	0.95	-0.31	-0.39	-0.43	-0.59	-0.62	-0.58	-0.87	-0.88	-0.55	0.94	-0.34	-0.41	-0.45	-0.72	-0.73	-0.67	0.15	0.36	0.32
	TT	0.57	0.34	0.49	-0.37	0.98	0.60	0.80	0.43		0.23	0.54	0.33	0.54	-0.33	0.96	0.71	0.81	0.49		0.25		-0.37	-0.36
	TD	0.47	0.32	0.72	-0.37	0.53	0.97	0.60	0.68	0.16	0.23	0.42	0.29	0.53	-0.33	0.44	0.60	0.59	0.51	0.20	0.29		-0.30	-0.24
	NC	0.66	0.32	0.55	-0.42	0.80	0.68	0.97	0.48		0.19	0.61	0.31	0.56	-0.38	0.76	0.57	0.95	0.51		0.29		-0.35	-0.33
	DM	0.63	0.59	0.88	-0.61	0.42	0.65	0.51	0.94	0.52	0.52	0.57	0.55	0.80	-0.56	0.35	0.55	0.51	0.82	0.50	0.52	-0.17	-0.33	-0.35
	DMN	0.48	0.61	0.25	-0.57		0.16		0.51	0.92	0.79	0.54	0.63	0.39	-0.63		0.15		0.60	0.86	0.77		-0.24	-0.19
	DMD	0.55	0.60	0.25	-0.58	0.23	0.25	0.29	0.48	0.76	0.88	0.62	0.64	0.53	-0.64	0.26	0.38	0.36	0.69	0.73	0.88		-0.31	-0.26
Weekend	ENT	0.79	0.65	0.22	-0.73	0.53	0.18	0.57	0.26	0.20	0.31	0.76	0.63	0.39	-0.69	0.54	0.32	0.56	0.38	0.26	0.36		-0.27	-0.26
	NENT	0.69	0.67	0.16	-0.71	0.34		0.32	0.23	0.30	0.33	0.69	0.66	0.37	-0.69	0.37	0.25	0.35	0.39	0.34	0.39			
	LV	0.46	0.32	0.55	-0.39	0.51	0.38	0.51	0.56	0.19	0.34	0.43	0.35	0.66	-0.36	0.46	0.46	0.51	0.58	0.21	0.39		-0.30	-0.35
	HS	-0.73	-0.62	-0.21	0.73	-0.44	-0.20	-0.46	-0.24	-0.18	-0.24	-0.68	-0.59	-0.35	0.68	-0.45	-0.34	-0.47	-0.31	-0.23	-0.31		0.19	0.17
	TT	0.52	0.29	0.38	-0.36	0.84	0.37	0.65	0.33		0.13	0.46	0.27	0.44	-0.28	0.79	0.51	0.64	0.31			0.13	-0.35	-0.35
	TD	0.31	0.13	0.49	-0.21	0.53	0.64	0.54	0.44		0.13	0.28	0.13	0.39	-0.17	0.48	0.46	0.51	0.34		0.26		-0.26	-0.24
	NC	0.68	0.42	0.35	-0.52	0.70	0.36	0.80	0.34		0.24	0.64	0.43	0.42	-0.49	0.68	0.39	0.77	0.40		0.26		-0.41	-0.44
	DM	0.48	0.35	0.51	-0.42	0.49	0.35	0.48	0.54	0.24	0.39	0.46	0.39	0.64	-0.41	0.45	0.44	0.49	0.58	0.28	0.44		-0.27	-0.33
	DMN	0.33	0.33	0.22	-0.31	0.21	0.13	0.24	0.29	0.30	0.36	0.33	0.35	0.37	-0.32	0.21	0.22	0.24	0.35	0.28	0.37		-0.21	-0.24
	DMD	0.37	0.35	0.29	-0.33	0.29	0.20	0.31	0.35	0.31	0.39	0.37	0.38	0.44	-0.33	0.27	0.28	0.32	0.40	0.27	0.44		-0.28	-0.28
Median	ENT		0.87	0.50	-0.93	0.59	0.47	0.71	0.57	0.44	0.50	0.96	0.83	0.62	-0.89	0.59	0.52	0.73	0.70	0.54	0.58	-0.13	-0.40	-0.34
	NENT			0.40	-0.93	0.34	0.29	0.36	0.53	0.59	0.57	0.87	0.95	0.54	-0.91	0.37	0.42	0.40	0.68	0.65	0.62		-0.35	-0.28
	LV				-0.44	0.49	0.75	0.55	0.94	0.21	0.22	0.39	0.32	0.75	-0.34	0.38	0.55	0.52	0.68	0.25	0.31	-0.17	-0.26	-0.30
	HS					-0.39	-0.36	-0.48	-0.54	-0.54	-0.53	-0.91	-0.89	-0.55	0.96	-0.41	-0.44	-0.50	-0.67	-0.65	-0.62	0.14	0.34	0.30
	TT						0.58	0.81	0.43		0.21	0.55	0.33	0.54	-0.33	0.97	0.69	0.81	0.47		0.23		-0.39	-0.39
	TD							0.65	0.70		0.20	0.41	0.26	0.54	-0.31	0.48	0.61	0.63	0.50	0.18	0.30		-0.32	-0.27
	NC								0.50		0.23	0.66	0.36	0.58	-0.44	0.77	0.56	0.97	0.53		0.32		-0.39	-0.38
	DM									0.50	0.49	0.51	0.47	0.79	-0.48	0.35	0.55	0.50	0.79	0.49	0.49	-0.17	-0.30	-0.33
	DMN										0.83	0.52	0.61	0.39	-0.62		0.15		0.60	0.86	0.67		-0.16	-0.13
	DMD											0.59	0.62	0.54	-0.62	0.25	0.42	0.32	0.69	0.72	0.76		-0.27	-0.24
Optimised	ENT												0.87	0.58	-0.93	0.59	0.49	0.70	0.71	0.61	0.65		-0.39	-0.32
	NENT													0.57	-0.91	0.37	0.39	0.40	0.70	0.67	0.66		-0.34	-0.27
	LV														-0.52	0.50	0.71	0.62	0.90	0.38	0.52	-0.14	-0.27	-0.25
	HS															-0.38	-0.40	-0.48	-0.68	-0.73	-0.68		0.33	0.29
	TT																0.69	0.79	0.46		0.26	0.13	-0.35	-0.35
	TD																	0.66	0.65	0.14	0.35		-0.21	-0.13
	NC																		0.59	0.16	0.37		-0.34	-0.31
	DM																			0.64	0.70		-0.30	-0.29
	DMN																				0.77		-0.17	-0.17
	DMD																						-0.24	-0.22

Appendix C

Full Results for Global Models of Mental Health

This appendix shows the full results from regression and classification tests run on geolocation data as described in chapter 6.

C.1 Identification of depression

This section gives the full results from classification of depression described in section 6.3.2.

Results are shown from classification performed using logistic regression; naïve Bayes; and the quadratic discriminant analysis models. Figures C.1 to C.3 on pages 288 and 289 show the main results (accuracy; sensitivity; specificity; and F_1 score) graphically for the three classification models with leave-one-participant-out cross validation. Tables C.1 to C.3 on pages 290 and 291 show full accuracy and AUC results for the three classification models with a) leave-one-participant-out cross validation where all data from one participant are left out for training; b) 10-fold and 5-fold cross-validation where all data from multiple participants are left out for training; and c) 3-fold cross-validation where a third of all the data are left out for training (see section 6.3.1.1 for details). In each fold of cross-validation in all models, 100 iterations of group equalisation were run (see section 3.5.1 for details) and all results in tables C.1 to C.3 are presented as median $\pm$ iqr calculated across all iterations.

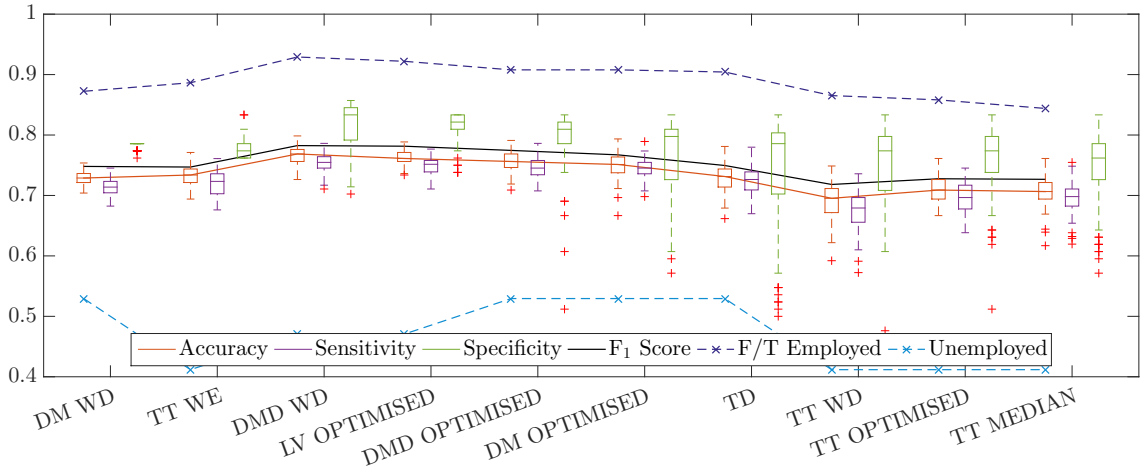


Figure C.1: Depression classification results using logistic regression classifier with leave-one-participant-out cross-validation.

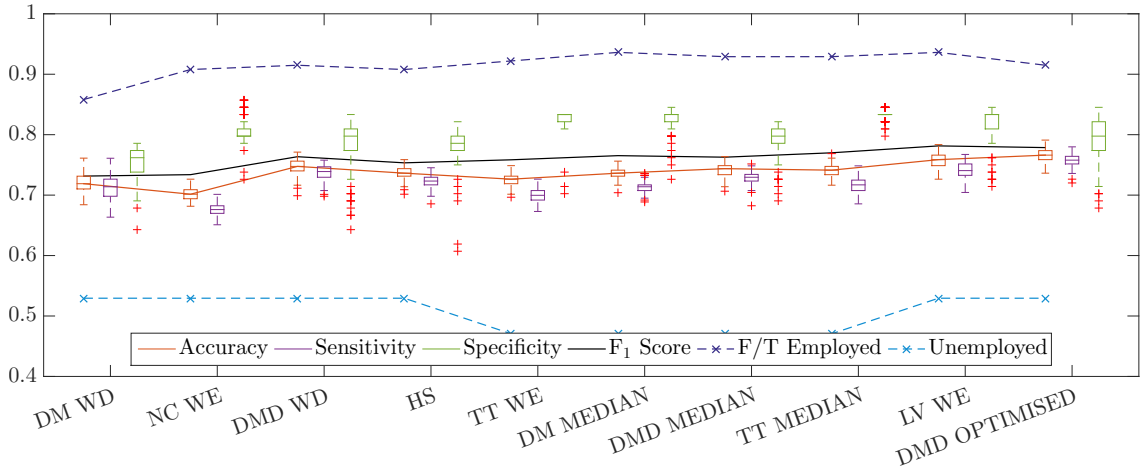


Figure C.2: Depression classification results using naïve Bayes classifier with leave-one-participant-out cross-validation.

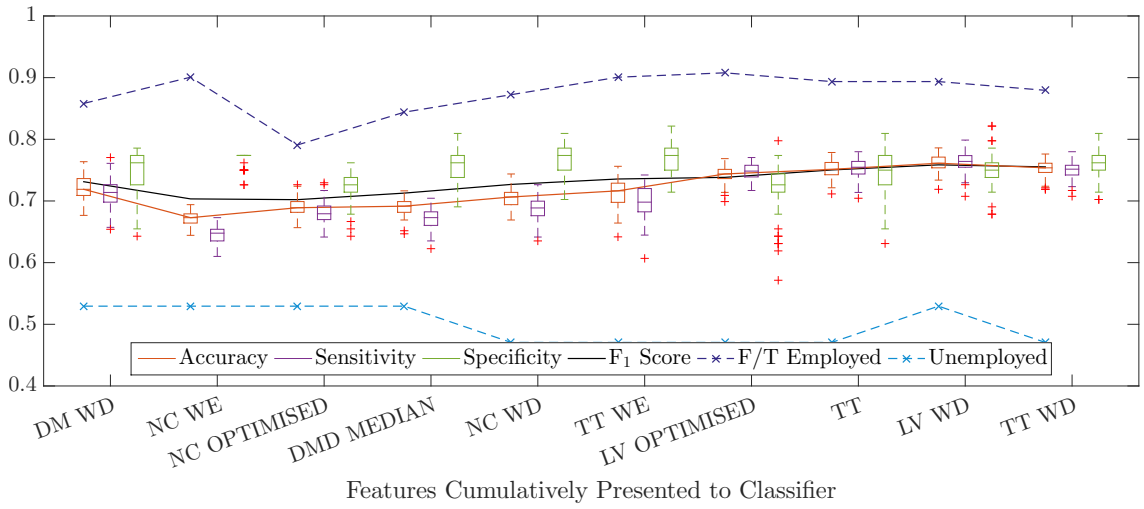


Figure C.3: Depression classification results using quadratic discriminant analysis classifier with leave-one-participant-out cross-validation.

Table C.1: Full depression classification results using logistic regression classifier.†

Features	LOO CV		10-fold CV		5-fold CV		3-fold CV	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
1	0.729 ± 0.015	0.755 ± 0.013	0.707 ± 0.016	0.798 ± 0.021	0.699 ± 0.046	0.798 ± 0.018	0.731 ± 0.022	0.772 ± 0.005
2	0.734 ± 0.022	$\mathbf{0.785} \pm 0.015$	0.754 ± 0.035	0.803 ± 0.025	$\mathbf{0.744} \pm 0.062$	$\mathbf{0.806} \pm 0.029$	0.789 ± 0.025	0.788 ± 0.009
3	$\mathbf{0.769} \pm 0.020$	0.775 ± 0.025	0.752 ± 0.022	$\mathbf{0.812} \pm 0.019$	0.724 ± 0.052	0.797 ± 0.035	0.776 ± 0.020	0.804 ± 0.008
4	0.761 ± 0.015	0.768 ± 0.028	0.746 ± 0.022	0.806 ± 0.027	0.719 ± 0.037	0.773 ± 0.036	0.789 ± 0.022	0.845 ± 0.008
5	0.756 ± 0.022	0.759 ± 0.033	0.744 ± 0.025	0.789 ± 0.026	0.729 ± 0.037	0.793 ± 0.030	0.789 ± 0.022	0.840 ± 0.009
6	0.751 ± 0.026	0.742 ± 0.035	0.751 ± 0.060	0.792 ± 0.121	0.714 ± 0.037	0.766 ± 0.050	$\mathbf{0.801} \pm 0.022$	0.841 ± 0.011
7	0.731 ± 0.030	0.730 ± 0.038	$\mathbf{0.764} \pm 0.027$	0.791 ± 0.022	0.714 ± 0.052	0.750 ± 0.057	0.798 ± 0.022	0.837 ± 0.013
8	0.695 ± 0.040	0.732 ± 0.041	0.754 ± 0.030	0.785 ± 0.022	0.716 ± 0.055	0.752 ± 0.051	0.798 ± 0.022	0.838 ± 0.013
9	0.709 ± 0.032	0.745 ± 0.029	0.740 ± 0.027	0.773 ± 0.033	0.729 ± 0.057	0.782 ± 0.039	0.793 ± 0.021	0.836 ± 0.012
10	0.707 ± 0.027	0.737 ± 0.038	0.749 ± 0.042	0.764 ± 0.067	0.741 ± 0.062	0.785 ± 0.043	0.791 ± 0.025	0.834 ± 0.016
11							0.789 ± 0.022	0.842 ± 0.013
12							$\mathbf{0.801} \pm 0.022$	$\mathbf{0.858} \pm 0.017$

Table C.2: Full depression classification results using naïve Bayes classifier.†

Features	LOO CV		10-fold CV		5-fold CV		3-fold CV	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
1	0.719 ± 0.021	0.743 ± 0.021	0.716 ± 0.027	0.748 ± 0.026	0.739 ± 0.035	0.759 ± 0.019	0.741 ± 0.030	0.774 ± 0.005
2	0.702 ± 0.015	0.775 ± 0.021	0.699 ± 0.021	0.772 ± 0.024	0.776 ± 0.022	0.747 ± 0.024	0.811 ± 0.017	0.771 ± 0.007
3	0.748 ± 0.016	0.769 ± 0.022	0.769 ± 0.025	0.779 ± 0.013	0.764 ± 0.035	0.771 ± 0.020	0.784 ± 0.030	0.761 ± 0.005
4	0.736 ± 0.012	0.752 ± 0.028	0.707 ± 0.030	0.772 ± 0.027	0.766 ± 0.027	0.778 ± 0.025	0.796 ± 0.015	0.779 ± 0.007
5	0.726 ± 0.012	0.770 ± 0.022	0.743 ± 0.027	0.780 ± 0.024	0.744 ± 0.045	0.784 ± 0.023	0.796 ± 0.012	0.776 ± 0.006
6	0.736 ± 0.010	0.774 ± 0.021	0.779 ± 0.017	0.798 ± 0.015	0.769 ± 0.022	0.783 ± 0.024	0.786 ± 0.022	0.802 ± 0.006
7	0.744 ± 0.015	0.779 ± 0.023	0.779 ± 0.019	0.791 ± 0.018	0.786 ± 0.027	0.789 ± 0.018	$\mathbf{0.821} \pm 0.012$	0.808 ± 0.006
8	0.741 ± 0.014	0.790 ± 0.024	0.751 ± 0.022	0.790 ± 0.018	0.784 ± 0.027	0.793 ± 0.033	0.818 ± 0.010	0.814 ± 0.006
9	0.759 ± 0.017	$\mathbf{0.793} \pm 0.018$	0.759 ± 0.017	0.789 ± 0.021	$\mathbf{0.793} \pm 0.027$	$\mathbf{0.802} \pm 0.034$	$\mathbf{0.821} \pm 0.010$	0.817 ± 0.007
10	$\mathbf{0.766} \pm 0.015$	0.791 ± 0.021	$\mathbf{0.786} \pm 0.040$	$\mathbf{0.809} \pm 0.022$	0.781 ± 0.047	0.801 ± 0.030	0.811 ± 0.017	0.823 ± 0.007
11							0.816 ± 0.012	0.823 ± 0.006
12							0.818 ± 0.012	$\mathbf{0.824} \pm 0.007$

Table C.3: Full depression classification results using quadratic discriminant analysis classifier.[†]

Features	LOO CV		10-fold CV		5-fold CV		3-fold CV	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
1	0.719 ± 0.027	0.740 ± 0.018	0.724 ± 0.040	0.731 ± 0.034	0.716 ± 0.035	0.741 ± 0.022	0.741 ± 0.030	0.774 ± 0.005
2	0.673 ± 0.015	0.738 ± 0.029	0.714 ± 0.042	0.716 ± 0.041	0.729 ± 0.037	0.722 ± 0.049	0.756 ± 0.025	0.777 ± 0.015
3	0.689 ± 0.017	0.742 ± 0.031	0.692 ± 0.049	0.681 ± 0.068	0.674 ± 0.050	0.718 ± 0.047	0.736 ± 0.017	0.809 ± 0.011
4	0.692 ± 0.017	0.770 ± 0.037	0.682 ± 0.035	0.705 ± 0.055	0.667 ± 0.037	0.771 ± 0.029	0.766 ± 0.020	0.848 ± 0.016
5	0.707 ± 0.020	0.778 ± 0.036	0.734 ± 0.027	0.771 ± 0.038	0.723 ± 0.065	0.749 ± 0.031	0.784 ± 0.015	0.867 ± 0.011
6	0.716 ± 0.031	0.784 ± 0.030	0.764 ± 0.027	0.776 ± 0.030	0.744 ± 0.040	0.747 ± 0.066	0.796 ± 0.019	0.880 ± 0.014
7	0.744 ± 0.015	0.776 ± 0.029	0.734 ± 0.024	0.772 ± 0.032	0.760 ± 0.047	0.764 ± 0.042	0.838 ± 0.015	0.868 ± 0.013
8	0.751 ± 0.020	0.794 ± 0.035	0.766 ± 0.020	0.785 ± 0.034	0.779 ± 0.043	0.773 ± 0.053	0.843 ± 0.012	0.865 ± 0.015
9	$\mathbf{0.761} \pm 0.017$	$\mathbf{0.798} \pm 0.026$	$\mathbf{0.781} \pm 0.020$	$\mathbf{0.797} \pm 0.029$	$\mathbf{0.793} \pm 0.040$	$\mathbf{0.782} \pm 0.052$	0.851 ± 0.015	0.870 ± 0.016
10	0.754 ± 0.015	0.796 ± 0.020	0.761 ± 0.027	0.783 ± 0.026	0.764 ± 0.027	0.757 ± 0.033	$\mathbf{0.856} \pm 0.015$	0.871 ± 0.015
11							$\mathbf{0.856} \pm 0.017$	$\mathbf{0.881} \pm 0.011$
12							0.846 ± 0.020	0.879 ± 0.015

[†] All results are presented as median $\pm$ iqr.

C.2 Differentiation of mental health conditions

This section gives the full results from classification of study cohorts described in section 6.3.3.

Results are shown from classification performed using logistic regression; naïve Bayes; and the quadratic discriminant analysis models. Figures C.4 to C.6 show the main results (overall accuracy; and sensitivity for each cohort) graphically for the three classification models with leave-one-participant-out cross validation. These results are also detailed fully in tables C.4 to C.6 on page 294. For each left out participant in all models, 20 iterations of group equalisation were run (see section 3.5.1 for details).

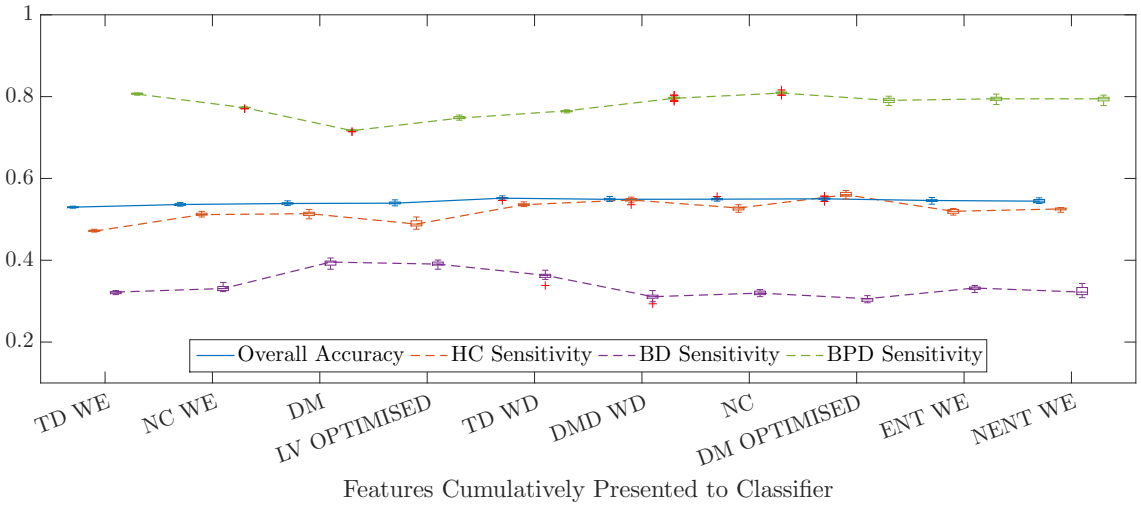


Figure C.4: Cohort classification results using logistic regression classifier with leave-one-participant-out cross-validation.

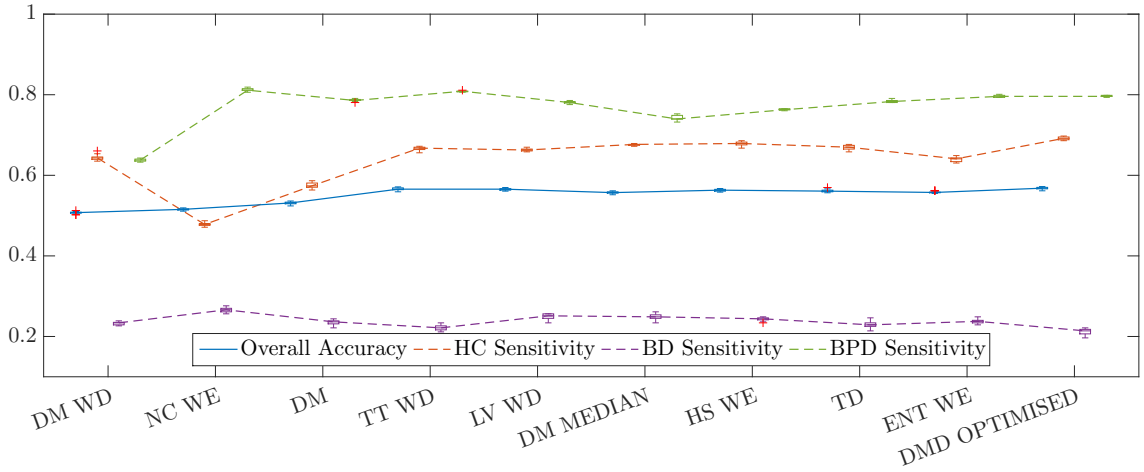


Figure C.5: Cohort classification results using naïve Bayes classifier with leave-one-participant-out cross-validation.

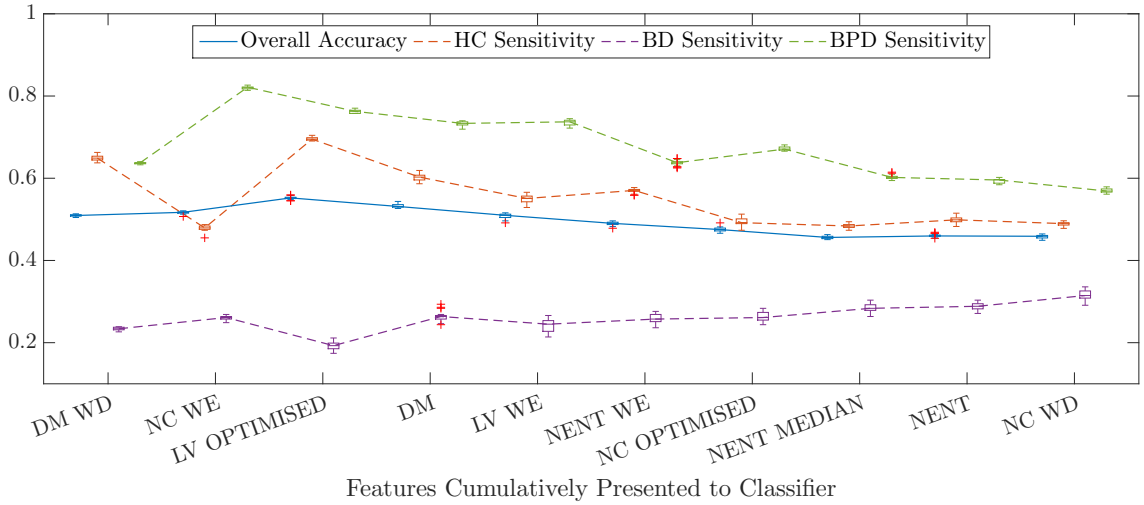


Figure C.6: Cohort classification results using quadratic discriminant analysis classifier with leave-one-participant-out cross-validation.

Table C.4: Full cohort classification results using logistic regression classifier.†

Features	Overall Accuracy	HC Sensitivity	BD Sensitivity	BPD Sensitivity
1	0.530 ± 0.002	0.471 ± 0.002	0.322 ± 0.005	0.806 ± 0.003
2	0.536 ± 0.004	0.511 ± 0.005	0.331 ± 0.010	0.773 ± 0.001
3	0.539 ± 0.004	0.514 ± 0.008	$\mathbf{0.396} \pm 0.010$	0.717 ± 0.001
4	0.539 ± 0.004	0.488 ± 0.013	0.391 ± 0.007	0.747 ± 0.005
5	$\mathbf{0.552} \pm 0.003$	0.536 ± 0.005	0.363 ± 0.007	0.765 ± 0.003
6	0.549 ± 0.004	0.547 ± 0.006	0.311 ± 0.007	0.796 ± 0.003
7	0.549 ± 0.003	0.528 ± 0.007	0.320 ± 0.009	$\mathbf{0.809} \pm 0.003$
8	0.550 ± 0.002	$\mathbf{0.560} \pm 0.009$	0.306 ± 0.006	0.791 ± 0.010
9	0.546 ± 0.004	0.520 ± 0.009	0.332 ± 0.006	0.795 ± 0.008
10	0.544 ± 0.008	0.525 ± 0.005	0.322 ± 0.017	0.795 ± 0.009

Table C.5: Full cohort classification results using naïve Bayes classifier.†

Features	Overall Accuracy	HC Sensitivity	BD Sensitivity	BPD Sensitivity
1	0.507 ± 0.002	0.643 ± 0.006	0.234 ± 0.006	0.638 ± 0.005
2	0.515 ± 0.003	0.478 ± 0.005	$\mathbf{0.266} \pm 0.009$	$\mathbf{0.811} \pm 0.004$
3	0.531 ± 0.004	0.574 ± 0.010	0.236 ± 0.007	0.786 ± 0.003
4	0.566 ± 0.004	0.667 ± 0.006	0.221 ± 0.010	0.809 ± 0.001
5	0.566 ± 0.004	0.663 ± 0.005	0.251 ± 0.011	0.781 ± 0.004
6	0.557 ± 0.003	0.677 ± 0.002	0.249 ± 0.009	0.740 ± 0.009
7	0.563 ± 0.004	0.679 ± 0.007	0.244 ± 0.005	0.763 ± 0.003
8	0.561 ± 0.004	0.670 ± 0.009	0.229 ± 0.009	0.783 ± 0.005
9	0.557 ± 0.002	0.641 ± 0.009	0.238 ± 0.006	0.796 ± 0.005
10	$\mathbf{0.568} \pm 0.004$	$\mathbf{0.692} \pm 0.007$	0.214 ± 0.011	0.796 ± 0.003

Table C.6: Full cohort classification results using quadratic discriminant analysis classifier.†

Features	Overall Accuracy	HC Sensitivity	BD Sensitivity	BPD Sensitivity
1	0.509 ± 0.004	0.649 ± 0.009	0.234 ± 0.006	0.638 ± 0.003
2	0.517 ± 0.003	0.479 ± 0.008	0.261 ± 0.006	$\mathbf{0.821} \pm 0.004$
3	$\mathbf{0.552} \pm 0.003$	$\mathbf{0.695} \pm 0.006$	0.193 ± 0.014	0.763 ± 0.008
4	0.531 ± 0.006	0.603 ± 0.011	0.264 ± 0.009	0.733 ± 0.008
5	0.510 ± 0.008	0.551 ± 0.014	0.245 ± 0.026	0.737 ± 0.011
6	0.490 ± 0.006	0.570 ± 0.005	0.258 ± 0.017	0.638 ± 0.005
7	0.475 ± 0.007	0.492 ± 0.011	0.261 ± 0.019	0.671 ± 0.008
8	0.456 ± 0.006	0.484 ± 0.007	0.284 ± 0.014	0.602 ± 0.005
9	0.460 ± 0.003	0.499 ± 0.009	0.289 ± 0.012	0.596 ± 0.009
10	0.459 ± 0.006	0.490 ± 0.006	$\mathbf{0.315} \pm 0.017$	0.569 ± 0.008

† All results are presented as median $\pm$ iqr.

C.3 Quantification of depression levels

This section gives the full results from quantification of depression levels described in section 6.4.3.

C.3.1 Individual features

The full regression error rates shown in figure 6.14 for regression on individual features calculated on the different data subsets are given in table C.7 across the following two pages. Leave-one-participant-out cross-validation was used to validate the models with all left out weeks used to calculate the regression coefficients, hence there are no error ranges. The regression models used are described in section 6.4.1.1, and the baseline model (described in section 6.4.2) uses the mean QIDS score of the left out weeks as the prediction model.

Table C.7: Full results of quantification of depression levels using individual geolocation features calculated on different data subsets for BD participants.

Feature	Data subset	Linear model		Quadratic Logistic GLM	
		RMSD	MAE	RMSD	MAE
Baseline model		5.531	4.625	5.531	4.625
ENT	Base	5.594	4.672	**5.308*****	**4.372**
	Weekday	5.617	4.671	5.337***	4.392
	Weekend	**5.513**	**4.642**	5.440**	4.562
	Median	5.580	4.661	5.313***	4.381
	Optimised	5.648	4.715	5.631	4.662
NENT	Base	5.700	4.830	5.437**	4.453
	Weekday	5.728	4.831	5.472*	4.494
	Weekend	**5.662**	**4.781**	5.583	4.679
	Median	5.711	4.841	**5.424*****	**4.444**
	Optimised	5.666	4.781	5.426***	4.484
LV	Base	5.383***	4.489	5.482*	4.538
	Weekday	5.406***	4.536	5.603	4.644
	Weekend	**5.267*****	**4.394**	**5.309*****	**4.387**
	Median	5.382***	4.485	5.481*	4.535
	Optimised	5.444***	4.577	5.641	4.658
HS	Base	5.619	4.734	5.430***	4.522
	Weekday	5.603	4.707	5.507	4.561
	Weekend	**5.576**	**4.699**	5.612	4.743
	Median	5.627	4.746	**5.416*****	**4.510**
	Optimised	5.666	4.764	5.717	4.772
TT	Base	5.246***	4.467	5.123***	4.312
	Weekday	5.298***	4.517	5.191***	4.386
	Weekend	5.259***	**4.412**	5.230***	4.418
	Median	**5.238*****	4.458	**5.113*****	**4.300**
	Optimised	5.318***	4.524	5.267***	4.462

(continued on the next page)

Table C.7 (cont.): Full results of quantification of depression levels using individual geolocation features calculated on different data subsets for BD participants.

Feature	Data subset	Linear Model		Quadratic Logistic GLM	
		RMSD	MAE	RMSD	MAE
Baseline model		5.531	4.625	5.531	4.625
TD	Base	5.437***	4.520	5.943	4.725
	Weekday	5.495*	4.576	6.024	4.802
	Weekend	**5.412****	**4.508**	**5.423****	**4.496**
	Median	5.451***	4.530	6.012	4.739
	Optimised	5.675	4.763	5.590	4.699
NC	Base	5.232***	4.383	5.327***	4.445
	Weekday	5.315***	4.445	5.402***	4.517
	Weekend	**5.076****	**4.206**	**5.105****	**4.167**
	Median	5.211***	4.369	5.307***	4.433
	Optimised	5.384***	4.524	5.466**	4.594
DM	Base	5.406***	4.564	5.470*	4.575
	Weekday	5.374***	4.552	5.470*	4.584
	Weekend	**5.325****	**4.456**	**5.318****	**4.376**
	Median	5.388***	4.543	5.454**	4.557
	Optimised	5.480**	4.632	5.540	4.559
DMN	Base	5.718	4.843	5.808	4.873
	Weekday	**5.635**	4.788	5.712	4.847
	Weekend	5.691	**4.692**	**5.343****	**4.464**
	Median	5.712	4.842	5.806	4.871
	Optimised	5.717	4.831	5.737	4.796
DMD	Base	5.573	4.710	5.687	4.700
	Weekday	5.460**	4.625	**5.590**	4.729
	Weekend	**5.445****	**4.534**	5.861	4.727
	Median	5.563	4.703	5.675	**4.690**
	Optimised	5.540	4.661	5.947	4.912

Significance from the baseline model indicated by asterisks: * < 0.05; ** < 0.01; *** < 0.001.

C.3.2 Cumulative features

The full regression error rates for features cumulatively selected and presented to the regression models shown in figure 6.16 are given in table C.8 below. Leave-one-participant-out cross-validation was used to validate the models with all left out weeks used to calculate the regression coefficients. The regression models used are described in section 6.4.1.1, and the baseline model (described in section 6.4.2) uses the mean QIDS score of the left out weeks as the prediction model.

Table C.8: Full results of quantification of depression levels using geolocation features for BD participants cumulatively presented to regression models.

Features	Linear Model			Quadratic Logistic GLM		
Selected	Feature	RMSD	MAE	Feature	RMSD	MAE
	Baseline model	5.531	4.625	Baseline model	5.531	4.625
1	NC WE	5.076	4.206	NC WE	5.105	4.167
2	LV MEDIAN	5.055	4.198	DMD WE	5.003	4.112
3	TD OPTIMISED	5.058	4.186	TD WE	5.050	4.164
4	TT WD	5.015	4.114	NC OPTIMISED	5.115	4.197
5	DMD MEDIAN	**4.903**	3.982	NC MEDIAN	4.976	4.007
6	TD WE	4.910	3.990	TT WD	4.995	4.031
7	NC OPTIMISED	4.923	3.970	HS WE	4.945	3.963
8	NENT WE	4.929	3.945	NC WD	4.912	3.944
9	NC MEDIAN	4.917	**3.929**	NC	4.921	3.890
10	LV	4.931	3.945	NENT OPTIMISED	4.960	3.902
11	ENT WE	4.964	3.968	DMN WD	4.983	3.885
12	TT OPTIMISED	5.008	4.004	DMN	4.778	3.775
13	TT MEDIAN	5.059	4.046	DMN MEDIAN	4.674	3.625
14	DMD WD	5.104	4.089	DMN OPTIMISED	4.677	3.632
15	TT WE	5.154	4.133	NENT WE	4.674	3.646
16	TT	5.205	4.183	LV WE	**4.504**	**3.593**
17	DM OPTIMISED	5.275	4.247	ENT WE	4.525	3.603
18	LV WD	5.338	4.315	TD OPTIMISED	4.618	3.638
19	LV WE	5.375	4.312	TT	4.684	3.696
20	HS	5.331	4.247	TT WE	4.772	3.727
21	DM WE	5.293	4.215	DMD	4.903	3.829
22	DMN	5.328	4.220	DMD MEDIAN	4.982	3.911
23	DMN MEDIAN	5.166	4.096	TT MEDIAN	5.090	3.963
24	ENT WD	5.201	4.099	DM WD	5.130	4.021
25	HS WE	5.155	4.023	DM	4.967	3.854

Appendix D

Feature Values for Selected Individuals

This appendix shows feature values for selected individuals on the raw feature values before filtering and smoothing (figure D.1 on the following page); the filtered feature values before smoothing (figure D.2 on page 301); and the final filtered and smoothed features (figure D.3 on page 302).

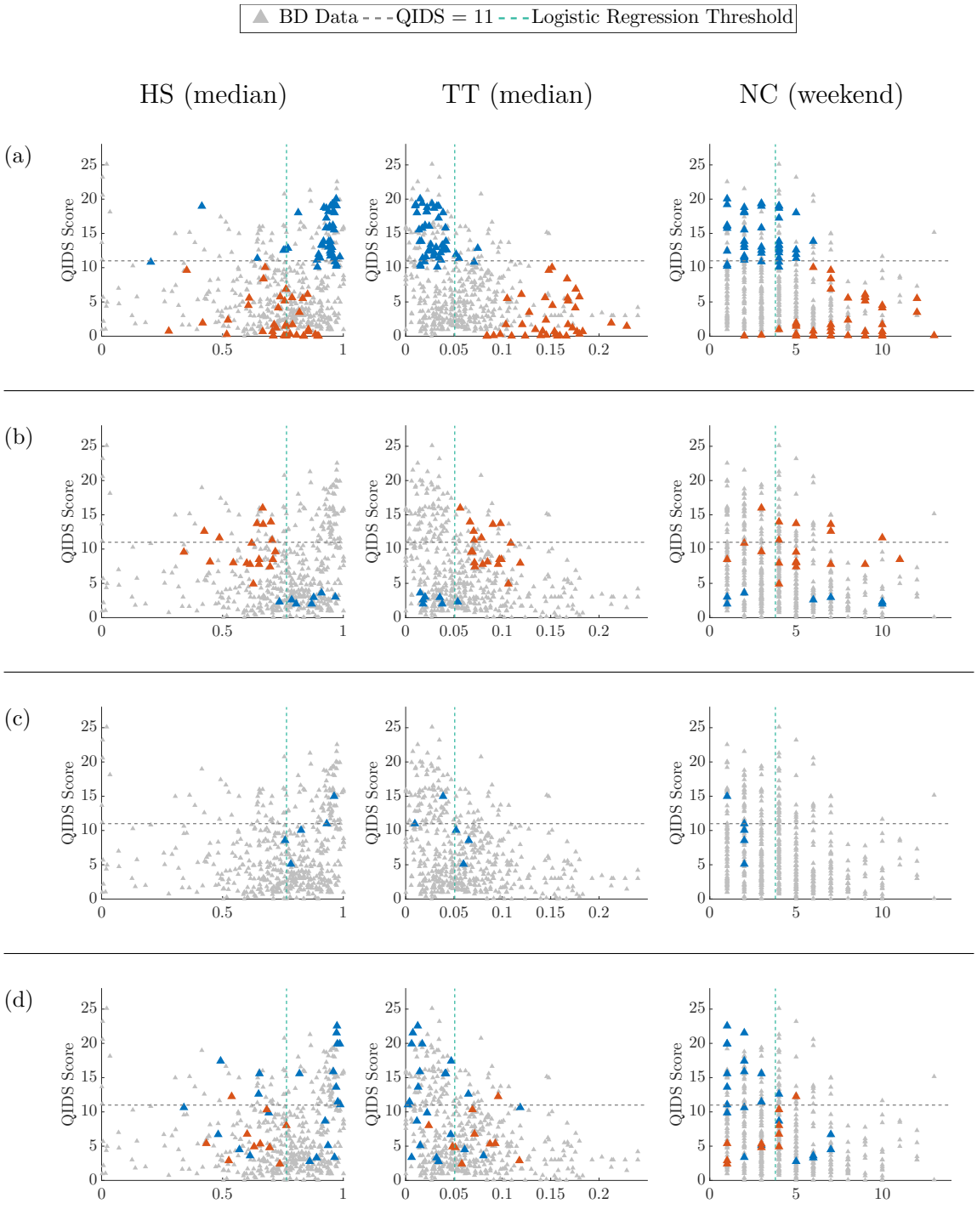


Figure D.1: Raw (unfiltered and unsmoothed) geolocation features for BD participants with individual participants highlighted.

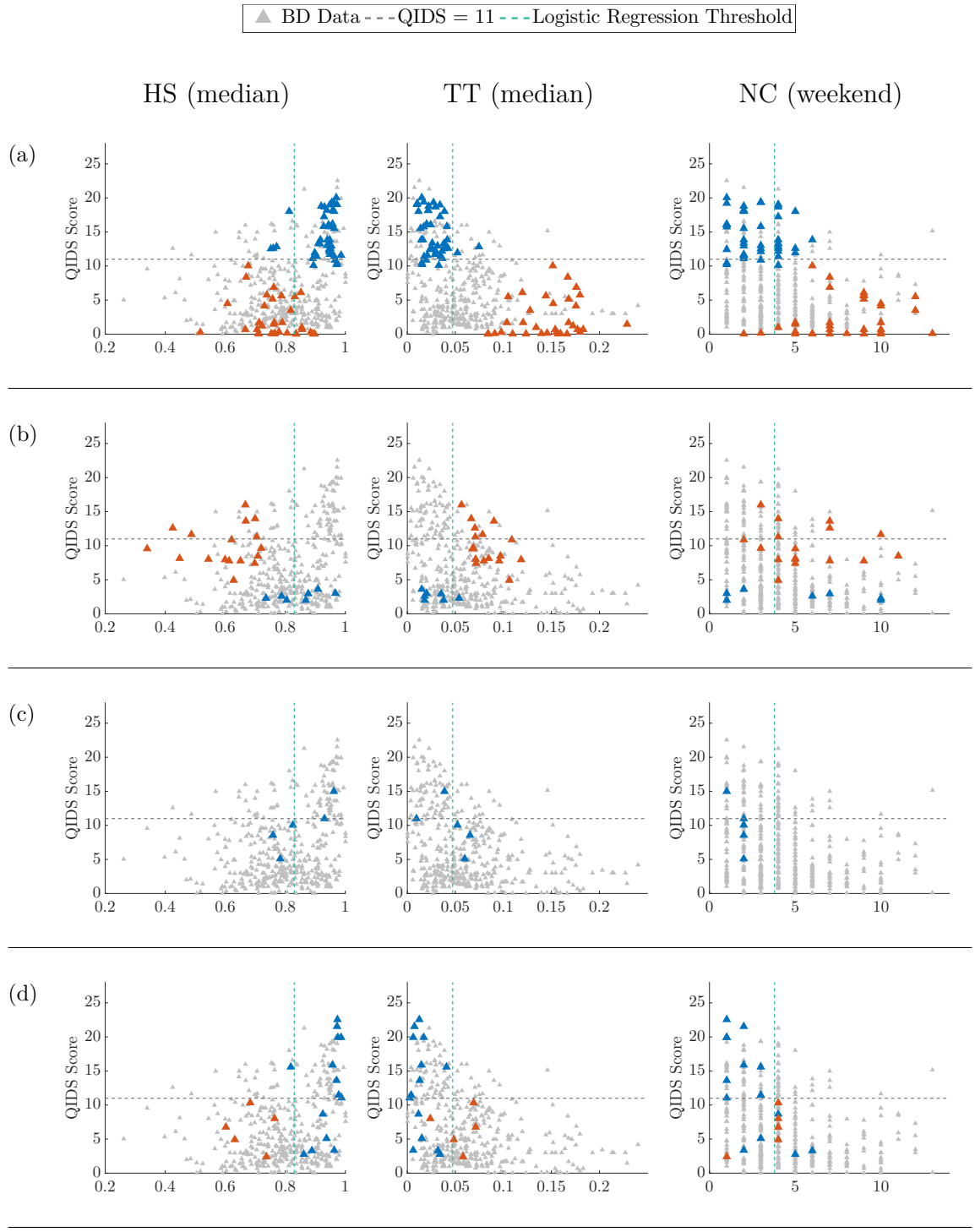


Figure D.2: Filtered (but unsmoothed) geolocation features for BD participants with individual participants highlighted.

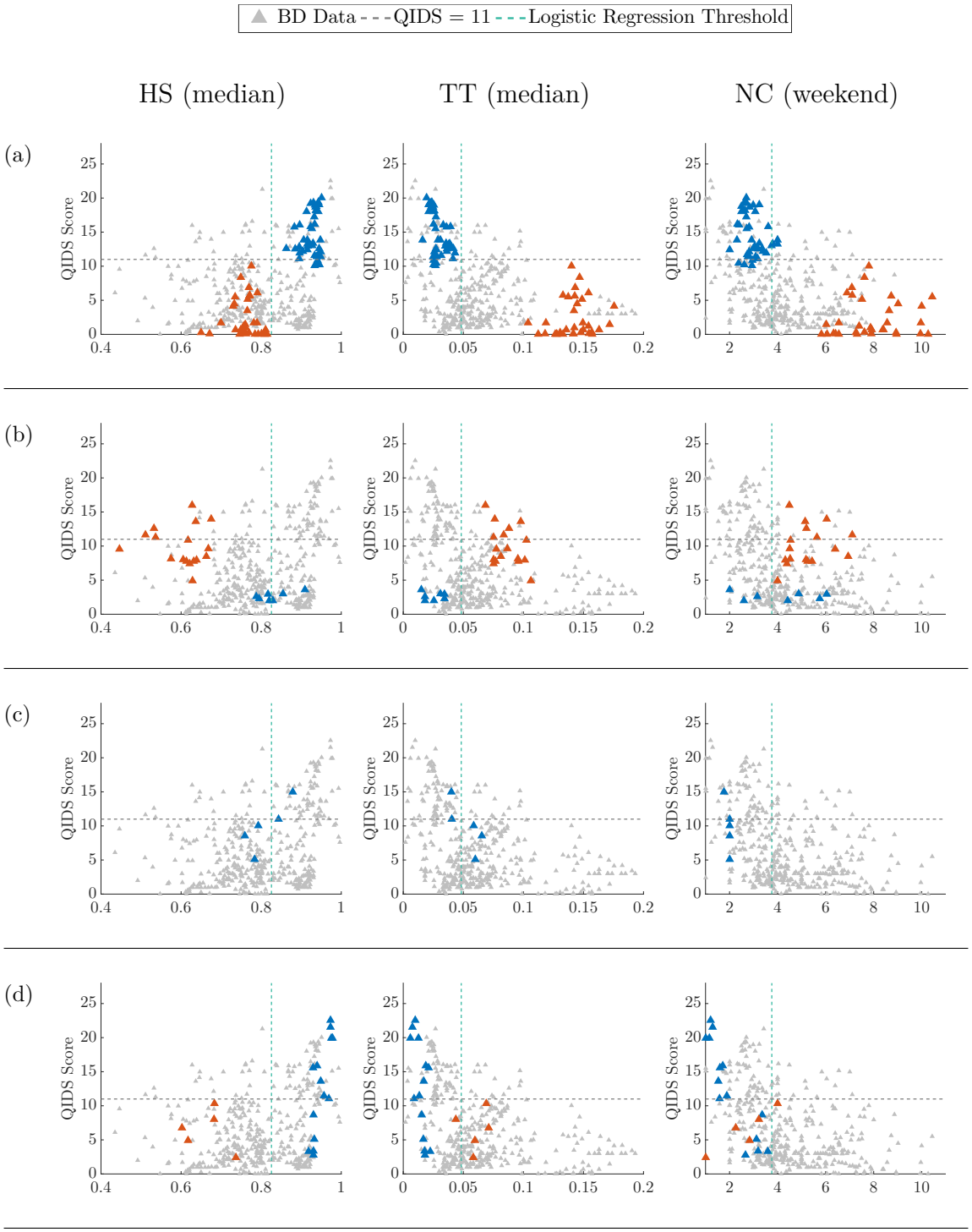


Figure D.3: Filtered and smoothed geolocation features for BD participants with individual participants highlighted.

Appendix E

Derivations

This appendix gives derivations for results given in the main thesis.

E.1 Laplace distribution represented as a scale mixture of Gaussians

This section provides the derivation for the representation of the Laplace distribution as a scale mixture of Gaussians.

The zero-centred Laplace (double exponential) distribution is defined as

$$\pi(\beta) = \frac{\lambda}{2} \exp\{-\lambda|\beta|\} \tag{E.1}$$

using scale parameter λ. This section considers how equation (E.1) can be represented as a scale mixture of Gaussians, defined as the hierarchical model

$$\beta \sim \mathcal{N}(0, \tau^2) \tag{E.2}$$

$$\tau^2 \sim \text{Exp}(\lambda'). \tag{E.3}$$

The full joint distribution for this hierarchical model is defined as

$$p(\beta, \tau^2) = \frac{1}{\sqrt{2\pi\tau^2}} \cdot \exp\left\{-\frac{\beta^2}{2\tau^2}\right\} \cdot \lambda' \cdot \exp\left\{-\lambda'\tau^2\right\}. \tag{E.4}$$

Marginalising the joint over τ^2 gives the density function

$$f_\beta(\beta) = \int_0^\infty \frac{1}{\sqrt{2\pi\tau^2}} \cdot \exp\left\{-\frac{\beta^2}{2\tau^2}\right\} \cdot \lambda' \cdot \exp\left\{-\lambda'\tau^2\right\} d\tau^2 \tag{E.5}$$

$$= \int_0^\infty \frac{\lambda'}{\sqrt{2\pi\tau^2}} \cdot \exp\left\{-\frac{\beta^2}{2\tau^2}\right\} \cdot \exp\left\{-\lambda'\tau^2\right\} d\tau^2. \tag{E.6}$$

The integral in equation (E.6) cannot be directly evaluated. Instead consider the moment generating function for this density, defined as

$$M_{\beta}(t) = \int_{-\infty}^{\infty} \exp\{t\beta\} \cdot f_{\beta}(\beta)\, d\beta \tag{E.7}$$

$$= \int_{-\infty}^{\infty} \exp\{t\beta\} \cdot \overbrace{\int_{0}^{\infty} \frac{\lambda'}{\sqrt{2\pi\tau^2}} \cdot \exp\left\{-\frac{\beta^2}{2\tau^2}\right\} \cdot \exp\left\{-\lambda'\tau^2\right\} d\tau^2}^{f_{\beta}(\beta) \text{ from equation (E.6)}}\, d\beta. \tag{E.8}$$

By swapping the order of integration and rearranging

$$= \int_{0}^{\infty} \exp\left\{-\lambda'\tau^2\right\} \cdot \frac{\lambda'}{\sqrt{2\pi\tau^2}} \cdot \int_{-\infty}^{\infty} \exp\{t\beta\} \cdot \exp\left\{-\frac{\beta^2}{2\tau^2}\right\} d\beta\; d\tau^2 \tag{E.9}$$

the integration over β can be performed using the Gaussian integral

$$= \int_{0}^{\infty} \exp\left\{-\lambda'\tau^2\right\} \cdot \frac{\lambda'}{\sqrt{2\pi\tau^2}} \cdot \sqrt{\frac{\pi}{\frac{1}{2\tau^2}}} \cdot \exp\left\{\frac{\left(-\frac{t}{2}\right)^2}{\frac{1}{2\tau^2}}\right\} d\tau^2 \tag{E.10}$$

$$= \int_{0}^{\infty} \exp\left\{-\lambda'\tau^2\right\} \cdot \frac{\lambda'}{\sqrt{2\pi\tau^2}} \cdot \sqrt{2\pi\tau^2} \cdot \exp\left\{\frac{2t^2\tau^2}{4}\right\} d\tau^2 \tag{E.11}$$

$$= \int_{0}^{\infty} \exp\left\{-\lambda'\tau^2\right\} \cdot \lambda' \cdot \exp\left\{\frac{t^2\tau^2}{2}\right\} d\tau^2 \tag{E.12}$$

$$= \int_{0}^{\infty} \lambda' \cdot \exp\left\{\frac{t^2\tau^2}{2} - \lambda'\tau^2\right\} d\tau^2 \tag{E.13}$$

$$= \lambda' \cdot \int_{0}^{\infty} \exp\left\{\frac{t^2\tau^2 - 2\lambda'\tau^2}{2}\right\} d\tau^2 \tag{E.14}$$

$$= \lambda' \cdot \int_{0}^{\infty} \exp\left\{\frac{t^2 - 2\lambda'}{2}\tau^2\right\} d\tau^2. \tag{E.15}$$

The final integration over τ^2 can also be performed using the Gaussian integral, hence

$$M_{\beta}(t) = \lambda' \cdot \frac{1}{-\frac{t^2-2\lambda'}{2}} = -\lambda' \cdot \frac{2}{t^2 - 2\lambda'} = \frac{-2\lambda'}{t^2 - 2\lambda'} \tag{E.16}$$

$$= \frac{1}{1 - \frac{1}{2\lambda'}t^2}. \tag{E.17}$$

This expression can be compared to the moment generating function for the zero-mean Laplace distribution (parametrised by rate parameter b rather than scale parameter λ used in equation (E.1)), defined as

$$M_X(t) = \frac{e^{t\mu}}{1 - b^2t^2} = \frac{1}{1 - b^2t^2} = \frac{1}{1 - \frac{1}{\lambda^2}t^2}. \tag{E.18}$$

The form of equation (E.17) matches equation (E.18), which means that (due to the uniqueness theorem of moment generating functions) the two distributions are equivalent. Therefore, the parameters can be mapped by observation as

$$b^2 = \frac{1}{2\lambda'} \implies \lambda' = \frac{1}{2b^2}, \tag{E.19}$$

which matches the result by Andrews & Mallows (1974). Parametrising the distribution instead by scale parameter $\lambda = \frac{1}{b}$ as in equation (E.1) gives

$$b^2 = \frac{1}{\lambda^2} = \frac{1}{2\lambda'} \implies \lambda' = \frac{\lambda^2}{2}. \tag{E.20}$$

Hence the following expression holds:

$$\frac{\lambda}{2} \exp\{-\lambda|\beta|\} = \int_0^\infty \frac{1}{\sqrt{2\pi\tau^2}} \cdot \exp\left\{-\frac{\beta^2}{2\tau^2}\right\} \cdot \frac{\lambda^2}{2} \cdot \exp\left\{-\frac{\lambda^2\tau^2}{2}\right\} d\tau^2. \tag{E.21}$$

Figure E.1 shows density functions of two Laplace distributions on the top row and samples from the hierarchical model in equations (E.2) and (E.3) on the bottom row.

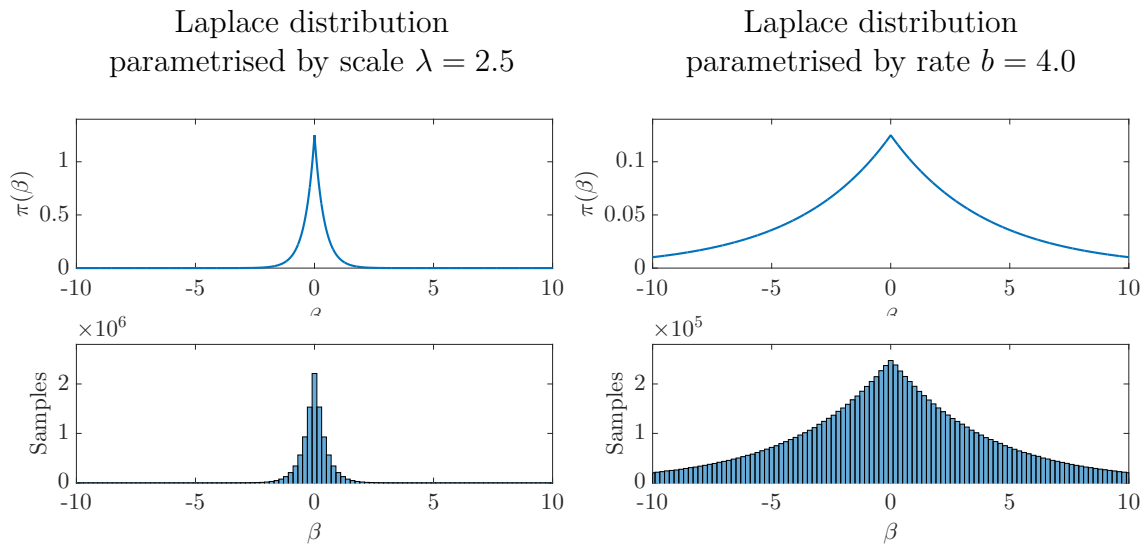


Figure E.1: Samples from the Laplace distribution represented as a scale mixture of Gaussians. The top row shows the density functions of two Laplace distributions paremetrised by the scale parameter λ on the left and the rate parameter b on the right. The bottom row shows histograms of 10 000 000 samples from the hierarchical model in equations (E.2) and (E.3) using parameters from equation (E.20) (left) and equation (E.19) (right).

E.2 Gibbs sampler distributions for Bayesian Lasso over multiple individuals

This section derives the Gibbs sampler distributions given in table 7.2, used in the Bayesian Lasso over multiple individuals and the full grouped Bayesian Lasso. Derivations are based on the full joint density in equation (7.24), defined as follows:

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau^2}, \beta_0, \sigma^2 \,|\, \mathbf{Z}) = &\left[\prod_{k=1}^{M} \frac{1}{(2\pi\sigma^2)^{N_k/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right\}\right] \\ &\times \frac{1}{(2\pi\sigma^2_{\beta_0})^{1/2}} \exp\left\{-\frac{1}{2\sigma^2_{\beta_0}}(\beta_0 - \mu_{\beta_0})^2\right\} \cdot \frac{\gamma^\alpha}{\Gamma(\alpha)} \left(\sigma^2\right)^{-\alpha-1} \exp\left\{-\frac{\gamma}{\sigma^2}\right\} \\ &\times \left[\prod_{j=1}^{p} \frac{1}{(2\pi\sigma^2\tau_j^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\} \cdot \frac{\lambda^2}{2} \exp\left\{-\frac{\lambda^2\tau_j^2}{2}\right\}\right] \end{aligned} \tag{E.22}$$

In the joint above and the derivations that follow, it is assumed that there are M individuals over which the probability is calculated.

E.2.1 τ_j^2 variables

The elements of the joint involving the τ_j^2 variable for a single component j are

$$p(\tau_j^2 \,|\, \cdot) \propto \frac{1}{(2\pi\sigma^2\tau_j^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\} \exp\left\{-\frac{\lambda^2\tau_j^2}{2}\right\} \tag{E.23}$$

$$\propto (\tau_j^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2 - \frac{\lambda^2\tau_j^2}{2}\right\} \tag{E.24}$$

$$\propto (\tau_j^2)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{\beta_j^2}{\sigma^2\tau_j^2} + \lambda^2\tau_j^2\right)\right\} \tag{E.25}$$

setting $\eta_j^2 = \frac{1}{\tau_j^2}$

$$p(\eta_j^2 \,|\, \cdot) \propto (\eta_j^2)^{-3/2} \exp\left\{-\frac{1}{2}\left(\frac{\beta_j^2\eta_j^2}{\sigma^2} + \frac{\lambda^2}{\eta_j^2}\right)\right\} \tag{E.26}$$

$$\propto (\eta_j^2)^{-3/2} \exp\left\{-\frac{1}{2}\left(\frac{\beta_j^2\eta_j^4 + \sigma^2\lambda^2}{\sigma^2\eta_j^2}\right)\right\} \tag{E.27}$$

$$\propto (\eta_j^2)^{-3/2} \exp\left\{-\frac{1}{2}\left(\frac{\eta_j^4 + \sigma^2\lambda^2/\beta_j^2}{\sigma^2\eta_j^2/\beta_j^2}\right)\right\} \tag{E.28}$$

completing the square on top

$$p(\eta_j^2 \,|\, \cdot\,) \propto (\eta_j^2)^{-3/2} \exp\left\{-\frac{1}{2}\left(\frac{\left(\eta_j^2 - \sqrt{\sigma^2\lambda^2/\beta_j^2}\right)^2 + 2\eta_j^2\sqrt{\sigma^2\lambda^2/\beta_j^2}}{\sigma^2\eta_j^2/\beta_j^2}\right)\right\} \tag{E.29}$$

$$\propto (\eta_j^2)^{-3/2} \exp\left\{-\frac{1}{2}\left(\frac{\left(\eta_j^2 - \sqrt{\sigma^2\lambda^2/\beta_j^2}\right)^2}{\sigma^2\eta_j^2/\beta_j^2} + \frac{2\eta_j^2\sqrt{\sigma^2\lambda^2/\beta_j^2}}{\sigma^2\eta_j^2/\beta_j^2}\right)\right\} \tag{E.30}$$

$$\propto (\eta_j^2)^{-3/2} \exp\left\{-\frac{1}{2}\left(\frac{\left(\eta_j^2 - \sqrt{\sigma^2\lambda^2/\beta_j^2}\right)^2}{\sigma^2\eta_j^2/\beta_j^2} + \frac{2\sqrt{\sigma^2\lambda^2/\beta_j^2}}{\sigma^2/\beta_j^2}\right)\right\} \tag{E.31}$$

removing constant terms not involving η_j^2

$$\propto (\eta_j^2)^{-3/2} \exp\left\{-\frac{1}{2}\left(\frac{\left(\eta_j^2 - \sqrt{\sigma^2\lambda^2/\beta_j^2}\right)^2}{\sigma^2\eta_j^2/\beta_j^2}\right)\right\} \tag{E.32}$$

$$\propto (\eta_j^2)^{-3/2} \exp\left\{-\frac{1}{2}\left(\frac{\lambda^2\left(\eta_j^2 - \sqrt{\sigma^2\lambda^2/\beta_j^2}\right)^2}{\left(\sigma^2\lambda^2/\beta_j^2\right)\eta_j^2}\right)\right\} \tag{E.33}$$

$$\propto (\eta_j^2)^{-3/2} \exp\left\{-\frac{1}{2}\left(\frac{\lambda^2\left(\eta_j^2 - \sqrt{\sigma^2\lambda^2/\beta_j^2}\right)^2}{\left(\sqrt{\sigma^2\lambda^2/\beta_j^2}\right)^2\eta_j^2}\right)\right\} \tag{E.34}$$

hence $\eta_j^2 = \frac{1}{\tau_j^2}$ is sampled from an inverse Gaussian distribution with

$$\text{mean parameter } \mu'_{\eta_j^2} = \sqrt{\frac{\sigma^2\lambda^2}{\beta_j^2}} \text{ and} \tag{E.35}$$

$$\text{scale parameter } \lambda'_{\eta_j^2} = \lambda^2. \tag{E.36}$$

E.2.2 $\boldsymbol{\beta}$ variable

The elements of the joint involving the $\boldsymbol{\beta}$ variable are

$$p(\boldsymbol{\beta} \mid \cdot) \propto \left[\prod_{k=1}^{M} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right\}\right] \times \prod_{j=1}^{p} \exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\} \tag{E.37}$$

setting $\tilde{\mathbf{y}}_k = \mathbf{y}_k - \beta_0 \mathbf{1}_{N_k}$

$$\propto \left[\prod_{k=1}^{M} \exp\left\{-\frac{1}{2\sigma^2}(\tilde{\mathbf{y}}_k - \mathbf{Z}_k\boldsymbol{\beta})^\top(\tilde{\mathbf{y}}_k - \mathbf{Z}_k\boldsymbol{\beta})\right\}\right] \prod_{j=1}^{p} \exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\} \tag{E.38}$$

setting $\mathbf{D} = \text{diag}(\tau_1^2, \ldots, \tau_p^2)$

$$\propto \left[\prod_{k=1}^{M} \exp\left\{-\frac{1}{2\sigma^2}(\tilde{\mathbf{y}}_k - \mathbf{Z}_k\boldsymbol{\beta})^\top(\tilde{\mathbf{y}}_k - \mathbf{Z}_k\boldsymbol{\beta})\right\}\right] \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\beta}^\top\mathbf{D}^{-1}\boldsymbol{\beta}\right\} \tag{E.39}$$

$$\propto \exp\left\{\sum_{k=1}^{M} -\frac{1}{2\sigma^2}(\tilde{\mathbf{y}}_k - \mathbf{Z}_k\boldsymbol{\beta})^\top(\tilde{\mathbf{y}}_k - \mathbf{Z}_k\boldsymbol{\beta})\right\} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\beta}^\top\mathbf{D}^{-1}\boldsymbol{\beta}\right\} \tag{E.40}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{M}(\tilde{\mathbf{y}}_k - \mathbf{Z}_k\boldsymbol{\beta})^\top(\tilde{\mathbf{y}}_k - \mathbf{Z}_k\boldsymbol{\beta})\right\} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\beta}^\top\mathbf{D}^{-1}\boldsymbol{\beta}\right\} \tag{E.41}$$

combining the exponentials

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top\mathbf{D}^{-1}\boldsymbol{\beta} + \sum_{k=1}^{M}(\tilde{\mathbf{y}}_k - \mathbf{Z}_k\boldsymbol{\beta})^\top(\tilde{\mathbf{y}}_k - \mathbf{Z}_k\boldsymbol{\beta})\right)\right\} \tag{E.42}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top\mathbf{D}^{-1}\boldsymbol{\beta} + \sum_{k=1}^{M}\left((\tilde{\mathbf{y}}_k^\top - \boldsymbol{\beta}^\top\mathbf{Z}_k^\top)(\tilde{\mathbf{y}}_k - \mathbf{Z}_k\boldsymbol{\beta})\right)\right)\right\} \tag{E.43}$$

multiplying out

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top\mathbf{D}^{-1}\boldsymbol{\beta} + \sum_{k=1}^{M}\left(\tilde{\mathbf{y}}_k^\top\tilde{\mathbf{y}}_k - \boldsymbol{\beta}^\top\mathbf{Z}_k^\top\tilde{\mathbf{y}}_k - \tilde{\mathbf{y}}_k^\top\mathbf{Z}_k\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{Z}_k^\top\mathbf{Z}_k\boldsymbol{\beta}\right)\right)\right\} \tag{E.44}$$

noting that $\tilde{\mathbf{y}}_k^\top\mathbf{Z}_k\boldsymbol{\beta} = \boldsymbol{\beta}^\top\mathbf{Z}_k^\top\tilde{\mathbf{y}}_k$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top\mathbf{D}^{-1}\boldsymbol{\beta} + \sum_{k=1}^{M}\left(\tilde{\mathbf{y}}_k^\top\tilde{\mathbf{y}}_k - 2\tilde{\mathbf{y}}_k^\top\mathbf{Z}_k\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{Z}_k^\top\mathbf{Z}_k\boldsymbol{\beta}\right)\right)\right\} \tag{E.45}$$

omitting the limits of the sum for brevity this can be re-written as

$$p(\boldsymbol{\beta} \mid \cdot) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top \mathbf{D}^{-1}\boldsymbol{\beta} + \sum \tilde{\mathbf{y}}_k^\top \tilde{\mathbf{y}}_k - \sum 2\tilde{\mathbf{y}}_k^\top \mathbf{Z}_k\boldsymbol{\beta} + \sum \boldsymbol{\beta}^\top \mathbf{Z}_k^\top \mathbf{Z}_k\boldsymbol{\beta}\right)\right\} \tag{E.46}$$

moving constants outside the sums and removing $\sum \tilde{\mathbf{y}}_k^\top \tilde{\mathbf{y}}_k$ term not involving $\boldsymbol{\beta}$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top \mathbf{D}^{-1}\boldsymbol{\beta} - 2\left(\sum \tilde{\mathbf{y}}_k^\top \mathbf{Z}_k\right)\boldsymbol{\beta} + \boldsymbol{\beta}^\top \left(\sum \mathbf{Z}_k^\top \mathbf{Z}_k\right)\boldsymbol{\beta}\right)\right\} \tag{E.47}$$

collecting terms and rearranging

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top \left(\mathbf{D}^{-1} + \sum \mathbf{Z}_k^\top \mathbf{Z}_k\right)\boldsymbol{\beta} - 2\left(\sum \tilde{\mathbf{y}}_k^\top \mathbf{Z}_k\right)\boldsymbol{\beta}\right)\right\} \tag{E.48}$$

setting $\mathbf{A} = \mathbf{D}^{-1} + \sum_{k=1}^{M} \mathbf{Z}_k^\top \mathbf{Z}_k$ and $\mathbf{B} = \sum_{k=1}^{M} \tilde{\mathbf{y}}_k^\top \mathbf{Z}_k$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top \mathbf{A}\boldsymbol{\beta} - 2\mathbf{B}\boldsymbol{\beta}\right)\right\} \tag{E.49}$$

completing the square (because $\mathbf{A}$ is symmetric)

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\left(\boldsymbol{\beta} - \mathbf{A}^{-1}\mathbf{B}^\top\right)^\top \mathbf{A}\left(\boldsymbol{\beta} - \mathbf{A}^{-1}\mathbf{B}^\top\right) - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top\right)\right\} \tag{E.50}$$

removing constant terms not involving $\boldsymbol{\beta}$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\left(\boldsymbol{\beta} - \mathbf{A}^{-1}\mathbf{B}^\top\right)^\top \mathbf{A}\left(\boldsymbol{\beta} - \mathbf{A}^{-1}\mathbf{B}^\top\right)\right)\right\} \tag{E.51}$$

hence $\boldsymbol{\beta}$ is sampled from a multivariate normal distribution with

$$\text{mean parameter } \boldsymbol{\mu}'_\beta = \mathbf{A}^{-1}\mathbf{B}^\top \text{ and} \tag{E.52}$$

$$\text{covariance parameter } \boldsymbol{\Sigma}'_\beta = \sigma^2\mathbf{A}^{-1} \tag{E.53}$$

where

$$\mathbf{A} = \mathbf{D}^{-1} + \sum_{k=1}^{M} \mathbf{Z}_k^\top \mathbf{Z}_k, \quad \mathbf{D} = \text{diag}(\tau_1^2, \ldots, \tau_p^2) \tag{E.54}$$

$$\mathbf{B} = \sum_{k=1}^{M} \tilde{\mathbf{y}}_k^\top \mathbf{Z}_k = \sum_{k=1}^{M} \left(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k}\right)^\top \mathbf{Z}_k. \tag{E.55}$$

E.2.3 β_0 variable

The elements of the joint involving the β_0 variable are

$$p(\beta_0 \,|\, \cdot) \propto \left[\prod_{k=1}^{M} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right\}\right]$$
$$\times \exp\left\{-\frac{1}{2\sigma^2_{\beta_0}}(\beta_0 - \mu_{\beta_0})^2\right\} \tag{E.56}$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{k=1}^{M}(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right]\right\}$$
$$\times \exp\left\{-\frac{1}{2\sigma^2_{\beta_0}}(\beta_0 - \mu_{\beta_0})^2\right\} \tag{E.57}$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{k=1}^{M}(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right]\right.$$
$$\left. -\frac{1}{2\sigma^2_{\beta_0}}(\beta_0 - \mu_{\beta_0})^2\right\} \tag{E.58}$$

setting $\boldsymbol{\Delta}_k = \mathbf{y}_k - \mathbf{Z}_k\boldsymbol{\beta}$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{k=1}^{M}(\boldsymbol{\Delta}_k - \beta_0 \mathbf{1}_{N_k})^\top(\boldsymbol{\Delta}_k - \beta_0 \mathbf{1}_{N_k})\right] - \frac{1}{2\sigma^2_{\beta_0}}(\beta_0 - \mu_{\beta_0})^2\right\} \tag{E.59}$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{k=1}^{M}(\beta_0 \mathbf{1}_{N_k}^\top \beta_0 \mathbf{1}_{N_k} - 2\beta_0 \mathbf{1}_{N_k}^\top \boldsymbol{\Delta}_k)\right] - \frac{1}{2\sigma^2_{\beta_0}}(\beta_0^2 - 2\beta_0\mu_{\beta_0})\right\} \tag{E.60}$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{k=1}^{M} N_k\beta_0^2 - \sum_{k=1}^{M} 2\beta_0 \mathbf{1}_{N_k}^\top \boldsymbol{\Delta}_k\right] - \frac{1}{2\sigma^2_{\beta_0}}(\beta_0^2 - 2\beta_0\mu_{\beta_0})\right\} \tag{E.61}$$
$$\propto \exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\sigma^2}\sum_{k=1}^{M} N_k\right)\beta_0^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{M}\mathbf{1}_{N_k}^\top\boldsymbol{\Delta}_k\right)\beta_0\right.\right.$$
$$\left.\left. +\frac{1}{\sigma^2_{\beta_0}}\beta_0^2 - 2\frac{\mu_{\beta_0}}{\sigma^2_{\beta_0}}\beta_0\right)\right\} \tag{E.62}$$

setting $A = \sum_{k=1}^{M} N_k$ and $B = \sum_{k=1}^{M}\mathbf{1}_{N_k}^\top\boldsymbol{\Delta}_k = \sum_{k=1}^{M}\boldsymbol{\Delta}_k^\top\mathbf{1}_{N_k} = \sum_{k=1}^{M}(\mathbf{y}_k - \mathbf{Z}_k\boldsymbol{\beta})^\top\mathbf{1}_{N_k}$

$$\propto \exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\sigma^2}A\right)\beta_0^2 - 2\left(\frac{1}{\sigma^2}B\right)\beta_0 + \left(\frac{1}{\sigma^2_{\beta_0}}\right)\beta_0^2 - 2\left(\frac{\mu_{\beta_0}}{\sigma^2_{\beta_0}}\right)\beta_0\right)\right\} \tag{E.63}$$

collecting terms and rearranging

$$p(\beta_0 \,|\, \cdot\,) \propto \exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\sigma^2_{\beta_0}}+\frac{1}{\sigma^2}A\right)\beta_0^2 - 2\left(\frac{\mu_{\beta_0}}{\sigma^2_{\beta_0}}+\frac{1}{\sigma^2}B\right)\beta_0\right)\right\} \tag{E.64}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\left(\frac{\sigma^2+\sigma^2_{\beta_0}A}{\sigma^2\sigma^2_{\beta_0}}\right)\beta_0^2 - 2\left(\frac{\sigma^2\mu_{\beta_0}+\sigma^2_{\beta_0}B}{\sigma^2\sigma^2_{\beta_0}}\right)\beta_0\right)\right\} \tag{E.65}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2\sigma^2_{\beta_0}}\left(\left(\sigma^2+\sigma^2_{\beta_0}A\right)\beta_0^2 - 2\left(\sigma^2\mu_{\beta_0}+\sigma^2_{\beta_0}B\right)\beta_0\right)\right\} \tag{E.66}$$

$$\propto \exp\left\{-\frac{\sigma^2+\sigma^2_{\beta_0}A}{2\sigma^2\sigma^2_{\beta_0}}\left(\beta_0^2 - 2\left(\frac{\sigma^2\mu_{\beta_0}+\sigma^2_{\beta_0}B}{\sigma^2+\sigma^2_{\beta_0}A}\right)\beta_0\right)\right\} \tag{E.67}$$

completing the square and omitting residual term not including β_0

$$\propto \exp\left\{-\frac{\sigma^2+\sigma^2_{\beta_0}A}{2\sigma^2\sigma^2_{\beta_0}}\left(\beta_0 - \left(\frac{\sigma^2\mu_{\beta_0}+\sigma^2_{\beta_0}B}{\sigma^2+\sigma^2_{\beta_0}A}\right)\right)^2\right\} \tag{E.68}$$

hence the β_0 parameter can be sampled from a normal distribution with

$$\begin{aligned}\text{mean parameter } \mu'_{\beta_0} &= \frac{\sigma^2\mu_{\beta_0}+\sigma^2_{\beta_0}B}{\sigma^2+\sigma^2_{\beta_0}A} \\ &= \frac{\sigma^2\mu_{\beta_0}+\sigma^2_{\beta_0}\sum_{k=1}^{M}\left(\mathbf{y}_k-\mathbf{Z}_k\boldsymbol{\beta}\right)^\top \mathbf{1}_{N_k}}{\sigma^2+\sigma^2_{\beta_0}\sum_{k=1}^{M}N_k} \quad \text{and}\end{aligned} \tag{E.69}$$

$$\begin{aligned}\text{standard deviation } \sigma'_{\beta_0} &= \sqrt{\frac{\sigma^2\sigma^2_{\beta_0}}{\sigma^2+\sigma^2_{\beta_0}A}} \\ &= \sqrt{\frac{\sigma^2\sigma^2_{\beta_0}}{\sigma^2+\sigma^2_{\beta_0}\sum_{k=1}^{M}N_k}}.\end{aligned} \tag{E.70}$$

E.2.4 σ^2 variable

The elements of the joint involving the σ^2 variable are

$$
\begin{aligned}
p(\sigma^2 \mid \cdot) \;\propto\; & \left[\prod_{k=1}^{M} \frac{1}{(2\pi\sigma^2)^{N_k/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right\}\right] \\
& \times \left[\prod_{j=1}^{p} \frac{1}{(2\pi\sigma^2\tau_j^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\}\right] \\
& \times \frac{\gamma^\alpha}{\Gamma(\alpha)} \left(\sigma^2\right)^{-\alpha-1} \exp\left\{-\gamma/\sigma_k^2\right\} \qquad \text{(E.71)} \\
\propto\; & \left[\prod_{k=1}^{M} (\sigma^2)^{-N_k/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right\}\right] \\
& \times \left[\prod_{j=1}^{p} (\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\}\right] \left(\sigma^2\right)^{-\alpha-1} \exp\left\{-\gamma/\sigma_k^2\right\} \qquad \text{(E.72)}
\end{aligned}
$$

replacing products with sums of exponents and setting $\mathbf{D} = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$

$$
\begin{aligned}
& \propto (\sigma^2)^{-\sum_{k=1}^{M} N_k/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{M}(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right\} \\
& \quad \times (\sigma^2)^{-p/2} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\beta}^\top\mathbf{D}\boldsymbol{\beta}\right\} \left(\sigma^2\right)^{-\alpha-1} \exp\left\{-\gamma/\sigma_k^2\right\} \qquad \text{(E.73)} \\
& \propto \left(\sigma^2\right)^{-\left(\sum_{k=1}^{M} N_k/2\right) - p/2 - \alpha - 1} \\
& \quad \times \exp\left\{-\frac{1}{\sigma^2} \cdot \frac{1}{2}\left(\left(\sum_{k=1}^{M}(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right)\right. \right. \\
& \qquad \left.\left. + \boldsymbol{\beta}^\top\mathbf{D}\boldsymbol{\beta} + 2\gamma\right)\right\} \qquad \text{(E.74)}
\end{aligned}
$$

hence σ^2 is sampled from an inverse gamma distribution with

$$
\begin{aligned}
\text{shape parameter } \alpha'_{\sigma^2} &= \left(\sum_{k=1}^{M} N_k/2\right) + p/2 + \alpha \quad \text{and} \qquad \text{(E.75)} \\
\text{scale parameter } \beta'_{\sigma^2} &= \frac{1}{2}\sum_{k=1}^{M}(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta}) \\
&\quad + \boldsymbol{\beta}^\top\mathbf{D}\boldsymbol{\beta}/2 + \gamma, \quad \mathbf{D} = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2). \qquad \text{(E.76)}
\end{aligned}
$$

E.3 Integration of joint distribution by $\boldsymbol{\tau^2}$

This section demonstrates the integration of the full joint distribution $p(\mathbf{y}_k, \boldsymbol{\beta}, \boldsymbol{\tau^2}, \beta_0, \sigma^2 \,|\, \mathbf{Z}_k)$ in equation (7.24) by $\boldsymbol{\tau^2}$ to derive the marginal joint $p(\mathbf{y}_k, \phi_c \,|\, \mathbf{Z}_k)$ in equation (7.29) used to allocate an individual k to group c in the DP Gibbs sampler.

The joint density in equation (7.24) expressed for a single individual k, marginalised over $\boldsymbol{\tau^2}$ is expressed as

$$p(\mathbf{y}_k, \phi_c \,|\, \mathbf{Z}_k) = \int_{-\infty}^{\infty} p(\mathbf{y}_k, \boldsymbol{\beta}, \boldsymbol{\tau^2}, \beta_0, \sigma^2 \,|\, \mathbf{Z}_k)\; d\boldsymbol{\tau^2} = \int_{0}^{\infty} p(\mathbf{y}_k, \boldsymbol{\beta}, \boldsymbol{\tau^2}, \beta_0, \sigma^2 \,|\, \mathbf{Z}_k)\; d\boldsymbol{\tau^2} \tag{E.77}$$

$$= \int_0^\infty \cdots \int_0^\infty p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau^2}, \beta_0, \sigma^2 \,|\, \mathbf{Z}_k)\, d\tau_1^2 \ldots d\tau_p^2 \tag{E.78}$$

$$= \int_0^\infty \cdots \int_0^\infty f(\mathbf{y}_k \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2, \mathbf{Z}_k) \cdot f(\boldsymbol{\beta} \,|\, \boldsymbol{\tau^2}, \sigma^2) \cdot \pi(\boldsymbol{\tau^2}) \cdot \pi(\beta_0) \cdot \pi(\sigma^2)\; d\tau_1^2 \ldots d\tau_p^2 \tag{E.79}$$

$$= f(\mathbf{y}_k \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2, \mathbf{Z}_k) \cdot \pi(\beta_0) \cdot \pi(\sigma^2) \cdot \int_0^\infty \cdots \int_0^\infty f(\boldsymbol{\beta} \,|\, \boldsymbol{\tau^2}, \sigma^2) \cdot \pi(\boldsymbol{\tau^2})\; d\tau_1^2 \ldots d\tau_p^2, \tag{E.80}$$

where the integrations by $d\tau_1^2, \ldots, d\tau_p^2$ are independent, hence

$$= f(\mathbf{y}_k \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2, \mathbf{Z}_k) \cdot \pi(\beta_0) \cdot \pi(\sigma^2) \cdot \prod_{j=1}^{p} \int_0^\infty f(\beta_j \,|\, \tau_j^2, \sigma^2) \cdot \pi(\tau_j^2)\; d\tau_j^2. \tag{E.81}$$

The integral over τ_j^2 has been shown to be equivalent to the Laplace distribution in section E.1, hence

$$= f(\mathbf{y}_k \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2, \mathbf{Z}_k) \cdot \pi(\beta_0) \cdot \pi(\sigma^2) \cdot \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left\{-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right\} \tag{E.82}$$

$$= \frac{1}{(2\pi\sigma^2)^{N_k/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})^\top (\mathbf{y}_k - \beta_0 \mathbf{1}_{N_k} - \mathbf{Z}_k\boldsymbol{\beta})\right\} \cdot \frac{1}{(2\pi\sigma_{\beta_0}^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma_{\beta_0}^2}(\beta_0 - \mu_{\beta_0})^2\right\}$$
$$\times \frac{\gamma^\alpha}{\Gamma(\alpha)} \left(\sigma^2\right)^{-\alpha-1} \exp\left\{-\frac{\gamma}{\sigma^2}\right\} \cdot \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left\{-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right\}. \tag{E.83}$$

E.4 Methods of performing principal component analysis

This section demonstrates two of the main methods for solving the PCA equation

$$\mathbf{ZW} = \mathbf{Y} \tag{E.84}$$

constrained such that

$$\operatorname{cov}(\mathbf{ZW}) = \operatorname{diag}(\mathbf{w}) \tag{E.85}$$

where $\mathbf{Z}$ is the $N \times p$ feature matrix to de-correlate; $\mathbf{W}$ is a $p \times p$ de-correlation matrix; $\mathbf{w} = [w_1, \ldots, w_p]^\top$ is a scalar vector of principal component variances; and $\mathbf{Y}$ is the $N \times p$ de-correlated feature matrix with the principal components on the columns. The covariance of matrix $\mathbf{A}$, $\operatorname{cov}(\mathbf{A}) = \frac{1}{N-1}(\mathbf{A}^\top\mathbf{A})$ if $\mathbf{A}$ is standardised.

E.4.1 Eigendecomposition method

One method for solving PCA uses matrix eigendecomposition (Shlens, 2014), where

$$\begin{aligned}
\operatorname{diag}(\mathbf{w}) &= \operatorname{cov}(\mathbf{ZW}) && \text{(E.86)}\\
&= \tfrac{1}{N-1}(\mathbf{ZW})^\top(\mathbf{ZW}) && \text{(E.87)}\\
&= \tfrac{1}{N-1}\mathbf{W}^\top\mathbf{Z}^\top\mathbf{ZW} && \text{(E.88)}\\
&= \mathbf{W}^\top \underbrace{\tfrac{1}{N-1}\mathbf{Z}^\top\mathbf{Z}}_{\operatorname{cov}(\mathbf{Z})}\mathbf{W}. && \text{(E.89)}
\end{aligned}$$

Because $\operatorname{cov}(\mathbf{Z})$ is, by definition, a $p \times p$ real symmetric matrix, its eigendecomposition is of the form $\operatorname{cov}(\mathbf{Z}) = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top$ where $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda})$ is a diagonal matrix of the p eigenvalues of $\operatorname{cov}(\mathbf{Z})$, $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_p]^\top$, and $\mathbf{Q}$ is an orthonormal matrix (meaning that $\mathbf{Q}^\top\mathbf{Q} = \mathbf{QQ}^\top = \mathbf{I}_p$) of the p eigenvectors of $\operatorname{cov}(\mathbf{Z})$. Therefore equation (E.89) becomes

$$\operatorname{diag}(\mathbf{w}) = \mathbf{W}^\top\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top\mathbf{W}. \tag{E.90}$$

Defining $\mathbf{W} = \mathbf{Q}$ gives

$$\text{diag}(\mathbf{w}) = \underbrace{\mathbf{Q}^\top \mathbf{Q}}_{=\mathbf{I}_p} \mathbf{\Lambda} \underbrace{\mathbf{Q}^\top \mathbf{Q}}_{=\mathbf{I}_p} \tag{E.91}$$

$$= \mathbf{\Lambda} \tag{E.92}$$

hence the PCA solution using eigendecomposition is

$$w_i = \mathbf{\Lambda}_{i,i} \quad 1 \leq i \leq p \quad \text{and} \tag{E.93}$$

$$\mathbf{W} = \mathbf{Q}. \tag{E.94}$$

E.4.2 Singular value decomposition method

Another, potentially more stable, method for solving PCA utilises the singular value decomposition (SVD), which factorises an $N \times p$ matrix $\mathbf{M}$ into the form $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U}$ is an $N \times N$ orthogonal matrix; $\mathbf{\Sigma}$ is an $N \times p$ diagonal matrix with the square roots of the non-zero eigenvalues of $\mathbf{M}^\top\mathbf{M}$ in descending order on the diagonal; and $\mathbf{V}$ is a $p \times p$ orthogonal matrix (Shlens, 2014; Barber, 2014).

Substituting $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ into equation (E.88) gives

$$\text{diag}(\mathbf{w}) = \tfrac{1}{N-1}\mathbf{W}^\top \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\right)^\top \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\right)\mathbf{W} \tag{E.95}$$

$$= \tfrac{1}{N-1}\mathbf{W}^\top\mathbf{V}\mathbf{\Sigma}^\top \underbrace{\mathbf{U}^\top\mathbf{U}}_{=\mathbf{I}_N}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{W} \tag{E.96}$$

$$= \tfrac{1}{N-1}\mathbf{W}^\top\mathbf{V}\mathbf{\Sigma}^\top\mathbf{\Sigma}\mathbf{V}^\top\mathbf{W}. \tag{E.97}$$

Similar to the eigendecomposition method described above, defining $\mathbf{W} = \mathbf{V}$ gives

$$\text{diag}(\mathbf{w}) = \tfrac{1}{N-1}\underbrace{\mathbf{V}^\top\mathbf{V}}_{=\mathbf{I}_N}\mathbf{\Sigma}^\top\mathbf{\Sigma}\underbrace{\mathbf{V}^\top\mathbf{V}}_{=\mathbf{I}_N} \tag{E.98}$$

$$= \tfrac{1}{N-1}\mathbf{\Sigma}^2 \tag{E.99}$$

hence the PCA solution using SVD is

$$w_i = \tfrac{1}{N-1}\mathbf{\Sigma}^2_{i,i} \quad 1 \leq i \leq p \quad \text{and} \tag{E.100}$$

$$\mathbf{W} = \mathbf{V}. \tag{E.101}$$

E.5 Expectation of an arbitrary distribution

This section derives the expectation of an arbitrary distribution in equation (7.42) from the definition of an expectation in a discrete distribution.

In the case of a continuous distribution, or a numerical discrete distribution, the expected value of a random variable, is defined as

$$\mathbb{E}(X) = \begin{cases} \int_{-\infty}^{\infty} xp(x)\,dx & \text{if } X \text{ is continuous; or} \\ \sum_{i=1}^{\infty} x_i p(x_i) & \text{if } X \text{ is discrete.} \end{cases} \tag{E.102}$$

In the case of having n samples from a distribution, the expectation is defined as

$$\hat{\mathbb{E}}(X) = c\sum_{i=1}^{n} x_i p(x_i), \tag{E.103}$$

where $c = 1/\sum_{i=1}^{n} p(x_i)$ is a normalisation constant to ensure that $\sum_{i=1}^{n} p(x_i) = 1$.

It is unclear from this how to calculate an expectation over an arbitrary distribution, such as discrete clusterings. Equation (E.103) uses a weighted sum of the the values of x_i to determine the expectation, but this does not make sense when x_i is a discrete clustering. An arbitrary offset, denoted x_j, can be applied to the expectation in equation (E.103) to write it as

$$\hat{\mathbb{E}}(X - x_j) = c\sum_{i=1}^{n} (x_i - x_j)p(x_i) \tag{E.104}$$

$$\hat{\mathbb{E}}(X) - x_j = c\sum_{i=1}^{n} (x_i - x_j)p(x_i) \tag{E.105}$$

$$\hat{\mathbb{E}}(X) = c\sum_{i=1}^{n} (x_i - x_j)p(x_i) + x_j. \tag{E.106}$$

By moving the expectation over to the right-hand side,

$$0 = c\sum_{i=1}^{n} (x_i - x_j)p(x_i) + x_j - \hat{\mathbb{E}}(X), \tag{E.107}$$

an alternative interpretation becomes clear, where the expectation can instead be framed as an optimisation problem. Denoting the optimisation variable x_j,

$$\hat{\mathbb{E}}(X) = \operatorname*{argmin}_{x_j} \left| c\sum_{i=1}^{n} (x_i - x_j)p(x_i) + x_j - x_j \right| \tag{E.108}$$

$$= \operatorname*{argmin}_{x_j} \left| \sum_{i=1}^{n} (x_i - x_j)p(x_i) \right|. \tag{E.109}$$

By replacing the absolute sum in equation (E.109) with equation (E.105), the equivalence of this optimisation to calculate an expectation can be demonstrated by writing equation (E.109) as

$$\hat{\mathbb{E}}(X) = \underset{x_j}{\operatorname{argmin}} \left| \hat{\mathbb{E}}(X) - x_j \right| \implies x_j = \hat{\mathbb{E}}(X) \text{ is a minimum.} \tag{E.110}$$

Furthermore, the optimisation function in equation (E.110) can be piecewise differentiated as

$$\frac{d\,\hat{\mathbb{E}}(X)}{dx_j} = \frac{d\,\hat{\mathbb{E}}(X)}{dx_j} \left| \hat{\mathbb{E}}(X) - x_j \right| \tag{E.111}$$

$$= \frac{d\,\hat{\mathbb{E}}(X)}{dx_j} \begin{cases} (\hat{\mathbb{E}}(X) - x_j) & \text{when } x_j < \hat{\mathbb{E}}(X) \\ -(\hat{\mathbb{E}}(X) - x_j) & \text{when } x_j > \hat{\mathbb{E}}(X) \end{cases} \tag{E.112}$$

$$= \begin{cases} -1 & \text{when } x_j < \hat{\mathbb{E}}(X) \\ 1 & \text{when } x_j > \hat{\mathbb{E}}(X) \end{cases} \tag{E.113}$$

demonstrating that $x_j = \hat{\mathbb{E}}(X)$ is the only minimum of equation (E.109).

Returning to equation (E.109), the term $(x_i - x_j)$ is just the difference between x_i and x_j, which can be substituted with $\Delta(x_i, x_j) = x_i - x_j$ in

$$\hat{\mathbb{E}}(X) = \underset{x_j}{\operatorname{argmin}} \left| \sum_{i=1}^{n} \Delta(x_i, x_j) p(x_i) \right|. \tag{E.114}$$

The function $\Delta(x_i, x_j)$ can now take any distance metric between pairs of points x_i and x_j.

It is now clear that with a suitable distance metric $\Delta(x_i, x_j)$, equation (7.42) can be used to calculate the expected value of an arbitrary distribution. This expected value is characteristic of the bulk of the posterior distribution.

References

Abdullah, S., Lane, N. D., & Choudhury, T. (2012). Towards Population Scale Activity Recognition: A Framework for Handling Data Diversity. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (pp. 851–857). Association for the Advancement of Artificial Intelligence.

Abdullah, S., Matthews, M., Frank, E., Doherty, G., Gay, G., & Choudhury, T. (2016). Automatic detection of social rhythms in bipolar disorder. *Journal of the American Medical Informatics Association*, *23*(3), 538–543.

Adams, K. B., Matto, H. C., & Sanders, S. (2004). Confirmatory Factor Analysis of the Geriatric Depression Scale. *The Gerontologist*, *44*(6), 818–826.

Alaa, A. M., Yoon, J., Hu, S., & van der Schaar, M. (2018). Personalized Risk Scoring for Critical Care Prognosis Using Mixtures of Gaussian Processes. *IEEE Transactions on Biomedical Engineering*, *65*(1), 207–218.

Al Agha, K., Pujolle, G., & Ali-Yahiya, T. (2016). *Mobile and Wireless Networks* (Vol. 2). Hoboken, NJ: Wiley-ISTE.

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, *28*(3), R1–R39.

Altman, E. G., Hedeker, D., Peterson, J. L., & Davis, J. M. (1997). The Altman Self-Rating Mania Scale. *Biological Psychiatry*, *42*(10), 948–955.

Alvarez-Lozano, J., Osmani, V., Mayora, O., Frost, M., Bardram, J., Faurholt-Jepsen, M., & Kessing, L. V. (2014). Tell me your apps and I will tell you your mood: Correlation of apps usage with Bipolar Disorder State. In *ACM proceedings of 7th International Conference on Pervasive Technologies Related to Assistive Environments*. ACM.

Amad, A., Ramoz, N., Thomas, P., Jardri, R., & Gorwood, P. (2014). Genetics of borderline personality disorder: Systematic review and proposal of an integrative model. *Neuroscience and Biobehavioral Reviews*, *40*, 6–19.

Andoni, A., & Indyk, P. (2008). Near-optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. *Communications of the ACM*, *51*(1), 117–122.

Andrews, D. F., & Mallows, C. L. (1974). Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(1), 99–102.

APA. (2000). *The Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR* (4th ed., text revision). Washington, DC: American Psychiatric Association.

APA. (2013). *The Diagnostic and Statistical Manual of Mental Disorders: DSM 5* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1027–1035). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Ashbrook, D., & Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, *7*(5), 275–286.

Askenfelt, A., & Sjölin-Nilsonne, Å. (1980). *Voice analysis in depressed patients: Rate of change of fundamental frequency related to mental state* (Quarterly Progress and Status Report (STL-QPSR), Vol. 21, No. 2-3, pp. 71–84). Stockholm, Sweden: Dept. for Speech, Music and Hearing, KTH Royal Institute of Technology.

Asmundson, G. J. G., Thibodeau, M. A., & Peluso, D. L. (2014). Somatic Symptom and Related Disorders. In D. C. Beidel, B. C. Frueh, & M. Hersen (Eds.), *Adult Psychopathology and Diagnosis* (pp. 451–471). Hoboken, NJ: Wiley.

Aylaz, R., Aktürk, Ü., Erci, B., Öztürk, H., & Aslan, H. (2012). Relationship between depression and loneliness in elderly and examination of influential factors. *Archives of Gerontology and Geriatrics*, *55*(3), 548–554.

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems* (Doctoral dissertation, School of Mathematical Sciences, Australian National University, Canberra, Australia). Retrieved from http://hdl.handle.net/1885/9449

Ball, J. S., & Links, P. S. (2009). Borderline personality disorder and childhood trauma: Evidence for a causal relationship. *Current Psychiatry Reports*, *11*(1), 63–68.

Barber, D. (2014). *Bayesian Reasoning and Machine Learning*. Cambridge: Cambridge University Press.

Barnett, I., & Onnela, J.-P. (2016, June 20). Inferring Mobility Measures from GPS Traces with Missing Data. *arXiv preprint arXiv:1606.06328v1*.

Barrio-Minton, C. (2017). Assessing Client Concerns. In J. S. Young & C. S. Cashwell (Eds.), *Clinical Mental Health Counseling: Elements of Effective Practice* (chap. 5). Thousand Oaks, CA: SAGE Publications.

Baryshnikov, I., Aaltonen, K., Koivisto, M., Näätänen, P., Karpov, B., Melartin, T., ... Isometsä, E. (2015). Differences and overlap in self-reported symptoms of bipolar disorder and borderline personality disorder. *European Psychiatry*, *30*(8), 914–919.

Beck, A. T. (2015). Theory of Personality Disorders. In A. T. Beck, A. Freeman, & D. D. Davis (Eds.), *Cognitive Therapy of Personality Disorders* (3rd ed., pp. 19–62). New York, NY: The Guilford Press.

Becker, A. E., & Kleinman, A. (2013). Mental Health and the Global Agenda. *New England Journal of Medicine*, *369*(1), 66–73.

Berwick, D. M., Nolan, T. W., & Whittington, J. (2008). The Triple Aim: Care, Health, And Cost. *Health Affairs*, *27*(3), 759–769.

Bickel, D. R., & Frühwirth, R. (2006). On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. *Computational Statistics & Data Analysis*, *50*(12), 3500–3530.

Bilderbeck, A. C., Saunders, K. E. A., Clifford, G. D., Tsanas, A., Ospiov, M., Harrison, P. J., ... Goodwin, G. M. (2015). Daily and weekly mood ratings: relative contributions to the differentiation of bipolar disorder and borderline personality disorder. *Bipolar Disorders*, *17*(S1), 129–130.

Biskin, R. S. (2015). The Lifetime Course of Borderline Personality Disorder. *The Canadian Journal of Psychiatry*, *60*(7), 303–308.

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, *46*, 109–132.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, *282*, 111–135.

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

Botella, C., Mira, A., Moragrega, I., García-Palacios, A., Bretón-López, J., Castilla, D., ... Baños, R. M. (2016). An Internet-based program for depression using activity and physiological sensors: efficacy, expectations, satisfaction, and ease of use. *Neuropsychiatric Disease and Treatment*, *12*, 393–406.

Botella, C., Moragrega, I., Baños, R., & García-Palacios, A. (2011). Online Predictive Tools for Intervention in Mental Illness: The OPTIMI Project. In J. D. Westwood et al. (Eds.), *Studies in Health Technology and Informatics: Vol. 163. Medicine Meets Virtual Reality 18* (pp. 86–92). Amsterdam, Netherlands: IOS Press.

Boullé, M., Guigourès, R., & Rossi, F. (2014). Nonparametric Hierarchical Clustering of Functional Data. In F. Guillet, B. Pinaud, G. Venturini, & D. A. Zighed (Eds.), *Advances in Knowledge Discovery and Management* (Vol. 4, pp. 15–35). Cham, Switzerland: Springer International Publishing.

Canzian, L., & Musolesi, M. (2015). Trajectories of Depression: Unobtrusive Monitoring of Depressive States by Means of Smartphone Mobility Traces Analysis. In *2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)* (pp. 1293–1304).

Carney, C. E., Edinger, J. D., Meyer, B., Lindman, L., & Istre, T. (2006). Daily activities and sleep quality in college students. *Chronobiology International*, *23*(3), 623–637.

Carr, O., Saunders, K. E. A., Bilderbeck, A. C., Tsanas, A., Palmius, N., Geddes, J. R., ... Goodwin, G. M. (2018b). Desynchronization of diurnal rhythms in bipolar disorder and borderline personality disorder. *Translational Psychiatry*, *8*(1), 79.

Carr, O., Saunders, K. E. A., Tsanas, A., Bilderbeck, A. C., Palmius, N., Geddes, J. R., ... De Vos, M. (2018a). Variability in phase and amplitude of diurnal rhythms is related to variation of mood in bipolar and borderline personality disorder. *Scientific Reports*, *8*(1), 1649.

Castro, J. C., Benseny, J., Colomé, J. M., & Martínez-Miranda, J. (2011). Help4Mood – Distributed System to Support Treatment of Patients with Major Depression. In M. Jordanova & F. Lievens (Eds.), *Proceedings of The International eHealth,*

Telemedicine and Health ICT Forum For Education, Networking and Business (MedeTel, 2011) (pp. 174–180).

Castro, J. C., Estevez, S., Aragonés, C., de Cerio, D. P. D., & Boqué, S. R. (2012). Help4Mood personal monitoring system: management of interoperatibilty in a wireless home for mental health treatment. In *Proceedings of the IADIS International Conferences on e-Health 2012* (pp. 139–145).

Castro, V. M., Minnier, J., Murphy, S. N., Kohane, I., Churchill, S. E., Gainer, V., ... Smoller, J. W., for the International Cohort Collection for Bipolar Disorder Consortium (2015). Validation of Electronic Health Record Phenotyping of Bipolar Disorder Cases and Controls. *American Journal of Psychiatry*, *172*(4), 363–372.

Cattrell, A., Harris, E. C., Palmer, K. T., Kim, M., Aylward, M., & Coggon, D. (2011). Regional trends in awards of incapacity benefit by cause. *Occupational Medicine*.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250.

Chen, M. Y., Sohn, T., Chmelev, D., Haehnel, D., Hightower, J., Hughes, J., ... Varshavsky, A. (2006). Practical Metropolitan-Scale Positioning for GSM Phones. In P. Dourish & A. Friday (Eds.), *UbiComp 2006: Ubiquitous Computing* (pp. 225–242). Springer Berlin Heidelberg.

Chow, P. I., Fua, K., Huang, Y., Bonelli, W., Xiong, H., Barnes, L. E., & Teachman, B. A. (2017). Using Mobile Sensing to Test Clinical Models of Depression, Social Anxiety, State Affect, and Social Isolation Among College Students. *Journal of Medical Internet Research*, *19*(3), e62.

Cipra, B. A. (2000). The Best of the 20th Century: Editors Name Top 10 Algorithms. *SIAM news*, *33*(4), 1–2.

Clifton, L., Clifton, D. A., Pimentel, M. A. F., Watkinson, P. J., & Tarassenko, L. (2012). Gaussian Process Regression in Vital-Sign Early Warning Systems. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 6161–6164).

Clifton, L., Clifton, D. A., Pimentel, M. A. F., Watkinson, P. J., & Tarassenko, L. (2013). Gaussian Processes for Personalized e-Health Monitoring With Wearable Sensors. *IEEE Transactions on Biomedical Engineering*, *60*(1), 193–197.

Conway, C. C., Bourgeois, M. L., & Brown, T. A. (2017). Clinical Interviewing with Adults. In S. G. Hofmann (Ed.), *Clinical Psychology: A Global Perspective* (pp. 29–41). Hoboken, NJ: Wiley-Blackwell.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Hoboken, NJ: Wiley-Blackwell.

CQC. (2017). *The state of care in mental health services 2014 to 2017: Findings from CQC's programme of comprehensive inspections of specialist mental health services* (Report No. CQC-380-072017). Newcastle upon Tyne, UK: Care Quality Commission.

Darby, J. K., & Hollien, H. (1977). Vocal and Speech Patterns of Depressive Patients. *Folia Phoniatrica et Logopaedica*, *29*(4), 279–291.

Dasgupta, A., Sun, Y. V., König, I. R., Bailey-Wilson, J. E., & Malley, J. D. (2011). Brief Review of Regression-Based and Machine Learning Methods in Genetic Epidemiology: The Genetic Analysis Workshop 17 Experience. *Genetic Epidemiology*, *35*(S1), S5–S11.

D'Avanzato, C., & Zimmerman, M. (2017). The Diagnosis and Assessment of Mood Disorders. In R. J. DeRubeis & D. R. Strunk (Eds.), *The Oxford Handbook of Mood Disorders* (pp. 95–107). Oxford, United Kingdom: Oxford University Press.

De Mello, M. T., Lemos, V. d. A., Antunes, H. K. M., Bittencourt, L., Santos-Silva, R., & Tufik, S. (2013). Relationship between physical activity and depression and anxiety symptoms: A population study. *Journal of Affective Disorders*, *149*(1-3), 241–246.

de Cerio, D. P.-D., Boqué, S. R., Rosell-Ferrer, J., Ramos-Castro, J., & Castro, J. C. (2012). A Wireless Sensor Network design for the Help4Mood european project. In *European Wireless 2012; 18th European Wireless Conference 2012* (pp. 1–6).

de Cerio, D. P.-D., Boqué, S. R., Rosell-Ferrer, J., Ramos-Castro, J., Valenzuela, J. L., & Colome, J. M. (2013). The Help4Mood Wearable Sensor Network for Inconspicuous Activity Measurement. *IEEE Wireless Communications*, *20*(4), 50–56.

de Montjoye, Y.-A., Quoidbach, J., Robic, F., & Pentland, A. S. (2013). Predicting Personality Using Novel Mobile Phone-Based Metrics. In A. M. Greenberg, W. G. Kennedy, & N. D. Bos (Eds.), *Social Computing, Behavioral-Cultural Modeling and Prediction* (Vol. 7812, pp. 48–55). Springer Berlin Heidelberg.

Deza, M. M., & Deza, E. (2016). *Encyclopedia of Distances* (4th ed.). Berlin, Heidelberg: Springer-Verlag.

Drake, G., Csipke, E., & Wykes, T. (2013). Assessing your mood online: acceptability and use of Moodscope. *Psychological Medicine*, *43*(7), 1455–1464.

Drucker, H. (1997). Improving Regressors Using Boosting Techniques. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)* (pp. 107–115). San Francisco, CA: Morgan Kaufmann Publishers.

Dunn, P., McKenna, H., & Murray, R. (2016, July). *Deficits in the NHS 2016.* The King's Fund.

Eckhardt, R. (1987). Stan Ulam, John von Neumann, and the Monte Carlo method. *Los Alamos Science*, *15*(Special Issue), 131–136.

Edel, A. (1995). *Aristotle and His Philosophy*. New Brunswick, New Jersey: Transaction Publishers.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499.

Ehlers, C. L., Frank, E., & Kupfer, D. J. (1988). Social zeitgebers and biological rhythms: A unified approach to understanding the etiology of depression. *Archives of General Psychiatry*, *45*(10), 948–952.

Ellgring, H., & Scherer, K. R. (1996). Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, *20*(2), 83–110.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* (pp. 226–231).

Estévez, P. A., Tesmer, M., Perez, C. A., & Zurada, J. M. (2009). Normalized Mutual Information Feature Selection. *IEEE Transactions on Neural Networks*, *20*(2), 189–201.

EuroQol Group. (1990). EuroQol – a new facility for the measurement of health-related quality of life. *Health Policy*, *16*(3), 199–208.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE: The Munich Versatile and Fast Open-source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 1459–1462). New York, NY: ACM.

Faggella, D. (2016, March 7). Valuing the Artificial Intelligence Market, Graphs and Predictions. *TechEmergence.com*. Retrieved from http://techemergence.com/valuing-the-artificial-intelligence-market-2016-and-beyond/

Faragher, R. (2012). Understanding the Basis of the Kalman Filter Via a Simple and Intuitive Derivation. *IEEE Signal Processing Magazine*, *29*(5), 128-132.

Farhan, A. A., Lu, J., Bi, J., Russell, A., & Wang, B. (2016b). *Multi-view Bi-clustering to Identify Smartphone Sensing Features Indicative of Depression.*

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., ... Wang, B. (2016a). *Behavior vs. Introspection: Refining prediction of clinical depression via smartphone sensing data.*

Fasel, M. (2011, October 14). *A Good Look at Android Location Data.* Retrieved from http://blog.shinetech.com/2011/10/14/a-good-look-at-android-location-data/

Faurholt-Jepsen, M., Brage, S., Vinberg, M., Christensen, E. M., Knorr, U., Jensen, H. M., & Kessing, L. V. (2012). Differences in psychomotor activity in patients suffering from unipolar and bipolar affective disorder in the remitted or mild/moderate depressive state. *Journal of Affective Disorders*, *141*(2), 457–463.

Faurholt-Jepsen, M., Busk, J., Frost, M., Vinberg, M., Christensen, E. M., Winther, O., ... Kessing, L. V. (2016a). Voice analysis as an objective state marker in bipolar disorder. *Translational Psychiatry*, *6*, e856.

Faurholt-Jepsen, M., Frost, M., Vinberg, M., Christensen, E. M., Bardram, J. E., & Kessing, L. V. (2014). Smartphone data as objective measures of bipolar disorder symptoms. *Psychiatry Research*, *217*(1), 124–127.

Faurholt-Jepsen, M., Vinberg, M., Christensen, E. M., Frost, M., Bardram, J., & Kessing, L. V. (2013). Daily electronic self-monitoring of subjective and objective symptoms in bipolar disorder—the MONARCA trial protocol (MONitoring, treAtment and pRediCtion of bipolAr disorder episodes): a randomised controlled single-blind trial. *BMJ Open*, *3*(7).

Faurholt-Jepsen, M., Vinberg, M., Frost, M., Christensen, E. M., Bardram, J. E., & Kessing, L. V. (2015). Smartphone data as an electronic biomarker of illness activity in bipolar disorder. *Bipolar Disorders*, *17*(7), 715–728.

Faurholt-Jepsen, M., Vinberg, M., Frost, M., Debel, S., Margrethe Christensen, E., Bardram, J. E., & Kessing, L. V. (2016b). Behavioral activities collected through

smartphones and the association with illness activity in bipolar disorder. *International Journal of Methods in Psychiatric Research*, *25*(4), 309–323.

Fava, M., Rush, A. J., Alpert, J. E., Balasubramani, G. K., Wisniewski, S. R., Carmin, C. N., ... Trivedi, M. H. (2008). Difference in Treatment Outcome in Outpatients With Anxious Versus Nonanxious Depression: A STAR*D Report. *American Journal of Psychiatry*, *165*(3), 342–351.

Feese, S., Muaremi, A., Arnrich, B., Troster, G., Meyer, B., & Jonas, K. (2011). Discriminating Individually Considerate and Authoritarian Leaders by Speech Activity Cues. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (pp. 1460–1465).

Feng, J. L. (2013). *Attack on WiFi-based Location Services and SSL Using Proxy Servers* (Master's thesis, University of Waterloo, Waterloo, Ontario, Canada). Retrieved from http://hdl.handle.net/10012/8116

Ferraty, F., & Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York, NY: Springer Science.

Frank, E., Gonzalez, J. M., & Fagiolini, A. (2006). The Importance of Routine for Preventing Recurrence in Bipolar Disorder. *American Journal of Psychiatry*, *163*(6), 981–985.

Frank, E., Hlastala, S., Ritenour, A., Houck, P., Tu, X. M., Monk, T. H., ... Kupfer, D. J. (1997). Inducing lifestyle regularity in recovering bipolar disorder patients: Results from the maintenance therapies in bipolar disorder protocol. *Biological Psychiatry*, *41*(12), 1165–1173.

Frank, E., Swartz, H. A., & Kupfer, D. J. (2000). Interpersonal and social rhythm therapy: managing the chaos of bipolar disorder. *Biological Psychiatry*, *48*(6), 593–604.

Franzen, P. L., & Buysse, D. J. (2008). Sleep disturbances and depression: Risk relationships for subsequent depression and therapeutic implications. *Dialogues in Clinical Neuroscience*, *10*(4), 473–481.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *1*(2), 302–332.

Fuster-Garcia, E., Bresó, A., Miranda, J. M., & García-Gómez, J. M. (2015). Actigraphy Pattern Analysis for Outpatient Monitoring. In C. Fernández-Llatas & J. M. García-Gómez (Eds.), *Data Mining in Clinical Medicine* (pp. 3–17). New York, NY: Springer New York.

Galper, D. I., Trivedi, M. H., Barlow, C. E., Dunn, A. L., & Kampert, J. B. (2006). Inverse Association between Physical Inactivity and Mental Health in Men and Women. *Medicine & Science in Sports & Exercise*, *38*(1), 173–178.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (Third ed.). Chapman & Hall/CRC.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741.

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for Depression in Medical Settings with the Patient Health Questionnaire (PHQ): A Diagnostic Meta-Analysis. *Journal of General Internal Medicine*, *22*(11), 1596–1602.

Gilburt, H. (2016, November). *Briefing: Mental health under pressure.* The King's Fund.

Gonzalez, R. (2014). The Relationship Between Bipolar Disorder and Biological Rhythms. *The Journal of Clinical Psychiatry*, *75*(4), e323–e331.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* Cambridge, Massachusetts: MIT Press.

Goodwin, R. D. (2003). Association between physical activity and mental disorders among adults in the United States. *Preventive Medicine*, *36*(6), 698–703.

Google. (2018, June 4). *Overview | Geolocation API | Google Developers.* Retrieved from https://web.archive.org/web/20180611091942/https://developers.google.com/maps/documentation/geolocation/intro

Groot, P. C. (2010). Patients can diagnose too: How continuous self-assessment aids diagnosis of, and recovery from, depression. *Journal of Mental Health*, *19*(4), 352–362.

Gruenerbl, A., Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., ... Lukowicz, P. (2014). Using Smart Phone Mobility Traces for the Diagnosis of Depressive and Manic Episodes in Bipolar Patients. In *Proceedings of the 5th Augmented Human International Conference (AH '14)* (pp. 38:1–38:8). ACM.

Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Ohler, S., Troster, G., ... Lukowicz, P. (2015). Smartphone-Based Recognition of States and State Changes in Bipolar Disorder Patients. *IEEE Journal of Biomedical and Health Informatics*, *19*(1), 140–148.

Gunderson, J. G. (2008). *Borderline Personality Disorder: A Clinical Guide* (2nd ed.). Arlington, VA: American Psychiatric Publishing.

Gunderson, J. G., Stout, R. L., McGlashan, T. H., Shea, M. T., Morey, L. C., Grilo, C. M., ... Skodol, A. E. (2011). Ten-Year Course of Borderline Personality Disorder: Psychopathology and Function From the Collaborative Longitudinal Personality Disorders Study. *Archives of General Psychiatry*, *68*(8), 827–837.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Haig, M. (2015). *Reasons to Stay Alive.* Edinburgh, UK: Canongate Books.

Ham, C., Dixon, A., & Brooke, B. (2012). *Transforming the Delivery of Health and Social Care: The case for fundamental change.* London, UK: The King's Fund.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second ed.). New York: Springer-Verlag.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations.* Chapman and Hall/CRC Press.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

Healy, D. (2010). From Mania to Bipolar Disorder. In L. N. Yatham & M. Maj (Eds.), *Bipolar Disorder: Clinical and Neurobiological Foundations* (chap. 1). Hoboken, NJ: Wiley-Blackwell.

Heckers, S. (2015). The value of psychiatric diagnoses. *JAMA Psychiatry*, *72*(12), 1165–1166.

Hensman, J., Rattray, M., & Lawrence, N. D. (2015). Fast Nonparametric Clustering of Structured Time-Series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(2), 383–393.

Hickie, I. B., Scott, J., Hermens, D. F., Scott, E. M., Naismith, S. L., Guastella, A. J., ... McGorry, P. D. (2013). Clinical classification in mental health at the cross-roads: which direction next? *BMC Medicine*, *11*(1), 125.

HMG/DH. (2011). *No health without mental: A cross-government mental health outcomes strategy for people of all ages*. London: Department of Health.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441.

Hu, Y., Wang, W., Suzuki, T., & Oyama-Higa, M. (2011). Characteristic Extraction of Mental Disease Patients by Nonlinear Analysis of Plethysmograms. In *AIP Conference Proceedings* (Vol. 1371, pp. 92–101). American Institute of Physics.

Hyman, S. E. (2010). The Diagnosis of Mental Disorders: The Problem of Reification. *Annual Review of Clinical Psychology*, *6*(1), 155–179.

Iacovides, A., Fountoulakis, K. N., Kaprinis, S., & Kaprinis, G. (2003). The relationship between job stress, burnout and clinical depression. *Journal of Affective Disorders*, *75*(3), 209–221.

Jacobi, F., Wittchen, H.-U., Hölting, C., Höfler, M., Pfister, H., Müller, N., & Lieb, R. (2004). Prevalence, co-morbidity and correlates of mental disorders in the general population: results from the German Health Interview and Examination Survey (GHS). *Psychological Medicine*, *34*(4), 597–611.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

Jaques, N., Taylor, S., Sano, A., & Picard, R. (2017). Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 202–208).

Johns, S. A., Kroenke, K., Krebs, E. E., Theobald, D. E., Wu, J., & Tu, W. (2013). Longitudinal Comparison of Three Depression Measures in Adult Cancer Patients. *Journal of Pain and Symptom Management*, *45*(1), 71–82.

Jones, K. D. (2010). The Unstructured Clinical Interview. *Journal of Counseling & Development*, *88*(2), 220–226.

Judd, L. L., Schettler, P. J., Akiskal, H. S., Maser, J., Coryell, W., Solomon, D., ... Keller, M. (2003). Long-term symptomatic status of bipolar I vs. bipolar II disorders. *International Journal of Neuropsychopharmacology*, *6*(2), 127–137.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Fluids Engineering*, *82*(1), 35–45.

Kendler, K. S. (2013). What psychiatric genetics has taught us about the nature of psychiatric illness and what is left to learn. *Mol Psychiatry*, *18*(10), 1058–1066.

Kessler, R. C. (2018). The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Current Opinion in Psychiatry*, *31*(1), 32–39.

Kessler, R. C., Merikangas, K. R., & Wang, P. S. (2007). Prevalence, Comorbidity, and Service Utilization for Mood Disorders in the United States at the Beginning of the Twenty-first Century. *Annual Review of Clinical Psychology*, *3*(1), 137–158.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, *220*(4598), 671–680.

Kolla, N. J., Eisenberg, H., & Links, P. S. (2008). Epidemiology, Risk Factors, and Psychopharmacological Management of Suicidal Behavior in Borderline Personality Disorder. *Archives of Suicide Research*, *12*(1), 1–19.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, *42*(8), 30–37.

Kraepelin, E. (1921). *Manic-Depressive Insanity and Paranoia* (G. M. Robertson, Ed. & R. Mary Barclay, Trans.). Edinburgh, United Kingdom: E. & S. Livingstone.

Kraskov, A., Stögbauer, H., Andrzejak, R. G., & Grassberger, P. (2005). Hierarchical clustering using mutual information. *EPL (Europhysics Letters)*, *70*(2), 278–284.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, *16*(9), 606–613.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1), 163–173.

LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., ... Schilit, B. (2005). Place Lab: Device Positioning Using Radio Beacons in the Wild. In H.-W. Gellersen, R. Want, & A. Schmidt (Eds.), *Pervasive Computing* (pp. 116–133). Springer Berlin Heidelberg.

Lamers, F., van Oppen, P., Comijs, H. C., Smit, J. H., Spinhoven, P., van Balkom, A. J. L. M., ... Penninx, B. W. J. H. (2011). Comorbidity Patterns of Anxiety and Depressive Disorders in a Large Cohort Study: the Netherlands Study of Depression and Anxiety (NESDA). *Journal of Clinical Psychiatry*, *72*(3), 341–348.

Lane, N. D., Xu, Y., Lu, H., Hu, S., Choudhury, T., Campbell, A. T., & Zhao, F. (2011). Enabling Large-scale Human Activity Inference on Smartphones using Community Similarity Networks (CSN). In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)* (pp. 355–364). New York, NY: ACM.

Lara, Ó. D., & Labrador, M. A. (2013). A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys & Tutorials*, *15*(3), 1192–1209.

Layard, R., Banerjee, S., Bell, S., Clark, D., Field, S., Knapp, M., ... Wessely, S. (2012, June). *How mental illness loses out in the NHS* (a report by The Centre for Economic Performance's Mental Health Policy Group). London, UK: London School of Economics.

Lee, S. I., Mortazavi, B., Hoffman, H. A., Lu, D. S., Li, C., Paak, B. H., ... Sarrafzadeh, M. (2016). A Prediction Model for Functional Outcomes in Spinal Cord Disorder Patients Using Gaussian Process Regression. *IEEE Journal of Biomedical and Health Informatics*, *20*(1), 91–99.

Leichsenring, F., Leibing, E., Kruse, J., New, A. S., & Leweke, F. (2011). Borderline personality disorder. *The Lancet*, *377*(9759), 74–84.

Leng, C., Tran, M.-N., & Nott, D. (2014). Bayesian adaptive Lasso. *Annals of the Institute of Statistical Mathematics*, *66*(2), 221–244.

Leszcz, M. (1989). Group Psychotherapy of the Characterologically Difficult Patient. *International Journal of Group Psychotherapy*, *39*(3), 311–335.

Lewis, S., & Higgins, N. (Eds.). (1996). *Brain imaging in psychiatry*. Oxford, United Kingdom: Blackwell Science.

Lieb, K., Zanarini, M. C., Schmahl, C., Linehan, M. M., & Bohus, M. (2004). Borderline personality disorder. *The Lancet*, *364*(9432), 453–461.

Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, *39*(2), 447–462.

Löwe, B., Spitzer, R. L., Gräfe, K., Kroenke, K., Quenter, A., Zipfel, S., ... Herzog, W. (2004). Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *Journal of Affective Disorders*, *78*(2), 131–140.

Lucas, C. P. (2003). Use of Structured Diagnostic Interviews in Clinical Child Psychiatric Practice. In M. B. First (Ed.), *Standardized Evaluation in Clinical Practice* (chap. 3). Washington, DC: American Psychiatric Publishing.

Lyalina, S., Percha, B., LePendu, P., Iyer, S. V., Altman, R. B., & Shah, N. H. (2013). Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. *Journal of the American Medical Informatics Association*, *20*(e2), e297–e305.

Ma, X., Yu, H., Wang, Y., & Wang, Y. (2015). Large-Scale Transportation Network Congestion Evolution Prediction Using Deep Learning Theory. *PLOS ONE*, *10*(3), e0119044.

Madsen, I. E. H., Nyberg, S. T., Magnusson Hanson, L. L., Ferrie, J. E., Ahola, K., Alfredsson, L., ... Kivimäki, M., for the IPD-Work Consortium (2017). Job strain as a risk factor for clinical depression: systematic review and meta-analysis with additional individual participant data. *Psychological Medicine*, *47*(8), 1342–1356.

Malik, A., Goodwin, G., & Holmes, E. (2011). Contemporary Approaches to Frequent Mood Monitoring in Bipolar Disorder. *Journal of Experimental Psychopathology*, *3*(4), 572–581.

Malkoff-Schwartz, S., Frank, E., Anderson, B., Sherrill, J. T., Siegel, L., Patterson, D., & Kupfer, D. J. (1998). Stressful life events and social rhythm disruption in the onset of manic and depressive bipolar episodes: A preliminary investigation. *Archives of General Psychiatry*, *55*(8), 702-707.

Malkoff-Schwartz, S., Frank, E., Anderson, B. P., Hlastala, S. A., Luther, J. F., Sherrill, J. T., ... Kupfer, D. J. (2000). Social rhythm disruption and stressful life events in the onset of bipolar and unipolar episodes. *Psychological Medicine*, *30*, 1005–1016.

Mathers, C., Fat, D. M., & Boerma, J. T. (2008). *The global burden of disease: 2004 update*. World Health Organization.

Matic, A., & Oliver, N. (2016). The Untapped Opportunity of Mobile Network Data for Mental Health. In *2016 10th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '16)* (pp. 285–288).

Matthews, M., Murnane, E., Snyder, J., Guha, S., Chang, P., Doherty, G., & Gay, G. (2017). The double-edged sword: A mixed methods study of the interplay between bipolar disorder and technology use. *Computers in Human Behavior*, *75*, 288–300.

Mayes, R., & Horwitz, A. V. (2005). DSM-III and the revolution in the classification of mental illness. *Journal of the History of the Behavioral Sciences*, *41*(3), 249–267.

Mayora, O., Arnrich, B., Bardram, J., Dräger, C., Finke, A., Frost, M., ... Wurzer, G. (2013). Personal Health Systems for Bipolar Disorder: Anecdotes, Challenges and Lessons Learnt from MONARCA project. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops* (pp. 424–429).

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, *16*(3), 285–292.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (Second ed.). Chapman & Hall/CRC.

McDaid, D., & Knapp, M. (2010). Black-skies planning? Prioritising mental health services in times of austerity. *The British Journal of Psychiatry*, *196*(6), 423 – 424.

Mehrotra, A., Hendley, R., & Musolesi, M. (2016). Towards Multi-modal Anticipatory Monitoring of Depressive States Through the Analysis of Human-smartphone Interaction. In *2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)* (pp. 1132–1138).

Mehrotra, A., & Musolesi, M. (2017). Designing Effective Movement Digital Biomarkers for Unobtrusive Emotional State Mobile Monitoring. In *2017 1st Workshop on Digital Biomarkers (DigitalBiomarkers '17)* (pp. 3–8).

Mehrotra, A., & Musolesi, M. (2018). Using Unsupervised Deep Autoencoders to Automatically Extract Mobility Features for Predicting Depressive States. In *2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '18).* (Accepted for presentation)

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, *98*(5), 873–895.

Mental Health Taskforce. (2016, February). *NHS Five Year Forward View for Mental Health: A report from the independent Mental Health Taskforce to the NHS in England.*

Merikangas, K. R., Jin, R., He, J.-P., Ronald C. Kessler, Lee, S., Sampson, N. A., ... Zarkov, Z. (2011). Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Archives of General Psychiatry*, *68*(3), 241–251.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science*, *15*(Special Issue), 125–130.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association*, *44*(247), 335–341.

Mikelsons, G., Smith, M., Mehrotra, A., & Musolesi, M. (2017). Towards Deep Learning Models for Psychological State Prediction using Smartphone Data: Challenges and Opportunities. In *Machine Learning for Health (ML4H) Workshop at the 31st Conference on Neural Information Processing Systems (NIPS 2017).*

Miller, C. J., Johnson, S. L., & Eisner, L. (2009). Assessment Tools for Adult Bipolar Disorder. *Clinical Psychology: Science and Practice*, *16*(2), 188–201.

Miller, G. (2010). Beyond DSM: Seeking a Brain-Based Classification of Mental Illness. *Science*, *327*(5972), 1437.

Millon, T., Grossman, S., Millon, C., Meagher, S., & Ramnath, R. (2004). *Personality Disorders in Modern Life* (2nd ed.). Hoboken, NJ: Wiley.

Moffitt, T. E., Caspi, A., Taylor, A., Kokaua, J., Milne, B. J., Polanczyk, G., & Poulton, R. (2010). How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. *Psychological Medicine*, *40*(6), 899–909.

Moffitt, T. E., Harrington, H., Caspi, A., Kim-Cohen, J., Goldberg, D., Gregory, A. M., & Poulton, R. (2007). Depression and generalized anxiety disorder: Cumulative and sequential comorbidity in a birth cohort followed prospectively to age 32 years. *Archives of General Psychiatry*, *64*(6), 651–660.

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology*, *13*(1), 23–47.

Monk, T. H., Reynolds III, C. F., Buysse, D. J., DeGrazia, J. M., & Kupfer, D. J. (2003). The Relationship Between Lifestyle Regularity and Subjective Sleep Quality. *Chronobiology International*, *20*(1), 97–107.

Monteith, S., Glenn, T., Geddes, J., Whybrow, P. C., & Bauer, M. (2016). Big data for bipolar disorder. *International Journal of Bipolar Disorders*, *4*(1), 10.

Moran, P., & Crawford, M. J. (2013). Assessing the severity of borderline personality disorder. *The British Journal of Psychiatry*, *203*(3), 163–164.

Moreno, C., Hasin, D. S., Arango, C., Oquendo, M. S., Vieta, E., Liu, S., ... Blanco, C. (2012). Depression in bipolar disorder versus major depressive disorder: Re-

sults from the National Epidemiologic Survey on Alcohol and Related Conditions. *Bipolar Disorders*, *14*(3), 271–282.

Morriss, R., Faizal, M. A., Jones, A. P., Williamson, P. R., Bolton, C. A., & McCarthy, J. P. (2007). Interventions for helping people recognise early signs of recurrence in bipolar disorder. *Cochrane Database of Systematic Reviews*(1).

Muaremi, A., Gravenhorst, F., Grünerbl, A., Arnrich, B., & Tröster, G. (2014). Assessing Bipolar Episodes Using Speech Cues Derived from Phone Calls. In P. Cipresso, A. Matic, & G. Lopez (Eds.), *4th International Symposium on Pervasive Computing Paradigms for Mental Health (MindCare 2014)* (pp. 103–114). Cham, Switzerland: Springer.

Murnane, E. L., Cosley, D., Chang, P., Guha, S., Frank, E., Gay, G., & Matthews, M. (2016). Self-monitoring practices, attitudes, and needs of individuals with bipolar disorder: Implications for the design of technologies to manage mental health. *Journal of the American Medical Informatics Association*, *23*(3), 477–484.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, Massachusetts: MIT Press.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1).

Nasrallah, H. A. (2013). Lab tests for psychiatric disorders: Few clinicians are aware of them. *Current Psychiatry*, *12*(2), 5–7.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384.

Nestler, E. J., Peña, C. J., Kundakovic, M., Mitchell, A., & Akbarian, S. (2016). Epigenetic Basis of Mental Illness. *The Neuroscientist*, *22*(5), 447–463.

Nguyen, X., & Gelfand, A. E. (2011). The Dirichlet labeling process for clustering functional data. *Statistica Sinica*, *21*(3), 1249–1289.

NHS. (2014, October). *Five Year Forward View*. NHS England. Retrieved from https://www.england.nhs.uk/ourwork/futurenhs/

NHS England. (2016, July). *Implementing the Five Year Forward View for Mental Health* (Gateway Reference No. 05574). Redditch, UK: NHS England.

NICE. (2009, November). *Promoting mental wellbeing at work* (NICE public health guidance No. 22). Manchester, UK: National Institute for Health and Clinical Excellence.

Nilsonne, Å., Sundberg, J., Ternström, S., & Askenfelt, A. (1988). Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. *The Journal of the Acoustical Society of America*, *83*(2), 716–728.

Osborne, M. R., Presnell, B., & Turlach, B. A. (2000). On the LASSO and its Dual. *Journal of Computational and Graphical Statistics*, *9*(2), 319–337.

Osipov, M. (2016). *Towards automated symptoms assessment in mental health* (Doctoral dissertation, University of Oxford, Oxford, UK). Retrieved from https://ora.ox.ac.uk/objects/uuid:42111684-8801-440e-8fbb-00f779d806ee

Owen, M. J. (2014). New Approaches to Psychiatric Diagnostic Classification. *Neuron*, *84*(3), 564–571.

Oyebode, F. (2015). *Sims' Symptoms in the Mind: Textbook of Descriptive Psychopathology* (5th ed.). Philadelphia, PA: Saunders/Elsevier.

Palmius, N., & De Vos, M. (2016). Non-Parametric Subset Selection for Personalised Time-Series Regression. In *Practical Bayesian Nonparametrics Workshop at the 30th Conference on Neural Information Processing Systems (NIPS 2016)*.

Palmius, N., Osipov, M., Bilderbeck, A. C., Goodwin, G. M., Saunders, K., Tsanas, A., & Clifford, G. D. (2014). A multi-sensor monitoring system for objective mental health management in resource constrained environments. In *Appropriate Healthcare Technologies for Low Resource Settings (AHT 2014)*. doi: 10.1049/cp.2014.0764

Palmius, N., Saunders, K. E. A., Carr, O., Geddes, J. R., Goodwin, G. M., & De Vos, M. (n.d.). Group-personalized regression models for prediction of mental health scores from objective mobile phone data streams. *Journal of Medical Internet Research*. doi: 10.2196/10194

Palmius, N., Tsanas, A., Saunders, K. E. A., Bilderbeck, A. C., Geddes, J. R., Goodwin, G. M., & De Vos, M. (2017). Detecting Bipolar Depression From Geographic Location Data. *IEEE Transactions on Biomedical Engineering*, *64*(8), 1761–1771. doi: 10.1109/TBME.2016.2611862

Paris, J. (2005). The Diagnosis of Borderline Personality Disorder: Problematic but Better Than the Alternatives. *Annals of Clinical Psychiatry*, *17*(1), 41–46.

Paris, J., Brown, R., & Nowlis, D. (1987). Long-term follow-up of borderline patients in a general hospital. *Comprehensive Psychiatry*, *28*(6), 530–535.

Paris, J., & Zweig-Frank, H. (2001). A 27-year follow-up of patients with borderline personality disorder. *Comprehensive Psychiatry*, *42*(6), 482–487.

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Pazzani, M. J., & Billsus, D. (2007). Content-Based Recommendation Systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web: Methods and Strategies of Web Personalization* (pp. 325–341). Berlin, Heidelberg: Springer Berlin Heidelberg.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, *2*(11), 559–572.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238.

Perlis, R. H., Ostacher, M. J., Uher, R., Nierenberg, A. A., Casamassima, F., Kansky, C., ... Sachs, G. S. (2009). Stability of symptoms across major depressive episodes in bipolar disorder. *Bipolar Disorders*, *11*(8), 867–875.

Perry, J. C. (2001). A pilot study of defenses in adults with personality disorders entering psychotherapy. *The Journal of nervous and mental disease*, *189*(10), 651–660.

Perry, J. C., Banon, E., & Ianni, F. (1999). Effectiveness of Psychotherapy for Personality Disorders. *American Journal of Psychiatry*, *156*(9), 1312–1321.

Petersen, J., Austin, D., Kaye, J. A., Pavel, M., & Hayes, T. L. (2014). Unobtrusive In-Home Detection of Time Spent Out-of-Home With Applications to Loneliness and Physical Activity. *IEEE Journal of Biomedical and Health Informatics*, *18*(5), 1590–1596.

Petersen, J., Thielke, S., Austin, D., & Kaye, J. (2016). Phone behaviour and its relationship to loneliness in older adults. *Aging & Mental Health*, *20*(10), 1084–1091.

Pham, T. D., Thang, T. C., Oyama-Higa, M., & Sugiyama, M. (2013). Mental-disorder detection using chaos and nonlinear dynamical analysis of photoplethysmographic signals. *Chaos, Solitons & Fractals*, *51*, 64–74.

Pimentel, M. A. F., Clifton, D. A., & Tarassenko, L. (2013). Gaussian Process Clustering for the Functional Characterisation of Vital-Sign Trajectories. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*.

Pitman, J. (2006). *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilitiés de Saint-Flour XXXII - 2002* (J. Picard, Ed.). Berlin: Springer Berlin Heidelberg.

Pollak, J. P., Adams, P., & Gay, G. (2011). PAM: A Photographic Affect Meter for Frequent, in Situ Measurement of Affect. In *SIGCHI Conference on Human Factors in Computing Systems (CHI '11)* (pp. 725–734).

Prigerson, H. G., Monk, T. H., Reynolds, C. F., Begley, A., Houck, P. R., Bierhals, A. J., & Kupfer, D. J. (1995). Lifestyle regularity and activity level as protective factors against bereavement-related depression in late-life. *Depression*, *3*(6), 297–302.

Prince, M., Patel, V., Saxena, S., Maj, M., Maselko, J., Phillips, M. R., & Rahman, A. (2007). No health without mental health. *The Lancet*, *370*(9590), 859–877.

Prince, S. A., Adamo, K. B., Hamel, M. E., Hardt, J., Gorber, S. C., & Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, *5*(1), 56.

Puiatti, A., Mudda, S., Giordano, S., & Mayora, O. (2011). Smartphone-centred wearable sensors network for monitoring patients with bipolar disorder. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 3644–3647).

Qin, L., & Zhu, X. (2013). Functional Dirichlet Process. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (CIKM '13)* (pp. 1331–1340). New York, NY: ACM.

Rahman, M. S., Park, Y., & Kim, K.-D. (2012). RSS-Based Indoor Localization Algorithm for Wireless Sensor Network Using Generalized Regression Neural Network. *Arabian Journal for Science and Engineering*, *37*(4), 1043–1053.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis* (2nd ed.). New York, NY: Springer Science.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT Press.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. *American Journal of Psychiatry*, *170*(1), 59–70.

Reid, S. C., Kauer, S. D., Dudgeon, P., Sanci, L. A., Shrier, L. A., & Patton, G. C. (2009). A mobile phone program to track young people's experiences of mood, stress and coping. *Social Psychiatry and Psychiatric Epidemiology*, *44*(6), 501–507.

Reid, S. C., Kauer, S. D., Hearps, S. J. C., Crooke, A. H. D., Khor, A. S., Sanci, L. A., & Patton, G. C. (2011). A mobile phone application for the assessment and management of youth mental health problems in primary care: a randomised controlled trial. *BMC Family Practice*, *12*(1), 131.

Rhee, I., Shin, M., Hong, S., Lee, K., & Chong, S. (2007). Human Mobility Patterns and Their Impact on Routing in Human-Driven Mobile Networks. In *Proceedings of the Sixth Workshop on Hot Topics in Networks (HotNets-VI).*

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J., & Chong, S. (2011). On the Levy-Walk Nature of Human Mobility. *IEEE/ACM Transactions on Networking*, *19*(3), 630–643.

Riva, G., Banos, R., Botella, C., Gaggioli, A., & Wiederhold, B. K. (2011). Personal Health Systems for Mental Health: The European Projects. In J. D. Westwood et al. (Eds.), *Studies in Health Technology and Informatics: Vol. 163. Medicine Meets Virtual Reality 18* (pp. 496–502). Amsterdam, Netherlands: IOS Press.

Robertson, T., & Cryer, J. D. (1974). An Iterative Procedure for Estimating the Mode. *Journal of the American Statistical Association*, *69*(348), 1012–1016.

Robinson, L. J., Thompson, J. M., Gallagher, P., Goswami, U., Young, A. H., Ferrier, I. N., & Moore, P. B. (2006). A meta-analysis of cognitive deficits in euthymic patients with bipolar disorder. *Journal of Affective Disorders*, *93*(1–3), 105–115.

Rosch, E. (1999). Principles of Categorization. In E. Margolis & S. Laurence (Eds.), *Concepts: Core Readings* (pp. 189–206). Cambridge, Massachusetts: MIT Press. A Bradford Book.

Rush, A., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., ... Keller, M. B. (2003). 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), 573–583.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sabatelli, M., Osmani, V., Mayora, O., Gruenerbl, A., & Lukowicz, P. (2014). Correlation of significant places with self-reported state of bipolar disorder patients. In *2014 4th International Conference on Wireless Mobile Communication and Healthcare - Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)* (pp. 116–119).

Saeb, S., Lattie, E. G., Kording, K. P., & Mohr, D. C. (2017). Mobile Phone Detection of Semantic Location and Its Relationship to Depression and Anxiety. *JMIR mHealth and uHealth*, *5*(8), e112.

Saeb, S., Lattie, E. G., Schueller, S. M., Kording, K. P., & Mohr, D. C. (2016). The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ*, *4*, e2537.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015a). Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Journal of Medical Internet Research*, *17*(7), e175.

Saeb, S., Zhang, M., Kwasny, M., Karr, C. J., Kording, K., & Mohr, D. C. (2015b). The relationship between clinical, momentary, and sensor-based assessment of depression. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2015)* (pp. 229–232).

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Sandwell, D. T. (1987). Biharmonic spline interpolation of GEOS-3 and SEASAT altimeter data. *Geophysical Research Letters*, *14*(2), 139–142.

Saunders, K., Bilderbeck, A. C., Panchal, P., Brett, D., Clifford, G. D., Harmer, C. J., ... Goodwin, G. M. (2015a). Acceptability and tolerability of ambulatory monitoring in bipolar disorder: a patient perspective. *Bipolar Disorders*, *17*(S1), 86.

Saunders, K., Tsanas, A., Bilderbeck, A. C., & Goodwin, G. (2017a). Clinical differentiation of bipolar disorder vs borderline personality disorder. The role of mood variability. *Bipolar Disorders*, *19*(S1), 35–36.

Saunders, K. E. A., Bilderbeck, A. C., Panchal, P., Atkinson, L. Z., Geddes, J. R., & Goodwin, G. M. (2017b). Experiences of remote mood and activity monitoring in bipolar disorder: A qualitative study. *European Psychiatry*, *41*, 115–121.

Saunders, K. E. A., Bilderbeck, A. C., Price, J., & Goodwin, G. M. (2015b). Distinguishing bipolar disorder from borderline personality disorder: A study of current clinical practice. *European Psychiatry*, *30*(8), 965–974.

Saunders, K. E. A., Bilderbeck, A. C., Tsanas, A., Harrison, P. J., Harmer, C. J., Nobre, A. C., ... Goodwin, G. M. (2016). Stable instability: characterising the nature and degree of mood instability in bipolar disorder and borderline personality disorder. *Bipolar Disorders*, *18*(S1), 153.

Scargle, J. D. (1982). Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, *263*, 835–853.

Schleidgen, S., Klingler, C., Bertram, T., Rogowski, W. H., & Marckmann, G. (2013). What is personalized medicine: sharpening a vague term based on a systematic literature review. *BMC Medical Ethics*, *14*(1), 55.

Schmidt-Dannert, A. (2010). Positioning Technologies and Mechanisms for mobile Devices. In *Seminar Master Module SNET2*. TU-Berlin.

Segal, D. L., & Williams, K. N. (2014). Structured and Semistructured Interviews for Differential Diagnosis: Fundamental Issues, Applications, and Features. In D. C. Beidel, B. C. Frueh, & M. Hersen (Eds.), *Adult Psychopathology and Diagnosis* (pp. 103–129). Hoboken, NJ: Wiley.

Servia-Rodríguez, S., Rachuri, K. K., Mascolo, C., Rentfrow, P. J., Lathia, N., & Sandstrom, G. M. (2017). Mobile Sensing at the Service of Mental Well-being: A Large-scale Longitudinal Study. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)* (pp. 103–112).

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shenkin, P. S., Erman, B., & Mastrandrea, L. D. (1991). Information-theoretical entropy as a measure of sequence variability. *Proteins: Structure, Function, and Bioinformatics*, *11*(4), 297–313.

Shlens, J. (2014). A Tutorial on Principal Component Analysis. *arXiv preprint arXiv: 1404.1100*.

Silver, D. (1983). Psychotherapy of the Characterologically Difficult Patient. *The Canadian Journal of Psychiatry*, *28*(7), 513–521.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Singh, I., & Rose, N. (2009). Biomarkers in psychiatry. *Nature*, *460*(7252), 202–207.

Singhal, A., Sinha, P., & Pant, R. (2017). Use of Deep Learning in Modern Recommendation System: A Summary of Recent Works. *International Journal of Computer Applications*, *180*(7), 17–22.

Sobrado, B. (2018). *Personalised Map Matched Imputation: imputing missing data from smartphone location logs* (Unpublished master's thesis). Department of Methodology & Statistics, Utrecht University, Utrecht, Netherlands.

Solomon, D. A., Leon, A. C., Coryell, W. H., Endicott, J., Li, C., Fiedorowicz, J. G., ... Keller, M. B. (2010). Longitudinal course of bipolar I disorder: Duration of mood episodes. *Archives of General Psychiatry*, *67*(4), 339–347.

Soreca, I., Frank, E., & Kupfer, D. J. (2009). The phenomenology of bipolar disorder: what drives the high rate of medical burden and determines long-term prognosis? *Depression and Anxiety*, *26*(1), 73–82.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.

Stachl, C., Hilbert, S., Au, J.-Q., Buschek, D., De Luca, A., Bischl, B., ... Bühner, M. (2017). Personality Traits Predict Smartphone Usage. *European Journal of Personality*.

Stern, A. (1983). Psychoanalytic Investigation of and Therapy in the Border Line Group of Neuroses. *Psychoanalytic Quarterly*, *7*, 467–489.

Stone, M. H. (1990). *The Fate of Borderline Patients: Successful Outcome and Psychiatric Practice*. New York, NJ: The Guilford Press.

Strejilevich, S. A., Martino, D. J., Murru, A., Teitelbaum, J., Fassi, G., Marengo, E., ... Colom, F. (2013). Mood instability and functional recovery in bipolar disorders. *Acta Psychiatrica Scandinavica*, *128*(3), 194–202.

Suhara, Y., Xu, Y., & Pentland, A. (2017). DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)* (pp. 715–724).

Sun, J., Lu, J., Xu, T., & Bi, J. (2015). Multi-view Sparse Co-clustering via Proximal Alternating Linearized Minimization. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning* (Vol. 37, pp. 757–766).

Suppes, T., & Dennehy, E. B. (2010). *Bipolar Disorder: Assessment and Treatment* (2nd ed.). Sudbury, MA: Jones & Bartlett Learning.

Svrakic, D. M., Draganic, S., Hill, K., Bayon, C., Przybeck, T. R., & Cloninger, C. R. (2002). Temperament, character, and personality disorders: etiologic, diagnostic, treatment issues. *Acta Psychiatrica Scandinavica*, *106*(3), 189–195.

Szasz, T. S. (1960). The myth of mental illness. *American Psychologist*, *15*(2), 113–118.

Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proceedings of the IEEE*, *105*(12), 2295–2329.

Thayer, J. F., & Lane, R. D. (2009). Claude Bernard and the heart–brain connection: Further elaboration of a model of neurovisceral integration. *Neuroscience & Biobehavioral Reviews*, *33*(2), 81–88.

Thielke, S. M., Mattek, N. C., Hayes, T. L., Dodge, H. H., Quiñones, A. R., Austin, D., ... Kaye, J. A. (2014). Associations Between Observed In-Home Behaviors and Self-Reported Low Mood in Community-Dwelling Older Adults. *Journal of the American Geriatrics Society*, *62*(4), 685–689.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tsakalidis, A., Liakata, M., Damoulas, T., Jellinek, B., Guo, W., & Cristea, A. I. (2016). Combining heterogeneous user generated data to sense well-being. In *26th International Conference on Computational Linguistics (COLING 2016)*.

Tsanas, A., Saunders, K., Bilderbeck, A., Palmius, N., Goodwin, G., & De Vos, M. (2017). Clinical Insight Into Latent Variables of Psychiatric Questionnaires for Mood Symptom Self-Assessment. *JMIR Mental Health*, *4*(2), e15.

Tsanas, A., Saunders, K. E. A., Bilderbeck, A. C., Palmius, N., Osipov, M., Clifford, G. D., ... De Vos, M. (2016). Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder. *Journal of Affective Disorders*, *205*, 225–233.

Turkington, C. (2002). *The Encyclopedia of the Brain and Brain Disorders* (2nd ed.). New York, NY: Facts on File.

Tuso, P. (2014). Treatment Progress Indicator: Application of a New Assessment Tool to Objectively Monitor the Therapeutic Progress of Patients With Depression, Anxiety, or Behavioral Health Impairment. *The Permanente journal*, *18*(3).

Uher, R. (2014). Gene–Environment Interactions in Severe Mental Illness. *Frontiers in Psychiatry*, *5*, 48.

UNC Vision Lab. (2015). *ILSVRC2015 Results.* Retrieved January 4th, 2018, from https://web.archive.org/web/20180104233932/http://image-net.org/challenges/LSVRC/2015/results

Valenza, G., Gentili, C., Lanatà, A., & Scilingo, E. P. (2013). Mood recognition in bipolar patients through the PSYCHE platform: Preliminary evaluations and perspectives. *Artificial Intelligence in Medicine*, *57*(1), 49–58.

van Beljouw, I. M. J., Verhaak, P. F. M., Cuijpers, P., van Marwijk, H. W. J., & Penninx, B. W. J. H. (2010). The course of untreated anxiety and depression, and determinants of poor one-year outcome: a one-year cohort study. *BMC Psychiatry*, *10*(1), 86.

van Loo, H. M., de Jonge, P., Romeijn, J.-W., Kessler, R. C., & Schoevers, R. A. (2012). Data-driven subtypes of major depressive disorder: a systematic review. *BMC Medicine*, *10*(1), 156.

van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworths.

van Rooij, J. G. J., Jhamai, M., Arp, P. P., Nouwens, S. C. A., Verkerk, M., Hofman, A., ... Kraaij, R. (2017). Population-specific genetic variation in large sequencing data sets: why more data is still better. *European Journal Of Human Genetics*, *25*, 1173–1175.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*(Oct), 2837–2854.

Wahle, F., Kowatsch, T., Fleisch, E., Rufer, M., & Weidt, S. (2016). Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild. *JMIR mHealth and uHealth*, *4*(3), e111.

Wallach, H., Jensen, S., Dicker, L., & Heller, K. (2010). An Alternative Prior Process for Nonparametric Bayesian Clustering. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 892–899).

Wang, H., & Leng, C. (2007). Unified LASSO Estimation by Least Squares Approximation. *Journal of the American Statistical Association*, *102*(479), 1039–1048.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., ... Campbell, A. T. (2014). StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)* (pp. 3–14).

Weyand, T., Kostrikov, I., & Philbin, J. (2016). PlaNet - Photo Geolocation with Convolutional Neural Networks. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 37–55).

Wielopolski, J., Reich, K., Clepce, M., Fischer, M., Sperling, W., Kornhuber, J., & Thuerauf, N. (2015). Physical activity and energy expenditure during depressive episodes of major depression. *Journal of Affective Disorders*, *174*(Supplement C), 310–316.

Wikberg, C., Pettersson, A., Westman, J., Björkelund, C., & Petersson, E.-L. (2016). Patients' perspectives on the use of the Montgomery-Asberg depression rating scale self-assessment version in primary care. *Scandinavian Journal of Primary Health Care*, *34*(4), 434–442.

Wikberg, C., Westman, J., Petersson, E.-L., Larsson, M. E. H., André, M., Eggertsen, R., ... Björkelund, C. (2017). Use of a self-rating scale to monitor depression severity in recurrent GP consultations in primary care – does it really make a difference? A randomised controlled study. *BMC Family Practice*, *18*(1), 6.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance . *Climate Research*, *30*(1), 79–82.

Wilson, S. J., Wong, D., Clifton, D., Fleming, S., Way, R., Pullinger, R., & Tarassenko, L. (2013). Track and trigger in an emergency department: an observational evaluation study. *Emergency Medicine Journal*, *30*(3), 186–191.

Wolfson, A. R., & Carskadon, M. A. (1998). Sleep Schedules and Daytime Functioning in Adolescents. *Child Development*, *69*(4), 875–887.

Wong, D., Bonnici, T., Knight, J., Gerry, S., Turton, J., & Watkinson, P. (2017). A ward-based time study of paper and electronic documentation for recording vital sign observations. *Journal of the American Medical Informatics Association*, *24*(4), 717–721.

World Health Organization. (2003). *Investing in mental health*. Geneva: World Health Organization.

World Health Organization. (2011). *Mental Health Atlas 2011*. Geneva: World Health Organization.

World Health Organization Europe. (2005). *Mental health: facing the challenges, building solutions – Report from the WHO European Ministerial Conference*. Copenhagen: World Health Organization Regional Office for Europe.

Wu, Z., & Fang, Y. (2014). Comorbidity of depressive and anxiety disorders: challenges in diagnosis and assessment. *Shanghai Archives of Psychiatry*, *26*(4), 227–231.

Xu, G., Gao, S., Daneshmand, M., Wang, C., & Liu, Y. (2017). A Survey for Mobility Big Data Analytics for Geolocation Prediction. *IEEE Wireless Communications*, *24*(1), 111–119.

Yang, J., Varshavsky, A., Liu, H., Chen, Y., & Gruteser, M. (2010). Accuracy Characterization of Cell Tower Localization. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (pp. 223–226). New York, NY, USA: ACM.

Yao, Y. Y. (2003). Information-Theoretic Measures for Knowledge Discovery and Data Mining. In Karmeshu (Ed.), *Entropy Measures, Maximum Entropy Principle and Emerging Applications* (pp. 115–136). Berlin, Heidelberg: Springer.

Yu, L., & Liu, H. (2004). Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research*, *5*, 1205–1224.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Yue, C., Ware, S., Morillo, R., Lu, J., Shang, C., Bi, J., ... Wang, B. (2017). Fusing Location Data for Depression Prediction. In *2017 IEEE 14th International Conference on Ubiquitous Intelligence and Computing (UIC 2017)*.

Zanarini, M. C., Frankenburg, F. R., Reich, D. B., Silk, K. R., Hudson, J. I., & McSweeney, L. B. (2007). The Subsyndromal Phenomenology of Borderline Personality Disorder: A 10-Year Follow-Up Study. *American Journal of Psychiatry*, *164*(6), 929–935.

Zanarini, M. C., Frankenburg, F. R., Ridolfi, M. E., Jager-Hyman, S., Hennen, J., & Gunderson, J. G. (2006). Reported Childhood Onset of Self-Mutilation Among Borderline Patients. *Journal of Personality Disorders*, *20*(1), 9–15.

Zanarini, M. C., Laudate, C. S., Frankenburg, F. R., Wedig, M. M., & Fitzmaurice, G. (2013). Reasons for Self-Mutilation Reported by Borderline Patients Over 16 Years of Prospective Follow-Up. *Journal of Personality Disorders*, *27*(6), 783–794.

Zanarini, M. C., Williams, A. A., Lewis, R. E., Bradford Reich, R., Vera, S. C., Marino, M. F., ... Frankenburg, F. R. (1997). Reported pathological childhood experiences associated with the development of borderline personality disorder. *American Journal of Psychiatry*, *154*(8), 1101–1106.

Zhang, J., Zheng, Y., & Qi, D. (2017). Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)* (pp. 1655–1661).

Zhou, C., Bhatnagar, N., Shekhar, S., & Terveen, L. (2007). Mining Personally Important Places from GPS Tracks. In *2007 IEEE 23rd International Conference on Data Engineering Workshop* (pp. 517–526).

Zignol, M., Cabibbe, A. M., Dean, A. S., Glaziou, P., Alikhanova, N., Ama, C., ... Raviglione, M. C. (2018). Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study. *The Lancet Infectious Diseases*, *18*(6), 675–683.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.