BUILDING TRAJECTORIES THROUGH CLINICAL DATA TO MODEL DISEASE PROGRESSION

A thesis submitted for the degree of
Doctor of Philosophy

By
Yuanxi Li

School of Information Systems, Computing and Mathematics
Brunel University
September 2013

Table of Contents

List of Figures

List of Tables

"If you build a model from a group of people who are kind of similar to the current patient, you might do better" Visweswaran 2012 [Sav12].

Abstract

Clinical trials are typically conducted over a population within a defined time period in order to illuminate certain characteristics of a health issue or disease process. These cross-sectional studies provide a snapshot of these disease processes over a large number of people but do not allow us to model the temporal nature of disease, which is essential for modeling detailed prognostic predictions. Longitudinal studies, on the other hand, are used to explore how these processes develop over time in a number of people but can be expensive and time-consuming, and many studies only cover a relatively small window within the disease process. This thesis describes the application of intelligent data analysis techniques for extracting information from time series generated by different diseases. The aim of this thesis is to identify intermediate stages in a disease process and sub-categories of the disease exhibiting subtly different symptoms. It explores the use of a bootstrap technique that fits trajectories through the data generating "pseudo time-series". It addresses issues including: how clinical variables interact as a disease progresses along the trajectories in the data; and how to automatically identify different disease states along these trajectories, as well as the transitions between them. The thesis documents how reliable time-series models can be created from large amounts of historical cross-sectional data and a novel relabling/latent variable approach has enabled the exploration of the temporal nature of disease progression. The proposed algorithms are tested extensively on simulated data and on three real clinical datasets. Finally, a study is carried out to explore whether we can "calibrate" pseudo time-series models with real longitudinal data in order to improve them. Plausible directions for future research are discussed at the end of the thesis.

Acknowledgements

Firstly I would like to thank my supervisor Dr Allan Tucker for constructive supervision throughout my PhD study. Without his support, encouragement and guidance, this thesis would have never been completed. I have learned a great deal under his direction not only academically but also philosophically in terms of dealing with challenges in both work and life. It has been a true privilege and a pleasure to work with him.

I am also very grateful to my second supervisor Dr Stephen Swift for all of his invaluable advice and feedback, which are instrumental in My PhD research and all of my publications.

I would also like to thank my colleagues (past and present) in the CIDA (Centre of Intelligent Data Analysis) at Brunel for their help, support, as well as interesting discussions, which made the lab fun and pleasurable to work in.

I own a big gratitude to my family for their love, encouragement and support throughout my PhD study.

Finally, I would like to express my sincere gratitude to everyone who has helped and encouraged me in many ways throughout the four years of my research life.

Publications

The following publications have resulted from the research presented in this thesis include:

1. X. Li, D. Garway-Heath and A. Tucker. (2009). Using pseudo time-series trajectories to explore disease regions in glaucoma, Fourteenth Workshop on Intelligent Data Analysis in bioMedicine and Pharmacology (IDAMAP 2009), IEEE.

2. Y. Li and A. Tucker (2010). Uncovering disease regions using pseudo time-series trajectories on clinical trial data, 3rd International Conference on BioMedical Engineering and Informatics (BMEI 2010)

3. Y. Li and S. Swift and A. Tucker (2013). Modelling and analysing the dynamics of disease progression from cross-sectional studies, Journal of Biomedical Informatics, Volume 46, Issue 2, Pages 266-274.

4. Y. Li and A. Tucker (2014). How much time do we need to learn disease progression, 27th International Symposium on Computer-based Medical Systems (CBMS 2014), IEEE.

Chapter 1

Introduction

1.1 Motivation

Clinical trials are typically conducted over a population within a defined time period in order to illuminate certain characteristics of a health issue or disease process. These cross-sectional studies give us a snapshot of such disease processes over a large number of people but do not allow us to model the temporal nature of disease. Longitudinal studies on the other hand, are used to explore how these processes develop over time in a number of people but can be expensive and time-consuming, and often only cover a relatively small window within the disease process. Thus, there is a need for effective computational techniques that can help people to gather information from those unknown datasets in order to enhance the knowledge of the underlying processes. In the biomedical domain, computational methods have become more and more important for the evaluation and analysis of experimentally generated data. There is already a large amount of work that explores data mining from cross-sectional biomedical data in order to make predictions or understand the relationships between

variables ([SSPGH06], [BZ08], [SS11]). Whilst some studies involve learning computational and statistical time-series models of progression from longitudinal data such as ([TVLGH05], [SL02], [HXW+10]), many datasets are cross-sectional and the time dimension is not measured, despite the inherently temporal nature of disease. This is due to the expensive nature of these studies across an entire population. Disease progression can take different forms with different trajectories starting from healthy condition and developing different symptoms, depending on the individual, before progressing to advanced stages of a disease. Figure 1.1 shows two examples of disease progression in breast cancer and glaucoma where possible trajectories (marked as arrows) denote progression from healthy (light blue circles) to diseased stages (red crosses). Data has been visualised using multi-dimensional scaling.

Expert clinicians generally believe that degenerative diseases are characterised by a continuing deterioration of organs or tissues over time. However, this monotonic increase in severity of symptoms is not always straightforward. The rate can vary in a single patient during the course of their disease so that sometimes rapid deterioration is observed while other times the symptoms of the sufferer may stabilise or even improve, when medication is used. Interventions such as medication or surgery can make a huge difference to quality of life and slow the process of disease progression, but they rarely change the long term prognosis. The characteristic of many degenerative diseases is therefore a general transition from healthy to early onset to advanced stages. Therefore, clinicians need an 'indicator' to identify diseases and their progression for more accurate medical prognosis. Medical prognosis (the prediction of the outcome of a disease) can be used to improve quality of life, slowing the tempo of disease progression

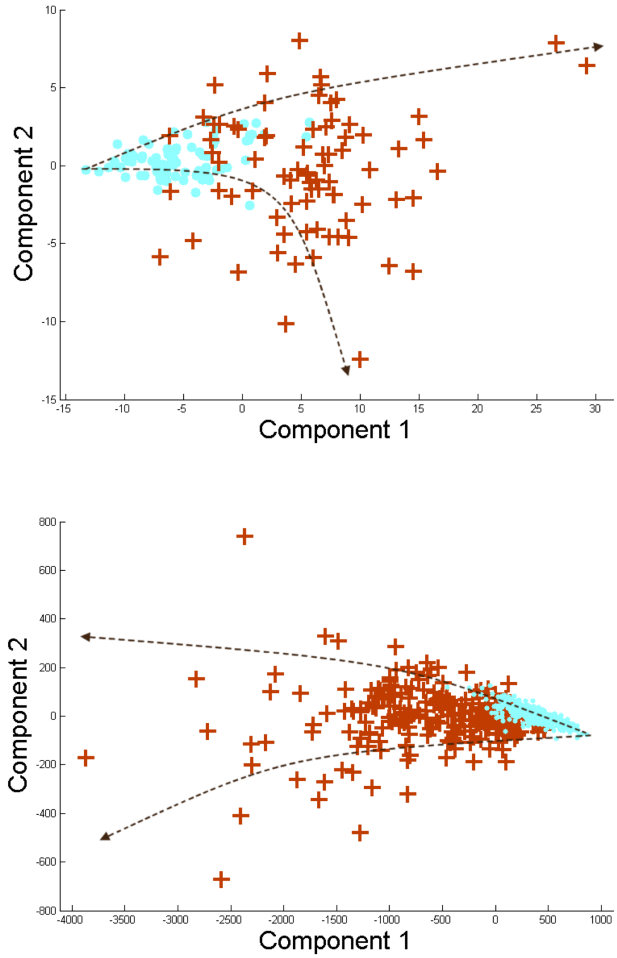

Figure 1.1: Trends in progression of degenerative disease. An example from glaucoma (top) and breast cancer (bottom).

as it can lead to early interventions. Furthermore, timely diagnosis can be extremely beneficial as it can reduce the potential risk for patients. Notably disease onset may take place before the first symptoms occur ([BSDP01], [BH05], [MMM+00]). Marcel et al. [vGTL08] depicts an interaction to represent how the current patient state may influence the next patient state (see as Figure 1.2). As Visweswaran said [Sav12] "if you build a model from a group of people who are kind of similar to the current patient,

you might do better".

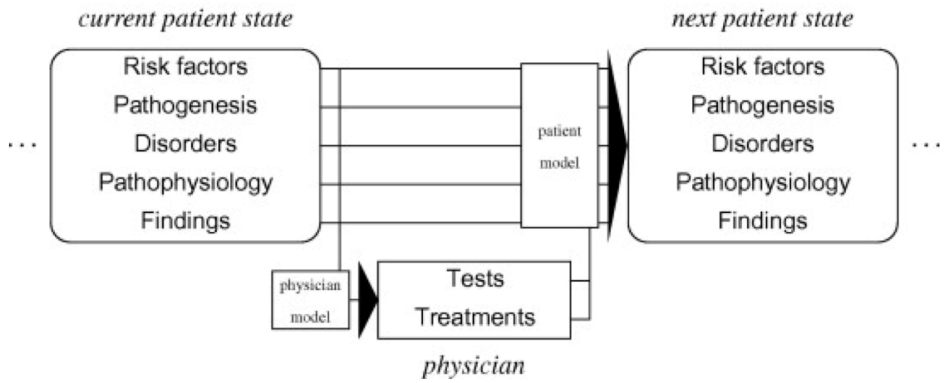

Figure 1.2: A clinical decision making in terms of a physician model and a patient model.

The aim of this thesis is to learn disease trajectories as well as to identify intermediate stages in the disease process by using a combination of bootstrapping, hidden Markov models (HMMs) and unsupervised machine learning. Thus, this thesis focusses on Machine learning [DRK04] methodologies for modelling disease progression, dealing with a number of issues inherent with this. It uses a combination of bootstrapping ([TGH10], [ET94]) and specialised hidden Markov models in order to model disease and identify important regions in the disease trajectories. The approach of relabeling trajectories has enabled the exploration of the temporal nature of disease progression, and has been tested on both real and simulated biomedical data.

1.2 Research contributions

Main contributions of this research include:

- A full formal definition of a pseudo-time-series is derived, along with the associated temporal bootstrap.

- Identified of key regions in disease processes learnt from cross-sectional data. An extension of the temporal bootstrap is explored, which can be used to identify intermediate stages in a disease process and sub-categories of known diseases with subtly different symptoms.

- Validation of the Pseudo-Times-Series model on a number of real-world biomedical datasets. The reliable time-series models have been created from large amounts of historical cross-sectional data from three very different diseases: Glaucoma, Cancer and Parkinson's disease using the temporal bootstrap technique and tested on real and simulated data.

- Exploration of the characteristics of the proposed algorithms, which is to demonstrate the effectiveness of the proposed method, the performance of the approach is explored on a number of simulated time-series generated from autoregressive hidden Markov models (HMMs).

- The new approach to analyse, visualise and model checking, which is compared with a clustering method for identifying key stages in disease progression based on the ability to explain the underlying dynamics.

- Integration of cross-sectional and time-series data within the Pseudo-Times-Series framework, which is includes exploring what degree pseudo time series model can be improved and finding out how many or how few time-series data is sufficient when cross-sectional data is abundant.

1.3 Organization of this thesis

Apart from this chapter, the remainder of this thesis is organised as follows:

- Chapter 2 presents the background knowledge and its role in disease progression modelling for the research conducted for this thesis. It reviews the state of the art methods, techniques and their performance. Gaps in the research area are also identified, some of which will form the basis for the objectives of this study.

- Chapter 3 focuses on a variety of key concepts including hidden Markov models, the temporal bootstrap developed using different methods and algorithms relevant to this research. The process of application of the new relabelling technique for extracting key stages in disease from both cross-sectional and longitudinal studies to build reliable models of disease progression, is demonstrated and explained. Finally, the real-world clinical datasets are introduced.

- In Chapter 4, results produced by using the proposed relabelling algorithm on the three different clinical datasets are presented and analysed. These results are used for exploring disease trajectories as well as simulated data. A comparison of the trajectories within a medical context is discussed.

- Chapter 5 explores the calibration of models learnt from cross-sectional data, the results using small samples of longitudinal data are presented.

- Chapter 6 summarises all key achievements and contributions towards the identified aim and objectives of this research project. Research limitations for further research and development in this area of research are discussed.

Chapter 2

Literature Review

This chapter reviews previous and current studies concerning the application of machine learning in medicine. Firstly, the nature of data associated with clinical trials is discussed with a focus on cross-sectional and longitudinal studies. This is to gain an insight into the advantages and disadvantages of these commonly used approaches when building predictive models. Next, different state-of-the-art machine learning techniques are introduced within a clinical data perspective discussing examples of their use and how the different techniques perform at predicting and understanding disease outcome. This review covers the subjects of machine learning, time-series modelling, classification and clustering using different techniques common in biomedical data analysis. Through this survey, gaps in this research area are identified, in particular concerning the weaknesses of cross-sectional and longitudinal data, on which the objectives for this thesis are centred.

2.1 Clinical Trials

Clinical trials are sets of tests taken from a wide range of medical areas that are typically conducted over a population, within a defined time period, in order to illuminate certain characteristics of a health issue or disease process. They can vary in size considerably. Clinical trials can be effectively used for identifying responses to interventions such as patients' response to drugs, the side effects of drugs and their concurrent diseases [SMB+03]. They can also be used to explore the variation in disease symptoms by applying clinical metrics to a large population exhibiting varying symptoms from the healthy through to the advanced disease stages, such as the use of multiple endpoints in interim analyses for disease treatments [SNA87].

A typical example of clinical data that is commonly used in large-scale clinical trials is the data obtained by Heidelberg Retinal Tomography (HRT) [GHFH00] which consists of measurements associated with the three dimensional shapes of the optic nerve head. Another test which will be used later in this thesis measures vocal impairment as early indicators for the onset of the Parkinson's disease (PD). A number of voice measurements have drawn significant attention for detecting and tracking the progression of symptoms of PD [LMR+07]. The final dataset that is used in this study concerns cancer [RSMJ85], [CYYK05], [PWM+99]. These studies are often based upon tumour examinations where characteristics of tumours are measured to explore the effect of a medical treatment.

Whilst many clinical trials are cross-sectional where only one measurement is made per individual, some involve longitudinal studies where a (relatively smaller) number of people are followed over a period of time. Albert [Alb99] discussed five different

clinical trials where the primary outcome is observed over time, allowing longitudinal data analysis. This study highlights that sequential monitoring is an important issue in clinical trials identifying sequence of events in individuals directly.

It is also worth noting that clinical trial data is often used in combination with historical data in order to improve the statistical validity of the conclusions. For example, randomised clinical trials with 6-12 months of clinical follow-up were used in [MLP+04] comparing different clinical interventions (here the insertion of a small tube - stent - into arteries for introducing drugs in a controlled manner). The study involved a large scale meta-analysis of 11 trials in order to quantify more accurately their effect on blood vessel narrowing (restenosis).

2.2 Cross-sectional and Longitudinal Studies

Cross-sectional studies are methods of clinical observation that only record information (such as clinical test results and demographics) across a sample of the population from subjects without modifying the environment. They do not consider past or future behaviour [Wor09] (a particular disease process but without any measurement of progression of the process over time), and all measurements on each subject are made at a single point in time [Man03]. For example, one may choose to measure insulin levels in diabetes patients across two age groups, over 40 and under 40, and compare them to insulin levels among controls. However, they do not normally consider the past or future insulin levels of diabetes patients. Cross-sectional studies can also be used to create subgroups for family diseases, for example gender or body mass index (BMI). A main advantage of cross-sectional studies is that they can compare many different

variables at the same time with fewer resources. They typically include people who are healthy as well as people at all stages of a particular disease process (if the sample is large enough) in the analysis. They are also relatively cheap and simple to perform [Man03], [Wor09], [JLTL81] in comparison to longitudinal studies. Cross-sectional studies give us a 'snapshot' of disease processes over a large number of people but do not allow us to model the temporal nature of disease, because the time dimension is not captured. For example, subgroups may exist that depend on temporal behaviour. Mann [Man03] indicates that "it is better to study a cross-sectional sample of patients who already have the disease" by employing longitudinal studies. This does however mean that many patients who are displaying early signs of disease onset may be missed.

Longitudinal studies [Dig02] are another observational approach that is used to explore how disease processes develop over time in a number of people. In clinical trials, longitudinal data are collected for three reasons [Alb99]: first, to obtain a more precise estimate of the outcome and hence the treatment effect; Second, to monitor clinical variables at a particular time; third, to evaluate the effect of treatment over time. Clinical test results are recorded, often without manipulating the study environment. Although the trial has the potential to provide definite information about "cause-end-effect" relationships, it may require monitoring the same subjects, over a long period of time [Wor09]. The results of multiple tests are recorded, generating Multivariate Time-Series data. This is common for patients who have high risk indicators of disease where they are monitored regularly prior to diagnosis. Cross-sectional studies cannot serve this purpose. For example, it may be chosen to look at the change in insulin levels among men over 40 who have been smoking for a period of more than

20 years. The longitudinal study design would account for insulin levels at the onset of a smoking regime and as the smoking behaviour continued over time. The advantages of a longitudinal study are that researchers can distinguish the individual level changes in the characteristics of the target subject [Dig02], and capture the temporal details of the disease progression beyond a single moment in time [Wor09]. However, the data is often limited in terms of the cohort size, due to the expensive nature of the studies. Albert [Alb99] proposed that "extensions of multiple endpoint methodology to the analysis of longitudinal data is another interesting area". For example, where similar early symptoms may end up following one of a number of potential disease trajectories. Such studies can be expensive and time-consuming, and many only cover a relatively small window within the disease process [YSA13].

Regardless of the advantages and disadvantages of cross-sectional and longitudinal studies, it seems that both have weaknesses when used on their own. Hence, there is an on-going effort to combine and use both approaches in conjunction ([LMH+88], [AZP06], [RPK+03], [AFS+04]). Cross-sectional studies are the best way to determine prevalence and variation in a wide population but do not provide an explanation for the findings [WDL+90], while longitudinal studies could provide "cause-end-effect" relationships but only under limited samples.

2.3 Machine Learning in Medicine

Machine learning (ML) [MBK98], [Mit97] belongs to the field of artificial intelligence [LCS+06], [Kon01]. It is used to extract useful and meaningful information from complex data. The aim of ML is to provide computational methods to learn from

data. This can be anything from the identification of patterns of similar behaviour to complex mappings between symptoms and diagnoses.

There are two major types of machine learning: supervised and unsupervised learning. Supervised learning is used to classify datasets by learning a mapping between the data (usually the results of some clinical test) and some predefined class label (representing some disease outcome) determined by an expert in the field. It involves learning a function from a set of training data and using this to predict the class for new input data (the test set). In contrast, unsupervised learning involves learning from data that has no class information by identifying similar data regions (or clusters). For example, it can be used to identify patients with similar symptoms.

Computational methods are becoming more and more important for the evaluation and analysis of experimental data, assisting us to extract useful information in as an automated as possible manner, especially when there are 98 large amounts of data. M-L has been widely applied in many areas: a natural language system was successfully developed for various tasks involving text processing and an automated script for data collection [HDA01]; Larranaga et al. [LCS$^+$06] used ML techniques to deal with gene identification problems. ML is also used to make predictions for experiments, such as Chen and Xu[CX04] applied ML for protein dispensability prediction. Furthermore, ML is often used to find the relationships between observed variables within datasets in a medical context, improving the efficiency and quality of medical decision making [KKG$^+$99] [MP01a]. A major research effort in this field is to automatically classify disease [Org] and predict future outcomes for patients [Sav12]. Clinicians may use

ML methods to generate new hypotheses to enhance their basic diagnostic and prognostic processes [MP01a], [BGP99], [Jan99], [RRL99]. Based on the number of ML publications and the increasing use of ML techniques in medicine, it appears that ML will play an ever more important role in clinical medicine. The next section documents some examples of successful machine learning applications in medicine.

2.3.1 Analysis of Time-series Focusing on Clinical Data

Time-series can be simply defined as a sequence of data points that is characterized by its continuous nature. This kind of data is widely used in various domains, such as econometrics, finance, earthquake prediction, weather forecasting, as well as biomedicine (in longitudinal data analysis). In general, time-series analysis is applied to experimental data measured over a period of time, in order to extract meaningful information from uncertain data, and to predict future values of variables ([LAR03], [Cha96]). Sometimes, data is also available on several related variables of interest, which can be defined as Multivariate Time-series (MTS) (more details in Section 3.1). MTS can be used to study the dynamic relationships between variables over time in order to predict the progression of degenerative diseases, such as Parkinson's disease, glaucoma or cancer. MTS often give better predictions than univariate time-series models. Specialists in medicine are interested in how a disease progresses across the time-series. Clearly these trends will depend upon a number of factors such as which clinical variables are selected, how much data there is available in the sample, and whether the disease process is generally monotonic. According to [Fu11]: the way to represent the time series data is the fundamental problem in the context of time series data mining. In order to infer the process of disease progression, an appropriate use of

a modelling approach is needed.

A time-series model can be built when time stamps (discrete or continuous) are available. These models can be used to try and predict future values of the data or the disease outcome. Tiao et al. [CMJ98] suggest two reasons for analyzing and modelling time-series data: first, to understand the dynamic relationships among variables (one series may lead the others), and second, to improve accuracy of forecasts. Linden et al. [LAR03] listed three basic steps for the development of a time-series model: (1) graphing the data; (2) choosing the appropriate model and fitting the data; (3) evaluating the model. There are a number of popular approaches to modelling time-series. For example, the Box-Jenkins method [BJR13] is known as the Autoregressive Integrated Moving Average (ARIMA) as it combines the autoregressive and moving average processes to model past observations and errors. It handles trends and cycles through differencing the data. The approach is attractive due to its ability to capture a diverse set of time-series behaviour. However, as a result of its flexibility it can risk overfitting data. Another very common time-series modelling approach is the hidden Markov model (HMM) [Rab89] (See section 3.4.1 for more details). It is a popular model for modelling sequential and time-series data which can deal with uncertainty and noise. A correct model is essential to predict future values based on previous observations and this can include forecasting, regression analysis and classification. Regression techniques aim to fit a model (linear or polynomial) through the data where time is used to predict the rate of change in some set of clinical variables [VFH97]. Linear regression attempts to find a straight line that best 'fits' the data. See Figure 2.1 for an example.

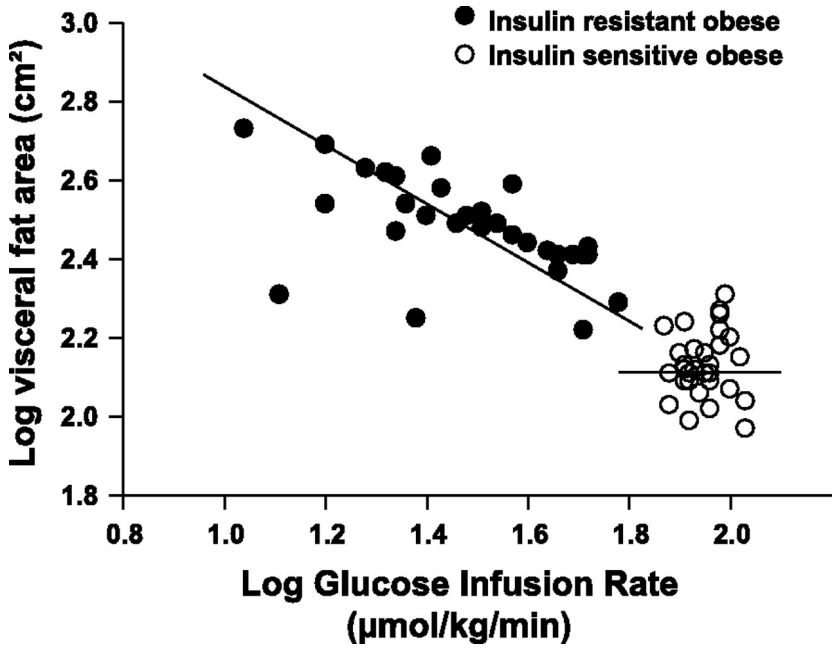

Figure 2.1: The example of clinical linear regression. Adapted from [KFD+10].

As well as classification methods can be regarded as types of data mining approaches to modelling time-series data in medicine (see section 2.3.2), however, care must be taken as observations are not independent of one another (symptoms at time *t* are typically dependent on the symptoms at $t-1$). Previous observations play an important role in time-series analysis, since they can be used as basis for future behaviour prediction [LAR03].

There has been some preliminary work conducted in an effort to create reliable time-series models from large amounts of historical cross-sectional data. For example, Broman et al. [BQW+08] attempts to estimate the rate of glaucoma progression from cross-sectional studies by using the mean damage in the eye and the mean age of disease onset. In [TGH10], a combination of distance metrics, graph theoretical operations and resampling is used to build trajectories through the dataspace. Bellazzi et al.

[BLM+99] combined time-series analysis and temporal abstraction techniques to the analysis of time-series data of Blood Glucose Levels, from insulin dependent diabetes patients. Diggle[Dig02] explored three certain longitudinal data sets from the biomedical domain, which present the challenges for analysis: First, in HIV, an immune cell called CD4+ is attacked by the virus, and its numbers decrease with time following infection, thus it can be used to monitor disease progression; Second, a three-period crossover trial of primary dysmenorrhoea produced data was used to examine the effectiveness of pain killers; the last, a dataset from a clinical trial of drug therapies for schizophrenia. Other studies involve learning computational and statistical time-series models of progression from longitudinal data ([TVLGH05], [SL02], [HXW+10]).

In conclusion, time-series analysis methods are important tools for analysis of medical data to anticipate disease [RM03]. Time-series analysis lies in the heart of the work presented in this thesis and we will explore probabilistic models like hidden Markov models (HMMs) in more detail in Chapter 3.

2.3.2 Intelligent Data Analysis in Medicine: Classification Methods

A set of data with known structure can be divided into classes. For example, if information about some labelling of the data is available then supervised classification algorithms can induce the rules from the data [LCS+06] in order to predict new cases. More formally, classification tasks involve learning a mapping from a vector of measurements to a categorical variable. The variable to be predicted (the class variable), takes values in the set $C = \{c_1, ..., c_m\}$. The observed variables $d_1, ..., d_i$ are referred to as the features. Classification takes are becoming increasingly popular due

to dramatic increases in the availability of large collections of medical data [PSS+09], [ZL10], [VPR+07]. Any improvements in prediction will result in better medical decision making which, can be very valuable for diagnosing future patients [VAH+10], [RKM09]. For example ML classification was explored in the study of breast cancer [BHK]. Here, Bontempi and Haibe-Kains used classification techniques to divide breast cancer patients into groups for different clinical therapy. Classification was based on tumours with similar histopathological appearance. The study examined various clinical sources and responses to therapy. The results revealed that biologists often failed to accurately classify breast cancer due to tumour metastasis, thus highlighting the ability of ML methods to assist clinical experts in making diagnoses. One other popular biomedical application of ML classification concerns the use gene expression data to classify disease. Many databases containing gene expression information for patients can be used to classify cohorts with different diseases from control groups. For example, in a study of tumours [vVDVDV+02], which applied classification techniques to identify a gene expression signature. This is known as feature selection where a small number of variables are identified as most informative in a classification task.

There are a number of approaches and methodologies used in classification in the clinical data analysis area. The main ones are discussed in the following section (based on a combination of predictive performance and their ability to explain the underlying relationships in a dataset): *Bayesian Networks (BNs), Naïve Bayes (NBs), Neural Networks (NNs) and Decision Tree (DTree).*

2.3.2.1 Bayesian Networks in Medicine

Bayesian Networks (BNs) (See Figure 2.2) are probabilistic graph-based models that represent the probabilistic interdependencies between variables. They are simple, commonly used and less prone to overfitting the data as they are biased towards simple networks (the data source can be stationary and not vary with time). BNs typically consist of two components: a directed acyclic graph (DAG) with nodes representing any variance of variables that exists in the real world, and a set of conditional interdependencies associated with each node (see Figure 3.2).

The networks can be used to represent the network qualitatively and quantitatively using a graphical structure involving nodes that have an associated conditional probability distribution. BNs can also be used to answer probabilistic queries about networks. BNs have become more and more popular, being used for the computational modelling of knowledge in many areas such as bioinformatics, medicine, and decision-making systems. They are particularly powerful in transparently modelling the relationship between variables and capturing the uncertainty in knowledge and data [LCS+06].

One of the key advantages of Bayesian networks is their ability to integrate data with human expertise. This can be achieved using the notion of an *informative prior* [CS00], where a model is constructed and then updated when more data becomes available using Bayesian updating techniques, resulting in a *posterior* model.

Microarray data analysis methods often fail to correctly deal with uncertainty, thus BNs have advantages for analyzing gene microarray data, especially for un-normalised cDNA microarray data [Wil07]. BN classifiers can also be used to analyze differential

gene expression, to provide useful information about biological pathways [Slo02].

BNs not only work on very different types of data but they have also proven very popular for modelling combinations of different types of data in a single model (such as gene expression and clinical data). For example, Gevaert et al. [GDST+06] learnt BNs on breast cancer patients data treating clinical and microarray data on an equal footing. Visweswaran et al. showed that patient-specific models could be improved through the use of a Bayesian model. Elsewhere, BNs were applied to the diagnosis and prognosis of first cerebral paroxysm [ZLNRP99]. In an effort to avoid a "wrong blood in tube" error ('a specimen of blood collected on Patient A, but for which the accompanying requisition and label is for Patient B'), which can kill people more often than another accidents or diseases, a BN was used to predict the mismatches of Glucose and HbA1c in two experiments [DS10].

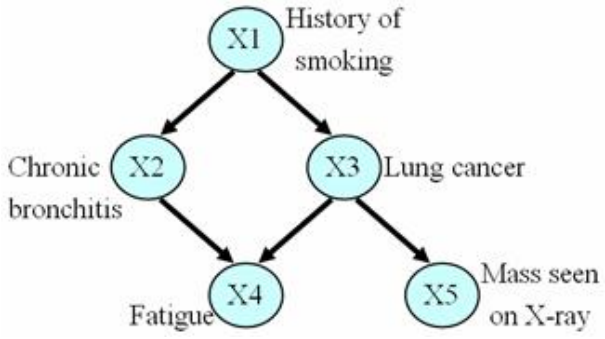

Figure 2.2: Example architecture of a Bayesian Network. The diagram adopted from [Coo99], which is hypothetically about the medical domain with 5 variables.

Dynamic Bayesian Networks (DBNs) (See Figure 2.3 [PdKJ+10]) can be defined as a special case of Bayesian Networks that can model sequential data or time-series [FGW99]. They are known as time-series analysis, which consist of a set of observations from the data source (a sequence of variables) dynamically changing over time.

Unlike BNs, DBNs can provide direct mechanisms for representing temporal dependencies. Therefore, DBNs enable users to update the system and also predict further events. More recently, they have become popular for modelling disease [TVLGH05] and are useful as both prognostic and diagnostic tools. This is due to the fact that they explicitly model temporal (dynamic) and non-temporal relationships among different variables in the real world and are flexible enough to model latent variables.

Several applications of DBNs have been proposed in medicine. For example, in monitoring the treatment of renal failure patients [CCC05], BNs are implemented to represent relationships between Hydration and Dialysis Sessions on Dry Weight. DBNs are used to model dynamic processes of the treatment in order to explore whether past events have an effect on the present state of the patient. Watt et al. [WB08] used DBNs to help predict early presence of Osteoarthritis (a knee disease that can cause knee pain, disability and decreased bone mass) and analyse the progression of the condition over time. Van Gerven et al. [vGTL08] used three individual patientś case to construct DBNs for prognosis. They demonstrated that DBNs not only serve for modelling disease progression, but can also identify the effect of treatment and the development of complications. In [CGHCT12] a form of DBN, that clusters sections of time series whilst simultaneously learning DBN structure and parameters, was used to model glaucoma progression. Furthermore, DBNs are also used to model Neuronal Interactivity for brain activation patterns [ZSAK$^+$05], and to optimise treatment in intensive care unit (ICU) [CVDGV$^+$09].

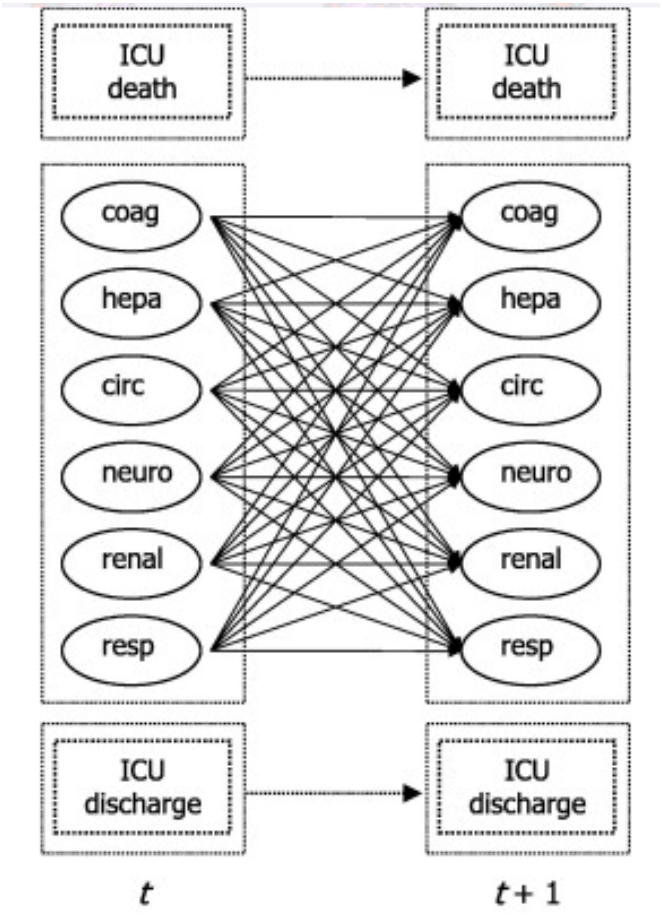

Figure 2.3: A Dynamic Bayesian Networks (Predicting Attacks). Possible transitions between variables at the same time-slice. The Figure shows an example of DBN with number of variables over time lags where each node represents a variable at a certain time slice and each link represents a conditional dependency between organ systems. Adapted from [PdKJ+10].

It is a non-trivial problem to create DBNs from data. Table 2.1 lists four different stratums of creating DBNs. 'Complete data' means all variables are known, while 'incomplete data' means some variables cannot be measured in some situations, referred to as missing variables. 'Unknown structure and full observability' refers to finding a way to learn the structure of DBNs from observable data. In this thesis, we focus on the 'unknown structure and partial observability' because some variables cannot be observed in the real world and it is hard to distinguish the structure when learning DBNs from real data. The EM algorithm (see 3.4.1 for more details) is a powerful method to tackle this problem, and it can improve the likelihood of the data given the model.

Structure/Observability	**Method**
Known/Full (complete data)	Simple statistics
Known/Partial (incomplete data)	EM or gradient ascent
Unknown/Full	Search through model space
Unknown/Partial	Structural EM

Table 2.1: Methods for creating DBNs structure and determining their parameters. Adapted from [MP01b].

2.3.2.2 The Naïve Bayes in Medicine

Naïve Bayes (NBs) (See Figure 2.4) is a simple probabilistic classifier using Bayes theorem with strong naïve independence assumptions [Zha04], where no hidden attributes influence the prediction process [GP95]. This classifier requires a small amount of training data to estimate the parameters necessary for classification, and is fast to train. It is also called an optimum classifier, because this classifier can minimize the cost of total misclassification. It has been successfully used in many applications, such as the

Rainbow program which employs a NBs classifier to perform statistical text classification [A.K96]. Yousef et al. [YJK+07] used the Rainbow program to train the NBs classifier for microRNA target gene prediction. Palaniappan et al. [PA08] used NBs along with another two predictive models, in order to discover the hidden patterns and relationships between different medical profiles and to develop an intelligent prediction system for heart disease. Although the NBs classifier is generally an effective and versatile classification approach, sometimes false predictions may occur because NBs can not gives appropriate descriptions for the relationships between the variables and the outcomes. The positional independence assumption of the NBs makes the computation of the joint probability value easier at the expense of the accuracy or the underlying reality (it is the strength as well as the weakness [PLV02]). NBs are the simplest form of Bayesian Network, too simple to model complex domains, where all attributes are independent. Therefore, only the variances of the variables for each class need to be determined.

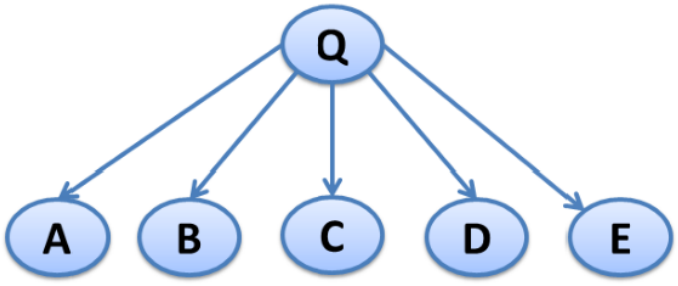

Figure 2.4: Example architecture of The Naïve Bayes. In the diagram, the Q represents a parent of the A, B, C, D and E. It is assumed that all the variables are independent for a given class Q. Given the values of A, B, C, D and E we can estimate the probability of the class, Q using the Bayes rule.

2.3.2.3 Neural Networks in Medicine

Neural Networks (NNs) (see Figure 2.5 [Hel13]) constitute a set of machine learning methodologies inspired by the structure of neurons in the brain [ZB08]. They are good at dealing with the complexity of experimentally generated data, which consists of nodes (neurons) that receive, process and transmit signals. The perceptron is the simplest neural network structure. It has the ability to learn from examples, that is a distinctive aspect of NNs over other classifiers. It contains two layers: the first layer is input layer and the second layer is the output layer. Both layers can have many nodes and units. NNs have the ability to give straightforward theoretical predictions on DNA sequence level, protein sequence level and protein structure level [BA06]. Odewahn et al. [OSP+92] pointed out that NNs are good at managing problems with a large amount of parameters, and at classifying objects distributed in complex high-dimensional space. For example, Nagl [Nag01] successfully applied NNs in a protein study, which were used to analyse the emergence of drug resistance in HIV-1 (human immunodeficiency virus 1). In another case, Lisboa and Taktak [LT06] explored the benefits of NNs as decision making tools (diagnosis and prognosis). Using 396 clinical trials (cancer), most of the studies showed an increase in benefit to healthcare provision. Furthermore, NNs are used to estimate medical outcomes and resource utilization in intensive care unit environments (ICUs)[FEST01].

NNs also have their limitations. For example, Satish and Gururaj [SG93] demonstrated the limitations of the NNs with a single layer NNs. The NNs could only classify linearly separable signals, and proved inadequate to achieve the required tasks.

Multilayer Neural Networks (MNNs), also known as the Multilayer Perceptron, have layers between the input and output layers. A MNN is a fully connected network where all nodes from one layer are connected to the next layer and can be extended to any number of hidden and output layers. This type of NN can learn complex functions. The MNN deals with non-linearly classifiable data by employing hidden layers (neurons are not directly connected to the output) and essentially linking numerous perceptrons (which are the individual neural network units) together. The big advantage of MNNs lies in the fact that all nodes from one layer are fully connected to the next layer and can be extended to any number of hidden and output layers. Back propagation is the main algorithm utilised in multilayer neural networks. The main drawback to back propagation is that it is susceptible to overfitting the training data at the cost of decreasing generalization accuracy over other new data [Mit97]. In the medical domain, some authors ([JAGRRJ$^+$03], [FWIB95], [Kiy11] have modelled systems for prognosis in breast cancer patients using MNNs. Chen et al. [CLKW97] implemented MNNs for the purpose of designing the dynamics of the mean arterial blood pressure system, in order to meet specified clinical constraints. Yan et al. [YJZ$^+$06] developed a decision support system to support the diagnosis of heart diseases based on MNNs. In order to achieve high diagnosis accuracy they used five different kinds of heart diseases from 352 clinical records to train and test the system. Li et al. [LLCJ00] demonstrated MNNs are appropriate for non-linear medical decision support system in traumatic brain injury (cause of death and disability that include falls or vehicle accidents). MNNs have also been used to classify low back pain [VCT$^+$00].

Whilst, artificial neural networks (See Figure 2.5 for an example) have good

predictive performance, they are quite slow in both the training phase and application phases [BZ08], compared to simpler modelling methods such as BNs. They are also very prone to overfitting, where a classifier has focussed on a small area of data that may be irrelevant to the classification task. This is often due to too many parameters.

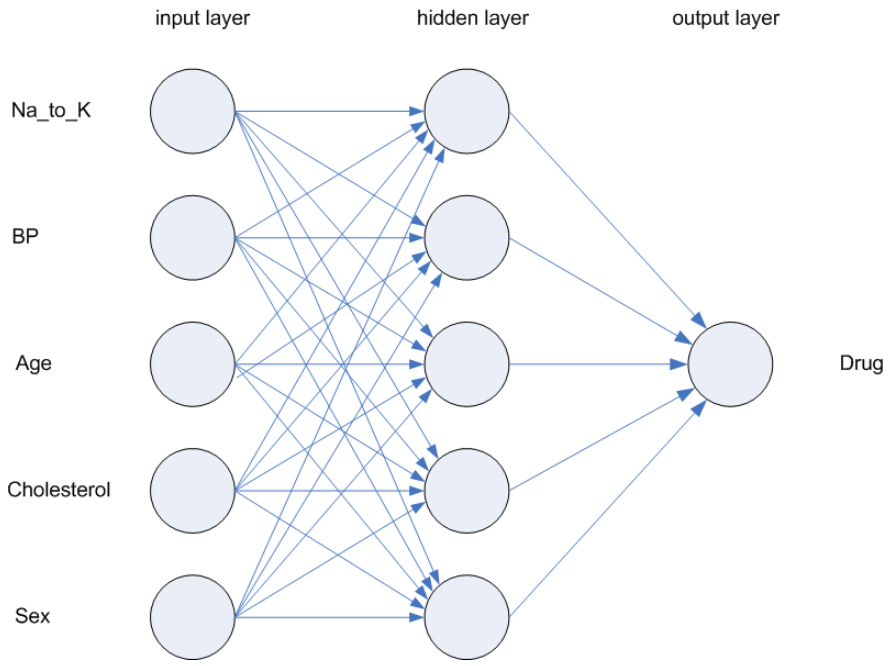

Figure 2.5: Example architecture of a Simple Neural Network. In the diagram, the inputs are separately transformed into a 3-dimensional vector hidden layer, which is finally transformed into the Drug (forward propagation). The output 'Drug' depends upon the random variables of the vector hidden layer, which depends upon the random variable inputs (back propagation). These two stages are independent of each other. Adapted from [Hel13].

2.3.2.4 Decision Tree in Medicine

A Decision Tree (see Figure 2.6 [WLH+09]) is a well documented machine learning method for classification, which is very easy to understand and widely used in classification problems, such as disease classification[Mit97]. In the traditional decision tree, the nodes represent decisions, arcs represent possible answers, and terminal nodes represent classification. The root node is located at the top of the tree and the tree is traversed starting at the root node. At each decision node, the different links represent the possible answers. This process is repeated from a root node until a terminal node (leaf node) is reached, where a class is allocated ([LKZ00], [BB01]). Links must be mutually distinct and exhaustive. In other words one and only one link will be followed [LCS+06]. Thus, all the decision nodes are working together, following along the path of the decision tree from root node to the leaf nodes [BZ08]. However when the training set is small, the terminal node of a decision tree is defined by chance if no one class can be identified clearly [Ber03].

The *Alternating Decision Tree (ADTree)* is a special case of a decision tree. It is a majority-weighted vote over very simple prediction rules, which deploys a type of machine learning method based on boosting for classification [FM99]. The ADTree is smaller and easier to interpret than other classifiers. The structure of an ADTree is similar to the standard decision tree, but can sometimes achieve better performance. One special feature of ADTree classifiers is that they give the classification margin as a measure of confidence. For example, Takada et al. [TSN+12] used an ADTree as a prediction model for predicting axillary lymph node metastasis in breast cancer patients (the node is examined for diagnosis of breast cancer). Using various clinical

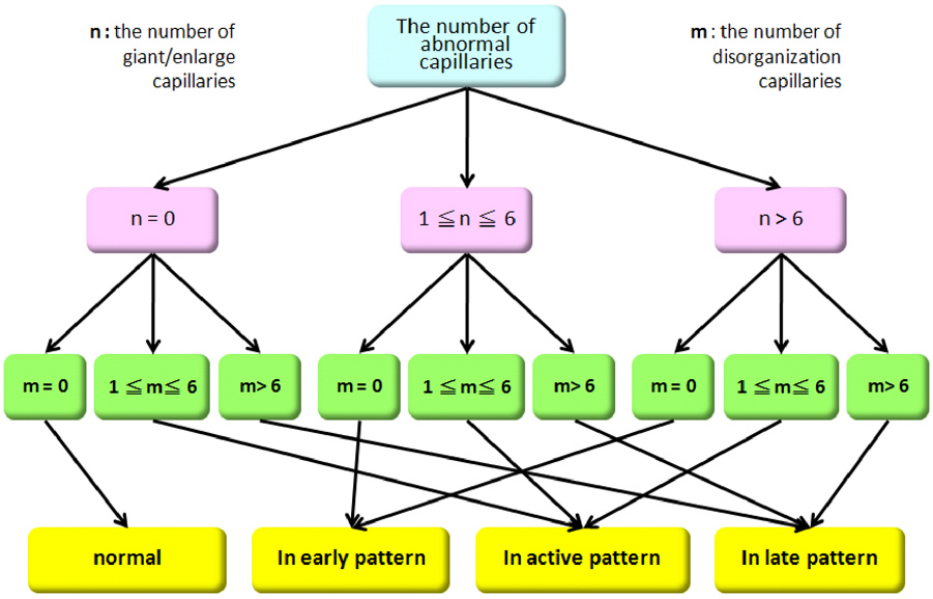

Figure 2.6: Example architecture of a Decision Tree, which is being used for determining disease progression based on the number of capillaries classified. From the diagram, it can be seen that there might be 4 types of symptoms that have been classified. Adapted from [WLH+09].

variables, they were able to achieve 95% accuracy on all the resulting values. The ADTree based approach has also been implemented for prediction of dengue fever (an infectious tropical disease that is transmitted by mosquito) with high accuracy [Kum13]. One special feature of ADTree classifiers is that they give the classification margin as a measure of confidence and each partition can be split multiple times. "The alternating tree maps each instance to a real valued prediction with the sum of the predictions of the base rules in its set. The classification of an instance is the sign of the prediction" [FM99]. ADTree has been used to predict survival of heart failure patients, by classifying them into groups that are expected to survive over a certain period of time or not [JKD13]. Experiments on a collection of datasets taken from the UCI machine-learning repository [NHBM], showed that the ADTree method is

efficient with small data sets, and is much easier to interpret than a standard decision tree with the same number of decision nodes.

Although decision trees are effective at solving classification problems, they also have limitations given that the typical top-down partitioning through a greedy split evaluation may result in quality loss [BBdC$^+$09] and may be sub-optimal. Classifiers such as decision trees, NBs and MNNs clearly have a role in medical informatics and in particular modelling disease when a diagnosis has been identified within a dataset. We will explore how the 'class data' (whether data has been labelled as healthy or diseased) can be used to build trajectories in the chapter 3.

2.3.3 Application of Machine Learning Methods to Medical Data Analysis: Clustering Methods

Clustering is a common technique for data analysis. It is an unsupervised machine learning method which deals with finding meaningful groups (the members are similar in some way) in a collection of unlabeled data. More formally, clustering can be defined as by given a representation of N objects $D = d_1, ..., d_n$, find k clusters $C_1, ..., C_k$ based on a measure of similarity. Each data point d_i is assigned to a unique cluster C_k. It has been extensively used in a variety of fields, including: marketing (finding groups of customers [HTO07]), biology (grouping gene expression data [CTC$^+$05]), medicine (evaluation of data from clinical trial [FSN$^+$00]) etc.. An important component of a clustering algorithm is the distance measure or similarity measure between data points. Data points that are close to each other according to a given metric or that have similar descriptive concepts belong to the same cluster. The efficiency of clustering algorithms and their application in a wide range of scientific fields is an area of active research.

The four most popular clustering algorithms for high dimensional data used in the biomedical domain are:

- K-means which is used to cluster the observations into k clusters, based upon distance of observations to k centroids (the allocation of data points and the centroids are iteratively updated). In gene expression data analysis [Slo02]. Bushel et al. [BWG07] explored an extension of the *k*-means algorithm, called mode *k*-prototypes for clustering heart disease samples.

- Fuzzy C-means clustering, where observations can belong to more than one cluster (centroid) with different degrees. It has been implemented for classification of oral cancer cell data [WGO03]). A modified fuzzy c-means algorithm has been utilised for bias field estimation and segmentation of magnetic resonance imaging data [AYM$^+$02].

- Hierarchical clustering which has two approaches: the first is 'bottom up', which starts at each observation, merging together data points or clusters, and then moves up the hierarchy; the other is 'top down', starting with one large cluster, then moving down the hierarchy, splitting clusters until all data is in its own cluster. Veer et al. [vVDVDV$^+$02] implemented hierarchical clustering for the purpose of clustering tumours on the basis of their similarities, to predict clinical outcome of breast cancer.

- Gaussian Mixture clustering, which uses a mixture model to represent the probability distribution of observations, and therefore cluster the data. For example, Gaussian Mixture Model-based segmentation has been applied in lung tumor

clinical studies [APMP07].

Despite the fact that the extensive research has been conducted to extend and improve clustering algorithms, they still have a number of limitations. Karypis et al. indicated that many advanced algorithms do not follow a preconceived model because they cannot faultlessly deal with highly variable clusters [KHK99]. For example, the K-means algorithm does not perform well with high dimensional data, when clusters in the data have different sizes, shapes and densities such as commonly in many clinical datasets. Hierarchial clustering also has limitations due to problems with defining a distance metric in high dimensional data [ESK02].

Clustering is clearly relevant to identifying regions of interest in a disease process. Therefore chapter 3 of this thesis will examine a method that involves using the Expectation Maximisation (EM) algorithm.

2.4 Summary

This chapter reviewed previous and current research work in the field of machine learning for biomedical data analysis. It discussed subjects including clinical trials, cross-sectional and longitudinal studies, machine learning, classification methods, clustering methods and time-series modelling. Four major models of classification and clustering were examined in a biomedical context.

The major conclusions arising from this review are:

(1) Cross-sectional studies do not allow us to model the temporal nature of disease and the time dimension is not captured as observations are taken at only one fixed point in time;

(2) Longitudinal studies are expensive due to their nature as individuals must be followed over time. In addition many studies only cover a relatively small window within the disease process, often missing the vital early stages.

(3) Building sequences through cross-sectional data is relatively unexplored, though Broman et al. [BQW$^+$08] have recently investigated estimating rates of progression in glaucoma from cross-sectional studies and Tucker [TGH10] has explored building trajectories through cross-sectional data - which is to be further explored in Chapter 4.

(4) Multiple endpoints were highlighted as an important issue (Albert [Alb99]) which again are relatively unexplored and will be further investigated in Chapter 4.

(5) Clustering data can help in understanding subcategories of disease such as different subpopulation of cancer sufferers as well as identifying important stages in a disease process (where data points in a series are clustered - for example into early, mid and late stages).

Based on the points above, the research in this thesis focuses on the use of sequence-building through cross-sectional data (including trajectories with multiple endpoints) by formalising and extending the pseudo time-series introduced by Tucker [TGH10]. It deals with clustering trajectories to identify important stages in a disease; and exploring how cross-sectional and longitudinal data can both be used to build more reliable models. The details of the methods undertaken, along with some preliminary results, are presented in the next chapter.

Chapter 3

Identifying Key Stages in a Disease Process from Cross-Sectional Data - Methods and Algorithms

This chapter deals with the different methods undertaken in this thesis as well as a formal definition of a Pseudo-Time-Series (PTS) is introduced for the first time. Multivariate Time-Series (MTS) modelling is explored in detail. The *Floyd-Warshall algorithm* [Flo62], a well-established algorithm that used to find the shortest path in a weighted graph (and therefore trajectories through data) will be described. In order to build more robust time-series models, the concept of a pseudo-time-series is introduced based upon the work in [TGH10], and a formal definition of this is further derived (as published in [LST12]). The details of the *temporal bootstrap* are illustrated with respect to this new formal definition. This is followed by a description of the performance of three typical models that are used for modelling time-series data: *hidden Markov models*, *Bayesian Networks* and *Dynamic Bayesian Networks*. Finally, a new algorithm for identifying key stages in pseudo-time-series is introduced (as published in [LT10]) which is focussed on in this thesis.

3.1 Multivariate Time-series Modelling with Pseudo Time-Series

A pseudo time-series is a sequence of observations measured over time, which aims to build multiple trajectories through cross-sectional data in order to approximate genuine longitudinal data. Building pseudo time-series involves plotting multiple trajectories through cross-sectional data based upon distances between data points, using prior knowledge of healthy and disease states to guide the trajectories. These trajectories can then be used to build approximate temporal models to make forecasts. However, models that can exploit multivariate time-series data can be very challenging to learn reliably. The formal definition of MTS is the following.

Definition: Let a dataset D be defined as a real valued matrix where m (rows) is the number of samples (here patients) - and n (columns) is the number of variables in the clinical test data. We define $D(i)$ as the ith row of matrix D. The vector $C = [c_1, c_2, ..., c_m]$ represents defined classes, where each $c_i \in \{0, 1\}$ corresponds to the sample i, $c_i = 0$ represents that sample i is a healthy case, and $c_i = 1$ represents that sample i corresponds to a diseased case. These classifications are based upon expert diagnoses.

A time-series is defined as a real valued T (row) by n (column) matrix where each row corresponds to an observation measured over T time points. If $T(i)$ was observed before $T(j)$ then $i < j$.

We define a set of pseudo time-series indices $P = \{p_1, p_2, ...p_k\}$ where each p_i is a T length vector where $T > 0$. We define p_{ij} as the jth element of p_i and each $p_{ij} \in \{1, ..., m\}$. We define the function $F(p_i) = [p_{i1}, ..., p_{iT}]$, creating a T by n

matrix where each row of $F(p_i) = D(p_{ij})$. Pseudo time-series can be constructed from each p_i using this operator. For example, if a pseudo time-series index vector $p_1 = [3, 7, 2]$ then $F(p_1)$ is a matrix where the first row is $D(3)$, the second row is $D(7)$ and the third row is $D(2)$. The corresponding class vector of each pseudo time-series generated by $F(p_i)$ is given by $G(p_i) = [C(p_{i1}), ..., C(p_{iT})]$.

To demonstrate this notation consider the following example:

Let the data matrix D be defined as:

$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \\ d_{41} & d_{42} & d_{43} \end{bmatrix}, d_{ij} \in \mathfrak{R}$$

Let the corresponding class vector be $C = [c_1, c_2, c_3, c_4]$. If $P = \{p_1, p_2\}$ where $p_1 = [1, 3, 1]$ and $p_2 = [2, 3, 1]$ then:

$$F(p_1) = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{31} & d_{32} & d_{33} \\ d_{11} & d_{12} & d_{13} \end{bmatrix}, G(p_1) = \begin{bmatrix} c_1, c_3, c_1 \end{bmatrix}$$

and

$$F(p_2) = \begin{bmatrix} d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \\ d_{11} & d_{12} & d_{13} \end{bmatrix}, G(p_2) = \begin{bmatrix} c_2, c_3, c_1 \end{bmatrix}$$

To summarise, we defined a set of k pseudo time-series with their associated class labels, sampled from the cross-sectional data D, indexed by the elements of p_i.

3.2 Floyd-Warshall Algorithm

We now briefly describe the Floyd-Warshall Algorithm [Flo62] which is used for the generation of pseudo time-series. Graph-theoretical approaches such as this are commonly used to find the shortest paths between all nodes for a weighted graph G. A weight matrix w_{ij} is an edge between node i and node j in graph G. An $m \times m$ matrix representing the edge weights of an n-node graph, where $W = (w_{ij})$. This algorithm is based upon a distance matrix $D^(k)$ which represents distance between data points $d(i)$ and $d(j)$, where $D^k = (d^k_{ij})$. A matrix d^k_{ij} is generated and represents the weight of the shortest path from i to j using a set of nodes $\{1,2,...k\}$ as intermediate nodes at iteration k. If k is not a node on the path, the shortest path has length d^{k-1}_{ij}, otherwise, the path is $d^{k-1}_{ik} + d^{k-1}_{kj}$. See Algorithm 1 for the full details.

Algorithm 1 The Pseudo code of Floyd-Warshall Algorithm.

Input: An $m \times m$ matrix W;

1: $D^0 = W$;
2: **for** $k = 1, ..., m$ **do**
3: $D^k = d^k_{ij}$;
4: **for** $i = 1, ..., m$ **do**
5: **for** $j = 1, ..., m$ **do**
6: $d^k_{ij} = min(d^{k-1}_{ij}, d^{k-1}_{ik} + d^{k-1}_{kj})$;
7: **end for**
8: **end for**
9: **end for**

Output: D^n

We will use this algorithm within the pseudo time-series construction in order to build trajectories between samples of cross-sectional data as detailed in the next section. We use Euclidean distance between data points to build the matrix W.

3.3 The Temporal Bootstrap

The temporal bootstrap is a resampling approach for building pseudo time-series as defined in Section 3.1. It involves resampling data from a cross-sectional study and repeatedly building trajectories through the samples in order to build more robust time-series models. Each trajectory begins at a randomly selected datum from a healthy individual and ends at a random datum classified as diseased. An extension of the temporal bootstrap is explored in this research, which allows us to identify intermediate stages in a disease process and sub-categories of known diseases with subtly different symptoms.

This method is compared to a strawman approach and its ability to inform us about the dynamics of the disease is examined. In particular, the ability of the method to explain the dynamics of disease progression, that results in trajectories through the data space starting at healthy data regions and ending at cases of advanced disease, is investigated. The hypothesis underlying the extended bootstrap approach is tested on biomedical data from three diseases in order to identify automatically the disease regions of interest at key junctions and the 'extreme' end of the trajectories. We use HMMs in conjunction with the EM algorithm for the identification of disease regions.

The data is firstly standardised to a mean μ of zero and a standard deviation σ of one as we found that this led to better HMM models. The elements of p_i (as described

in Section 3.1) are determined based upon a uniform random sampling procedure with replacement. The ordering of the elements in p_i is based upon randomly selecting a *start* and an *end* in p_i such that the associated classifications are $c_{start} = 0$ and $c_{end} = 1$. This means that the time-series will progress from a healthy state to a disease state. The ordering is then determined by the shortest path, calculated based upon the Floyd Warshall algorithm [Flo62] (Algorithm 1) applied to the Euclidean distance matrix, R_{ij} between samples in $F(p_i)$. See Algorithm 2 for the full details.

Algorithm 2 The Temporal Bootstrap for Learning Pseudo Time-Series Models from Cross-Sectional data.

Input: Cross-section data *D*; class labels *C*, sample size *T*; number of pseudo time-series *k*;

1: Standardise dataset D to $\mu = 0$ and $\sigma = 1$;
2: **for** i=1 **to** *k* **do**
3: Uniformly randomly sample (with replacement) *T* row indices from *D* to create d_i such that there is at least one healthy and one diseased class (in *C*) corresponding to any of the indices in d_i;
4: Uniformly randomly select a row index from d_i, start, from where $1 \leq start \leq$ T and an endpoint, end, where $1 \leq end \leq T$ where $C(d_i, start)$ represents a healthy class and $C(d_i, end)$ represents a diseased class;
5: Construct a $T \times T$ matrix, W_i, of Euclidean distances between each $D(d_{ia})$ and $D(d_{ib})$ for all combinations of indices in d_i;
6: Order d_i to create d_i^* based upon the shortest path between $D(d_i, start)$ and $D(d_i, end)$ given the weighted graph W_i using the Floyd-Warshall algorithm constrained so that every index in d_i is included in the path;
7: Add the ordered d_i^* to the set of pseudo time-series P;
8: **end for**
9: Use the set *P* of *k* pseudo time-series to train a time-series model

Output: Pseudo Time-Series Model

As an example we can explore how well multivariate time-series models can be reverse-engineered from cross-sectional data by simulating the cross-sectional study

process. Figure 3.1 shows the result of simulating a varying number of time-series from an autoregressive hidden Markov model (ARHMM) with two disease states and one healthy state. The data shown is a result of sampling a single point from each series randomly (essentially generating a cross section of the population of time-series).

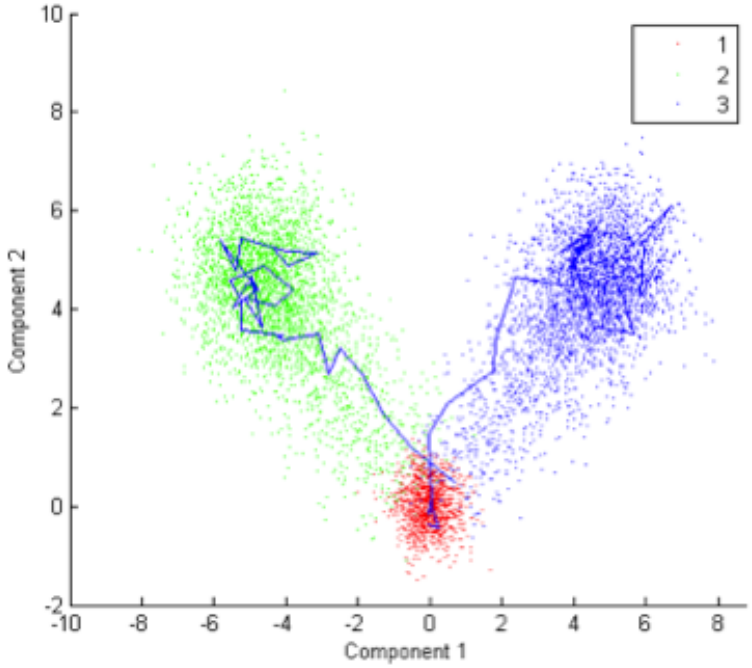

Figure 3.1: Scatter plot of the first two components using multidimensional scaling on simulated data (generated from an Auto-regressive hidden Markov model (ARHMM) with 3 states, one representing healthy control patients - red dots, and two representing different disease symptoms - green and blue dots). Two of the original MTS are plotted along with the full cross sectional data (one sampled form each MTS).

We can then use the temporal bootstrap to learn pseudo time-series prior to building a pseudo temporal model. The error rate (Table 3.1) and classification accuracy (Table 3.2) resulting from the pseudo time-series models by using Temporal Bootstrap (TBS) are shown, compared with the statistics generated from a model learnt from the original

multivariate time-series (Full MTS). It should be noticed that the TBS results actually appear to be better than the model inferred from the full MTS. This is because the resampling process in the TBS procedure smooths the data and also shown the results of the full MTS after smoothing, which are the most accurate (as would be expected - it is highly unlikely that the pseudo time-series will generate more accurate models). However, as the sample size increases and approaches 500, the statistics appear to almost converge (results taken form [TGH10]).

We have been testing this approach on real cross-sectional datasets (VF, BC and PD's), where we can validate the outcome using longitudinal data, revealed similar results. The next section illustrates the sample pseudo time-series generated from those three cross-sectional datasets.

Length	Full MTS	TBS	MTS smoothed
50	0.129 ± 0.039	0.251 ± 0.228	0.095 ± 0.055
100	0.126 ± 0.023	0.158 ± 0.121	0.086 ± 0.012
250	0.126 ± 0.013	0.079 ± 0.034	0.084 ± 0.014
500	0.125 ± 0.015	0.067 ± 0.023	0.084 ± 0.014

Table 3.1: Mean Forecast Sum Squared Error and 95% Confidence for Model Learnt using the Temporal Bootstrap on Cross-Sectional Data (TBS), the Original Time-Series with smoothing (MTS smoothed) and without (MTS).

Length	Full MTS	TBS	MTS smoothed
50	0.907 ± 0.047	0.897 ± 0.092	0.903 ± 0.055
100	0.905 ± 0.045	0.912 ± 0.044	0.905 ± 0.048
250	0.905 ± 0.046	0.910 ± 0.046	0.905 ± 0.048
500	0.905 ± 0.046	0.912 ± 0.044	0.904 ± 0.049

Table 3.2: Mean Classification Forecast Accuracy and 95% Confidence for Model Learnt using the Temporal Bootstrap on Cross-Sectional Data (TBS), the Original Time-Series with smoothing (MTS smoothed) and without (MTS).

3.4 Predictive Models and Their Mathematical Description

Having built trajectories from cross-sectional data using pseudo time-series approaches, we then explored different techniques to model these sequences. As discussed in Section 2.3.1, statistical approaches such as the Box Jenkins model [BJR13] are data intensive and can be problematic when modelling uncertain and noisy data. Probabilistic graphical models, however, have the natural ability to deal with this sort of data and also allow for combining data with expert knowledge or, in case there are no data, rely entirely on expert knowledge. This makes probabilistic graphical models such as hidden Markov models (HMMs) and dynamic Bayesian networks (DBNs) particularly attractive and popular in practice.

3.4.1 Mathematical Description of Hidden Markov Model (HMM)

A popular probabilistic model for modelling sequential and time-series data is known as the Hidden Markov Model (HMM)[Rab89], which assumes a single discrete hidden state, H and a continuous observed process, X. HMMs have widespread application in a variety of tracking scenarios, from speech recognition to clinical analysis ([Rab89], [WP02], [ZOM99], [Edd98], [LKBJ08], [SBR07]). Two characteristics of HMMs provide [Rab89] strong support for the choice to use them in this research: firstly, the models have a 'complete' mathematical architecture and refer to a wide range of applications; Secondly, they are a very accurate approach for some important applications when applied properly. See Figure 3.2 shows the general architecture of HMMs, where directed links determine conditional probability distributions. It involves one hidden

variable that is conditioned on the hidden variable at the previous time point and *n* observed variables that are conditioned upon the hidden variable at the same time point. The transition equation is modelled using a discrete distribution of the state *H* at time *t*, H^t conditioned upon the state at $t-1$. This is written as $p(H^t|H^{t-1})$ (and gives rise to the link $H^{t-1} \rightarrow H^t$ in Figure 3.2). The measurement equation is captured using the distribution of each variable at time *t* conditioned upon the hidden state at time *t*, written as $p(X^t|H^t)$ (giving rise to the links, $H^t \rightarrow X_1^t$, $H^t \rightarrow X_2^t$, ..., $H^t \rightarrow X_N^t$). The forward algorithm allows the probability of the hidden state at time t, H^t to be estimated from the previous and current values of measured variables, $X^{1\ldots t}$. To predict $p(H^t|X^{1\ldots t})$, H^t represents the hidden state at time *t* and X^t represents the variables in the time series. This process is known as filtering and can be used to estimate future probabilities of disease states from historical longitudinal data.

There are three basic problems associated with learning HMMs which are useful in real-world applications [Rab89]:

- Problem 1: Given the observation sequence and a model, how to efficiently compute the probability of the observation sequence, given the model?
- Problem 2: Given the observation sequence and the model, how to choose a corresponding state sequence which is optimal in some meaningful sense?
- Problem 3: How to adjust the model parameters to maximize the probability of the observation sequence, given the model?

Predictive models such as HMMs contain unknown parameters, known observed data and latent variables. The forward or backward procedure of HMMs enable to be

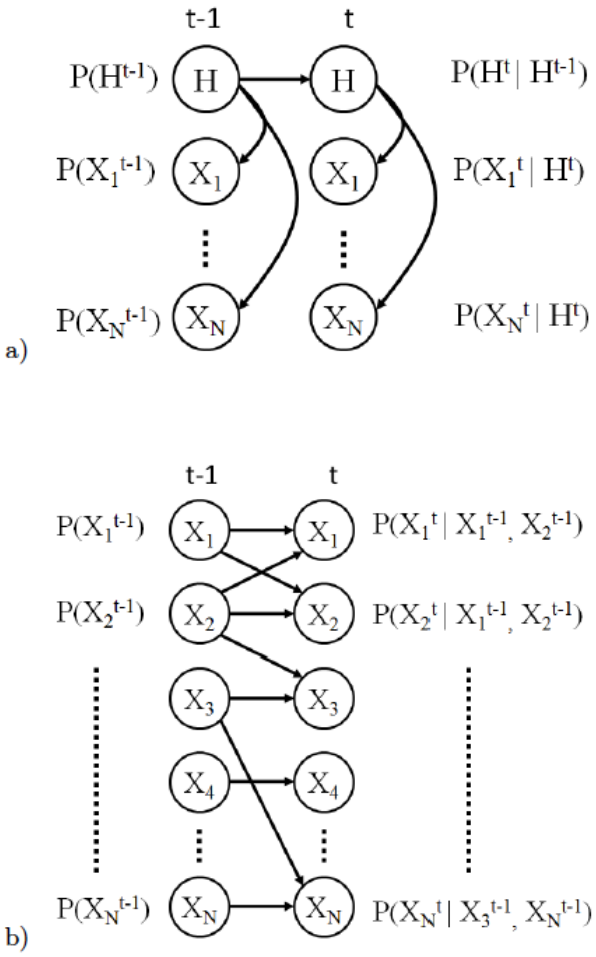

Figure 3.2: Architectures of Hidden Markov Models (top) and Dynamic Bayesian Netowrks (bottom).

used because it is more efficient than direct evaluation. In order to find optimal or the best state sequences in continuous nature, the sources are non-stationary vary over time. The Baum-Welch algorithm can be used to solve the three problems where are listed above in the learning process of HMMs. The Baum-Welch algorithm [Bag01] is basically the application of expectation-maximization (EM) algorithm [B$^+$98] to HMMs. EM is an iterative method, which can be used to find the maximum likelihood

of parameters in statistical model with unobserved latent variables. Implementing the EM algorithm for HMMs is straightforward and consists of two steps: first, the estimate of the parameters for a HMM given a set of observed data (expectation step - E step), is used to allow the evaluation of expectation of the log-likelihood; then the maximum likelihood is estimated (maximization step - M step), in order to maximize the expected log-likelihood found in the E step. The particular application of the EM algorithm for this work can be described as ([Wik13],[Bor09]):

Expectation Step: Calculate the expected value of the log-likelihood function.

$$Q(\theta \mid \theta^t) = E_{Z|X,\theta^t}\left[logL\left(\theta; X, Z\right)\right] \tag{3.4.1}$$

Maximization Step: Maximize the quantity of the parameter.

$$\theta^{(t+1)} = \underset{\theta}{argmax} Q(\theta \mid \theta^t) \tag{3.4.2}$$

where,

- X is a set of observed data associated with each data point.
- Z is a set of unknown latent variables, extract from a fixed number of values and each observed data point has a latent variable.
- θ is a continuous vector of unknown parameters, which are associated with all data points, and also with a particular value of a latent variable.
- $Q(\theta \mid \theta^t)$ represents the conditional distribution of Z given X under the current estimate of the parameters θ.

3.4.2 Mathematical Description of Bayesian Networks (BNs)

Bayesian Networks [Bar12] are probabilistic networks that provide a way to represent the independence assumptions made in a distribution. A Bayesian Networks defined as follows [MP01b]:

$$P(X_1, ..., X_t) = \prod_{i=1}^{N} P(X_i \mid Pa(X_i)) \tag{3.4.3}$$

where $Pa(X_i)$ is the parent set of a node X_i.

The three elements of DAG (the conditional probabilities, the structure and joint probability distribution) can be used to estimate the probability or likelihood of each variable or state. Mihajlovic [MP01b] points out three rationales why BNs are useful:

- *From known causes to unknown effects (causal reasoning)*(e.g we know the disease exist and the current state, and aim to explore and predict the end state of the disease - longitudinal studies.), and
- *From known effects to unknown causes (diagnostic reasoning)* (e.g we known the several end states of disease, but we aim to explore the causes of the disease and the progression - cross-sectional studies.), or
- *For any combination of these two.*

If we want to know all of the conditional probabilities throughout the network we need to adjust the parameters of the network, so that 'Learning' plays an important role to enable us to overcome the problems. Many different approaches have been used to learn good BNs structures, such as the K2/K3 algorithms ([CH92], [Bou93]),

the Branch and Bound technique [Suz93], and evolutionary methods ([LPY$^+$96], [WLL99]). K2/K3 can be described as greedy search which explores the effect of adding each of the possible links to the current structure based on an empty structure with no links and the one that finishes with the best score is selected. K2/K3 use this algorithm with a log likelihood metric and a description length metric, respectively. The Branch and Bound technique is used to perform an effective thoroughgoing search by stopping any further exploration along a search route found on an edge which is deliberate on the scoring metric. When evolutionary methods are applied to static BNs the application of various operators is compulsory to prevent the generation of recurrent event within the network.

Learning BNs is the process of scoring candidate network structures. The log likelihood ([CH92], [Gei92]) and the Description Length (DL) ([LB94], [Suz93]) are two scoring metrics. The log likelihood is calculated by using expression 3.4.4, and the higher score indicates the best structure that can be obtained to fit the dataset. The Description Length metric is constructed from the summation of the description length of a network structure (expression 3.4.5) and the description length of encoding the dataset given that model (expression 3.4.6). In contrast to the log likelihood, the lower this score is, the better the structure fits the dataset.

$$log \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(F_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} F_{ijk} \tag{3.4.4}$$

$$\sum_{i=1}^{n} |\pi_i| \, log(n) + \left((r_i - 1) \prod_{j \in \pi_i} r_j \right) \tag{3.4.5}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} -F_{ijk} \times log\left(\frac{F_{ijk}}{F_{ij}}\right) \tag{3.4.6}$$

where,

- n is the number of nodes.
- F_{ijk} is the frequency of occurrences in the dataset that the node x_i takes on the value v_{ik} (where there are r_i possible instantiations).
- the parent nodes take on the instantiation w_{ij} (where there are q_i possible instantiations) and the parent nodes π_i take on the instantiation w_{ij} (where there are q_i possible instantiations).
- and $F_{ij} = \sum_{k=1}^{r_i} F_{ijk}$.

3.4.3 Mathematical Description of Dynamic Bayesian Networks (DBNs)

Dynamic Bayesian networks (DBNs) ([FMR98], [Gha98]) are an extension of the standard Bayes network as discussed in the Section 2.3.2.1, which are used to model probability distributions of random variables. DBNs have similar hidden process as hidden Markov models, but they are more general than hidden Markov models, which assume a single discrete variable representing a hidden state and possible multiple observation variables. Each state at time t depends on its past given state at time $t-1$. For complex structure, it may also depends on more past states in the same time occurrence. A Dynamic Bayesian network is defined as follows[vGTL08]:

$$P(X_{1\ldots t}) = \prod_{t=1}^{T} \prod_{i=1}^{N} P(X_t^i) \mid Pa(X_t^i)) \tag{3.4.7}$$

where,

- X_t^i is the ith node representing the variable at time slice t with nodes corresponding to a set of random hidden-state variables $X(t)$,
- $Pa(X_t^i)$ are the parent of X_t^i in the graph,
- P is the joint probability distribution (JPD) of variables in $X(t)$.

DBNs are the probabilistic directed graphical models in which each node represents a variable at a particular time-slice. Each arc in the graph represents a probabilistic relationship. The lack of arcs between nodes represents conditional independence assumptions. The arcs connect parent nodes to child nodes and form a directed acyclic graph (DAG), i.e. no directed cycles are admitted. Links can occur both in the same time slice and between different time slices. Each node is associated with a conditional probability distribution (CPD), which describes the probability of each possible value of the variables given their parents. Once the structure of the network and the CPDs are obtained, it is possible to infer the value of any node. In fact, all the CPDs of the DBN provide an efficient factorization of the joint probability of the variables in the model. A simple example is shown in Fig 3.2. To build a DBN, the structure of the network and the three set of parameters (state transition, observation probability distribution function and initial state distribution) (i.e. the CPDs) of all the variables must be obtained. Typically, the CPDs are learned from data by maximizing the posterior probability of the parameters given the data [HGC95]. When the structure is fixed and

the prior distribution of the parameters is uniform, this corresponds to maximizing the likelihood function.

In the learning process of DBNs, a similar approach to standard BNs can be adopted (such as a search and score with an appropriate metric e.g. log likelihood). For example, the REVEAL algorithm [LFS+98] is a modified equivalent of the K2 algorithm. In [TLOS01], a number of existing BN learning algorithms were adapted for DBNs. However, in many time-series, there are changes in the underlying distributions and standard DBNs cannot take this into account as they are time-invariant. More recently, non-stationary DBNs have been explored where both a model parametrisation and a segmentation process are performed to identify these changes in structure. However, the search space is usually limited by constraining one or more degrees of freedom, i.e. the segmentation points of the time series, the parameters of the variables, the dependencies between the variables and the number of segments for the model. Among the most recent and complete work, Talin and Hengartner [TH05] used a MCMC approach to estimate the variance structure of the data, but the search space was limited to a fixed number of segments and for learning undirected edges only. Xuan and Murphy [XM07] proposed an approach to model changing dependency structures from multivariate time series, but also in this case the search was limited to undirected edges. Robinson and Hartemink [RH10] formalized the concept and proposed a solution that tackles all the degrees of freedom described except for the parameters. Grzegorczyk and Husmeier [GH09] instead retained the stationarity of the structure in favour of the parameters flexibility, arguing that structure changes lead almost certainly to over-flexibility of the model with short time-series. While the first approach

may only capture parameter changes that are strong enough to give rise to a structural change, the latter may not model correctly underlying conditional dependencies over the stages. The ability to assess both weak and strong changes in variable distributions and explicitly model the evolution of their relationships would be extremely useful from an informative point of view, especially in unknown processes such as glaucoma. In [CGHCT12] a form of DBN that clusters sections of time series whilst simultaneously learning DBN structure and parameters was used to model glaucoma patients. The Bayesian Information Criterion (BIC) [Sch78] was used in conjunction with Simulated Annealing (SA)[KJV83] for learning both the BNs and the clusters. BIC incorporates a penalizing factor that is proportional to the number of parameters in the model and the number of cases in the data, and helps to prevent overfitting.

3.4.4 A proposed Algorithm for Identifying Key Stages in a Disease Process

An algorithm is proposed in this PhD study for identifying key stages in a disease process. A key novelty of this work is the use of the temporal bootstrap in conjunction with Algorithm 3 described, below, to generate and explore the transitions of the different states within the trajectories that are discovered from the selected data. As shown in the algorithm, essentially, it starts by searching for hidden states $h = 3$ (one more than the original 'healthy' and 'disease'), whilst learning a HMM. The HMM is chosen here in conjunction with the Expectation Maximisation algorithm for its ability to learn hidden underlying states in a temporal process.

The Expectation Maximisation (EM) algorithm [B$^+$98] can be used to cluster a sequence of data into different sections. It consists of two steps:

- The E step becomes the sum of expectations of sufficient statistics
- the M step involves maximizing a linear function.

In order to cluster the data into increasingly fine-grain regions using the EM algorithm, the EM model fitting process is repeated for increasing values of h and the transition matrix of the HMM is explored manually each time for interesting features. Thus, this iterative interactive approach is a good way to ensure interesting features are identified. In addition, it is can be used to minimize the cost function of the network. The process is detailed in Algorithm 3 below:

Algorithm 3 An iterative algorithm for identifying key stages in a disease process.

Input: A set of pseudo time-series, P generated from cross section data D and associated labels C;

1: Remove the class labels C from P;
2: Set $h = max(C)$ (the number of classes) + 1;
3: **repeat**
4: Train a HMM on the P with h hidden states using the EM algorithm [B$^+$98];
5: $h = h + 1$;
6: **until**
7: Transition table in the parameterised HMM captures disease features of interest (e.g. it features more than one clear end state);

Output: HMM with new intermediate or end states

By applying this approach we aim to identify different key stages in a disease process [LT10]. Furthermore, the ordering of the discovered sequences (i.e. the pseudo time-series) should lead to more informative clusters and transitions than simply clustering the unordered cross-sectional data using, for example, standard clustering such as K-means [HW79].

3.5 Summary

In this chapter, different approaches to modelling trajectories through clinical data have been examined. Firstly, two methods to infer time-series from data are discussed. This includes hidden Markov models and dynamic Bayesian networks which are especially good at dealing with uncertainty and noise. Secondly, some new techniques for building trajectories through cross-sectional data are explored with a focus on sequence reconstruction. It is clear that many of the longstanding approaches to modelling disease progression are proving inadequate to deal with issues of uncertainty in the dynamic and measurement processes and the ability to integrate cross-sectional studies with longitudinal studies. The approaches and techniques discussed in this chapter will be further discussed in the next two chapters. Examples of real cross-sectional data and simulated data that are designed to mimic the cross-sectional study data collection process will be described with analyses of pros and cons in the next chapter.

Chapter 4

Identifying Key Stages in a Disease Process from Cross-Sectional Data: Experiments and Results

This chapter presents the results from the experiments of using methods and algorithms described in chapter 3, utilising both simulated data and the three clinical datasets are discussed in this chapter. The *simulated results* consist of time-series generated by an autoregressive HMM. The *sample trajectories* are constructed using the temporal bootstrap on each dataset, plotting the first two components, following multi-dimensional scaling. The comparison of the trajectories within a medical context is illustrated. Following that, the *State-End diagrams* used to represent the transition probabilities between the different states are discovered. The mean values of the clinical data are presented for both healthy and diseased patients for comparative purposes. Additionally, the expected data values for each state associated with the state-end diagrams are also discussed, based upon the HMM learnt from the unlabelled pseudo time-series. All the results in this chapter have been published on number 2 [LT10] and 3 [YSA13] publications.

4.1 Simulated Data and Experiment Setups

In order to assess the performance of the different approaches explored in this section, a number of simulated datasets are used which are now described. In power analysis, simulation is a simplified imitation of the operation of a real-world process through time, that follow a particular distribution and calculating the test statistic from each sample. So that the significance level and power of the procedure may be investigated. In this research, a number of simulated time-series are generated from an autoregressive HMM (ARHMM) ([HDA01], [KIM03], [KT02]) with two variables for ease of visualisation (essentially it is a bivariate Gaussian). 200 time-series datasets are generated with four or five discrete states in order to determine the trajectory of each time-series. The number of samples was chosen to reflect the typical sample sizes from the available clinical data. An autoregressive hidden Markov model (ARHMM) was used as to capture the relatively smooth transition from healthy to disease states through autoregressive dependencies. This smooth transition is typical in many medical data and particularly in progressive diseases such as glaucoma. For some diseases, this may not hold if their symptoms fluctuate, appearing and disappearing over the course of the disease.

The states and the transitions between 'healthy' and 'diseases' are based upon the discussion on multiple end points in Chapter 2 (Albert [Alb99]). The states in the first simulated dataset represent a starting healthy region and one diseased region with two intermediate states, whilst the second dataset involves one starting state, two end states and two intermediate states. The observed variables are Gaussian and conditioned upon the hidden state and the same variable at the previous time point. The length of

each multivariate time-series is set to 30, as this reflects common longitudinal studies in the biological and medical literature. The hidden variable always starts in the healthy state and has a probability distribution that determines the probability of change to the intermediate and end states. One point from each time-series is then sampled to form a cross-sectional dataset. This process is similar to that used in [CK02] and an example of the cross-sectional data is plotted in Figure 4.1 including some time-series examples.

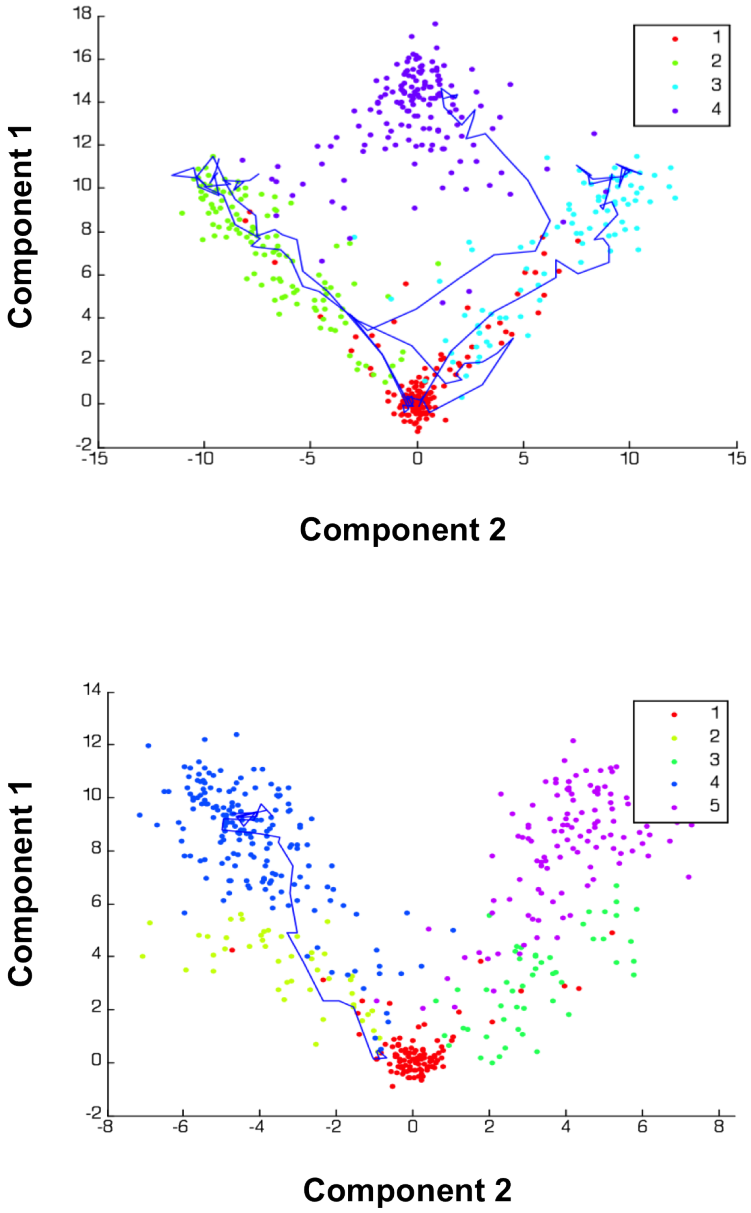

Figure 4.1: Two simulated datasets generated using autoregressive HMMs with two variables to model disease processes. The plots show a single sampled point from each time-series (dots) along with some of the original time-series (lines). One dataset has 1 healthy state, 1 disease state and 2 intermediate states (top); the second dataset has 1 healthy state, 2 disease states and 2 intermediate states (bottom).

4.2 Simulated Results

To demonstrate the power of the proposed methods, the two simulated time-series datasets are generated (see Figure 4.1) using an autoregressive HMM with two variables to model a disease process with both healthy state and different disease states. The plots show a single sampled point from each time-series (dots) along with some of the original time-series (lines). One of the datasets has 1 healthy state, 1 disease state and 2 intermediate states; the other one has 1 healthy state, 2 disease states and 2 intermediate states. The underlying reason for that choice is to capture complex advance disease scenarios as highlighted in the literature (specifically [Alb99]). We sample one point from each time-series generated in this data to represent cross-sectional data and use it to generate Multivariate Pseudo Time-series (MPTS).

The resulting transition matrices generated from applying Algorithm 2 (see Section 3.3 in Chapter 3) to the two simulated datasets are shown in Tables 4.2 and 4.4, and the visualisation of transition matrix are shown in Figure 4.3 and 4.5. By comparing with the original matrices in Tables 4.1 and 4.3, it can be seen that although the precise probabilities are not discovered, the general characteristics of many of the states are actually found (see the visualisation of original matrix from Figure 4.2 and 4.4). The stable end states (state 4 in Table 4.2, and states 4 and 5 in Table 4.4) with spurious dynamics are the examples for this. Testing the discovered model using the original (unseen) full time-series data gives a mean accuracy of 95% (see Section 3.3 in Chapter 3), indicating that a very reliable time-series model can indeed be learnt from cross-sectional data if the sample is sufficiently large enough. For example, states marked with an asterisk correspond to zero probabilities in the original model. Although these

spurious correlations are small, it is found though that as the sample size of the original data is reduced (to 50 or less), they become more of an issue. They are likely to have occurred because the learnt HMM overfitted spurious relationships in the MPTS through implicit correlations. In summary, the main characteristics of the transitions are preserved when the sample size is appropriately high. A small number of spurious correlations between impossible transitions were generally observed but these were always very low values. This gives a confidence in exploring the real-world clinical datasets with similar sample sizes.

$H_{t-1} \setminus h_t$	1	2	3	4
1	0.800	0.100	0.100	0.000
2	0.050	0.900	0.000	0.050
3	0.050	0.000	0.900	0.050
4	0.000	0.000	0.000	1.000

Table 4.1: Transition matrix for hand-coded simulated data with 4 states.

$H_{t-1} \setminus h_t$	1	2	3	4
1	0.833	0.061	0.091	0.015*
2	0.277	0.566	0.000	0.157
3	0.213	0.000	0.729	0.058
4	0.000	0.025*	0.077*	0.898

Table 4.2: Transition matrix learnt from PTS discovered from cross-sectional sample of simulated time-series with 4 states.

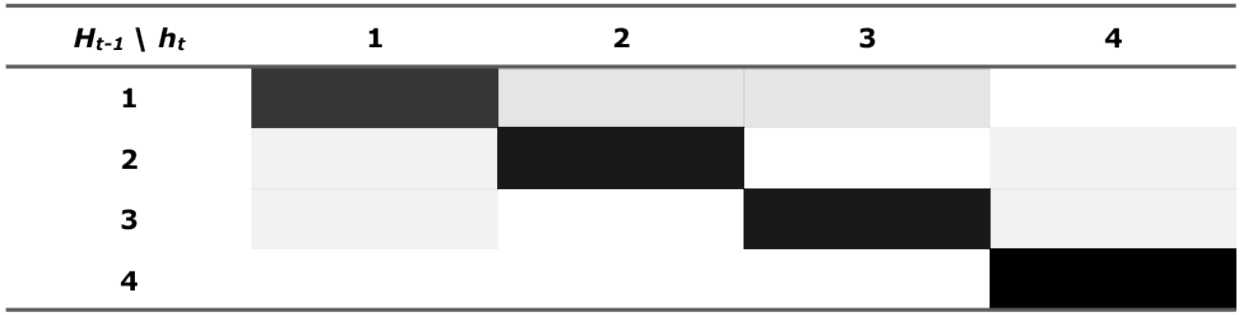

Figure 4.2: The visualisation of transition matrix for hand-coded simulated data with 4 states.

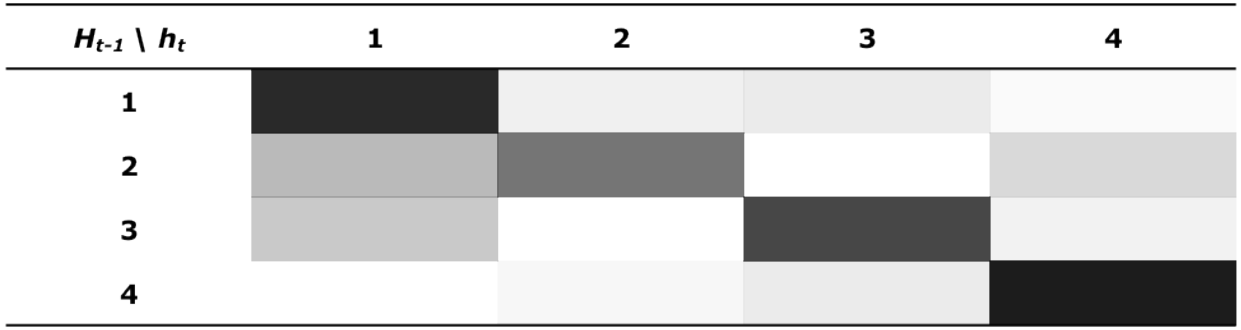

Figure 4.3: The visualisation of transition matrix learnt from PTS discovered from cross-sectional sample of simulated time-series with 4 states.

$H_{t-1} \setminus h_t$	1	2	3	4	5
1	0.800	0.050	0.050	0.050	0.050
2	0.050	0.900	0.000	0.050	0.000
3	0.050	0.000	0.900	0.000	0.050
4	0.000	0.000	0.000	1.000	0.000
5	0.000	0.000	0.000	0.000	1.000

Table 4.3: Transition matrix for hand-coded simulated data with 5 states.

$H_{t-1} \setminus h_t$	1	2	3	4	5
1	0.892	0.028	0.046	0.015	0.020
2	0.076	0.686	0.000	0.238	0.000
3	0.119	0.000	0.577	0.000	0.304
4	0.002^*	0.016^*	0.000	0.951	0.031^*
5	0.086	0.000	0.070^*	0.004^*	0.841

Table 4.4: Transition matrix learnt from PTS discovered from cross-sectional sample of simulated time-series with 5 states.

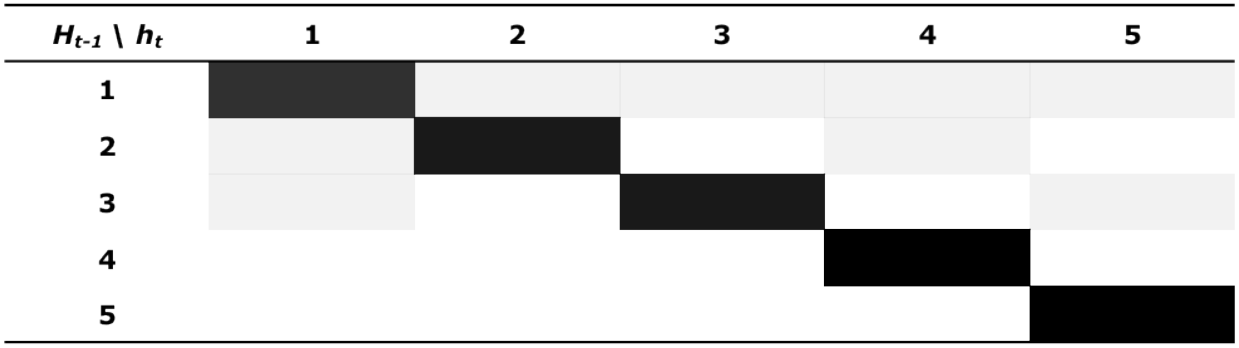

Figure 4.4: The visualisation of transition matrix for hand-coded simulated data with 5 states.

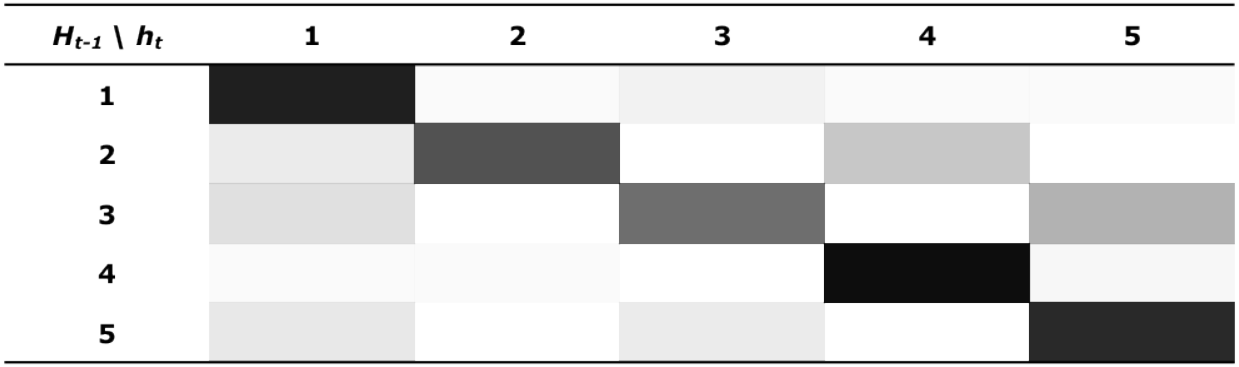

Figure 4.5: The visualisation of transition matrix learnt from PTS discovered from cross-sectional sample of simulated time-series with 5 states.

4.3 Real-World Cross-sectional Datasets

The three datasets used in this study are named after the disease conditions.The first dataset - Visual Field (VF) and Heidelberg Retina Tomography (HRT) data is from glaucoma sufferers. Glaucoma is an eye disease, characterised by progressive loss of vision and is the second major cause of blindness worldwide. The dataset is built based upon two kinds of medical tests for glaucoma. The second dataset - Breast Cancer (BC) data is from breast cancer sufferers. Breast Cancer is a kind of cancer from breast tissue that supply the ducts with milk. The majority of disease cases occur in women, the incidence of breast cancer is rising among women in many European countries, profoundly affecting up to 1 in 16 women [Org10]. The dataset is set upbased upon tumour examinations for breast cancer. The last dataset - Parkinson's disease (PD) is from a Parkinson's disease sufferers. Parkinson's disease is idiopathic disease, characterised by a degenerative of the central nervous system and having no known cause, most of cases occurring in the old people. Because of the ageing of the world, parkinson's disease has become an increasing public health issue due to the ageing of the world's population [Org98]. The dataset is built based on speech pattern data for Parkinson's disease. Although, these diseases are very different, the datasets share a similarity in terms of how data has been collected, as all of them come from cross-sectional studies over a relatively large population. These sampled cross-sectional data are used to generate the Pseudo Time-Series. The three clinical datasets are described individually in the next sections. Table 4.5 gives a summary of the datasets:

Dataset	Number of Variables	Number of Cases	%age of Healthy Controls
Visual Field	12	163	51.9
Breast Cancer	10	565	62.7
Parkinson Disease	22	195	24.6

Table 4.5: Summary table of the 3 datasets.

4.3.1 Visual Field Test and Heidelberg Retina Tomography Data

The Visual Field (VF) test assesses the sensitivity of the retina to light. It is typically measured by automated perimetry, a technique in which the subject views a dim background as brighter spots of light are shone onto the background at various locations in a regular grid pattern. The brightness at which the subject sees the spots of light is related to the retinal sensitivity. There are many diseases and conditions that affect the VF, the most common being neurological disease and glaucoma. For this study, the data are aggregated into average values based upon their association with one of 6 nerve fibre bundles based upon the mappings in [KCO02]. The other type of data that we explore are obtained by Heidelberg Retinal Tomography (HRT) [LCS$^+$06] which involves generating images of the retina in order to calculate certain measurements associated with the three dimensional shape of the optic nerve head. These include neuro-retinal rim area measurements, which are used for the experiments in this study. The measurements are calculated for 6 different segments of the retina: nasal (n), nasal inferior (ni) and superior (s), temporal (t), temporal inferior (ti) and temporal superior (ts). For the experiments in this study, we combine the VF and HRT datasets to see if trajectories that are identified capture the interaction between the two data types as glaucoma progresses. The VF and HRT data are taken from a study of approximately 163 people [Log05] and each patient is classified into healthy or glaucomatous based

upon the VF data, using a pre-defined algorithm (AGIS). This may result in a bias towards the VF data for the building of the pseudo time-series, however as the relabelling algorithm involves learning the states from scratch we do not envisage it biasing our final disease stages.

4.3.2 Breast Cancer Data

The 565 Breast Cancer data (BC) are classified into 212 malignant and 357 benign cases. Ten real-valued features are computed for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The original dataset is described in more detail in Wolberg and Mangasarian and is available from the UCI machine-learning repository [A.K96].

4.3.3 Parkinson's Disease Data

Parkinson's disease (PD) affects movement and motor-related symptoms including speech. Vocal impairment can be the earliest indicator for the onset of PD. Therefore, a number of voice measurements have drawn significant attention for detecting and tracking the progression of symptoms of PD [Mit97]:

- The average vocal fundamental frequency (MDVP:Fo(Hz))
- The maximum vocal fundamental frequency (MDVP:Fhi(Hz))
- The minimum vocal fundamental frequency (MDVP:Flo(Hz))
- Several measures of variation in fundamental frequency (MDVP: Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP)

- Several measures of variation in amplitude (MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA)
- Two measures of ratio of noise to tonal components in the voice status (NHR, HNR)
- Two nonlinear dynamical complexity measures (RPDE,D2)
- The signal fractal scaling exponent (DFA)
- Three nonlinear measures of fundamental frequency variation (spread1, spread2, PPE)

The PD dataset is composed of 195 biomedical voice measurements, 147 of which are from Parkinsons disease patients and 48 from controls. The original dataset was obtained by McSharry and Roberts and is available on the UCI machine-learning repository [MSP05].

4.4 Biomedical Experiments and Results

In this research, the disease region identification is compared with standard clustering techniques on real-world medical cross-sectional data. For the real world data, the true underlying state transitions are unknown. However, the discovered transitions can be explored in the medical context. The biomedical results provide some sample trajectories when using the temporal bootstrap on each dataset. Points are identified as healthy and diseased using red/dark grey and blue/light grey respectively. These figures are generated using multidimensional scaling (calculated with Euclidean distances). This enables us to visualise many more variables in two dimensions. The 'state diagrams' (it is not a graphical model, it displays the non-zero entries of the transition matrix) are given to represent the transition probabilities (p) between the different states discovered. Associated with these diagrams, the expected values and clustering values for the data for each state are produced, based upon the HMM learnt from the unlabelled pseudo time-series. The mean values for the data for healthy patients and diseased patients are shown for comparison. Note that Algorithm 2 given in Chapter 3 involves increasing values for the hidden state h. Results provided here are only for the models with the number of states that are the best compromise between finding new interesting end and intermediate states, but not with trivial states that are simply side-effects of the data. For the glaucoma data, 4 states were found which best met this balance; similarly for the breast cancer, 5 states were gained; and for the Parkinson's data with 3 states. A larger value for h led to the splitting of interpretable states into ones with less clear significance.

4.4.1 Glaucoma

4.4.1.1 The Trajectories - Glaucoma

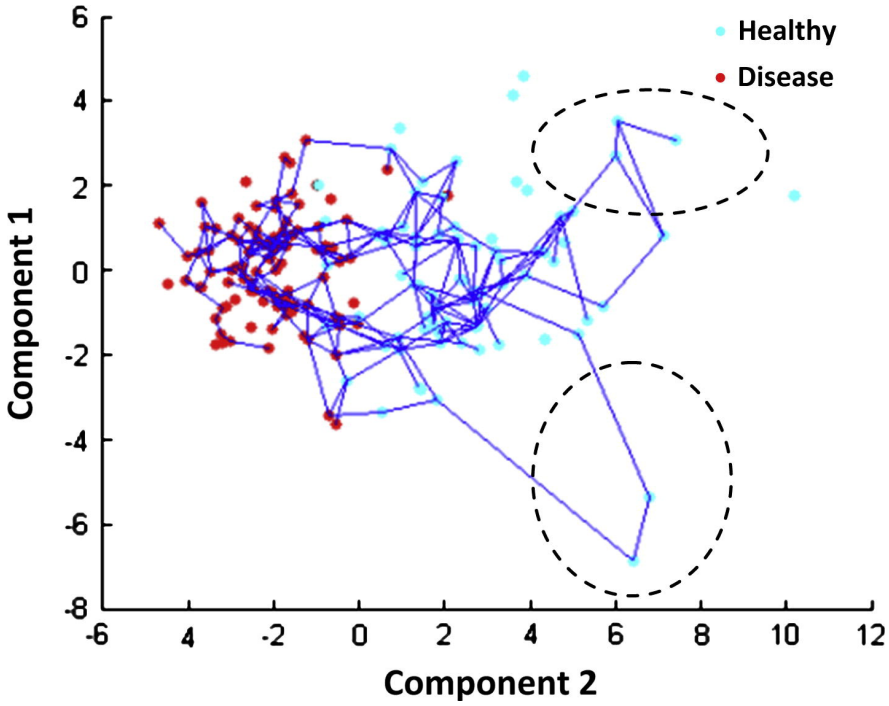

Figure 4.6: Typical trajectories learnt from the combined VF and HRT data plotted using multidimensional scaling with Euclidean distance. Normal VFs are marked in red and glaucomatous in blue.

Figure 4.6 shows sample trajectories that are discovered when building the pseudo time-series from healthy to disease regions. It also highlights that there could be two distinct regions of diseased state (in the top-right and bottom-right of the plot, marked with two dashed ellipses). Although this is may not be shown the states properly at this stage, the analysis below will confirm it through the relabelling scheme which is used to identify the sequence transitions and their medical context.

4.4.1.2 End-State Analysis - Glaucoma

The glaucoma state transition diagram generated from the transition matrix of the learnt HMM is displayed in Figure 4.7. The figure shows transitions with a $p > 0.15$ as solid lines and $p > 0.05$ as dashed. The full transition matrix is given in Table 4.6 where h_t represents the hidden state h at time t.

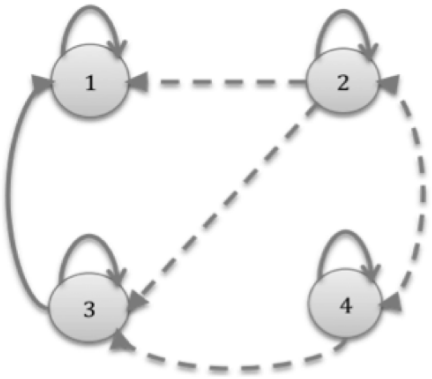

Figure 4.7: State Transitions for Glaucoma data. State 4 coincides with the starting healthy state, 1 and 2 appear to represent relatively stable end states and 3 appears a transitory state (a $p > 0.15$ is shown as a solid line and $p > 0.05$ as dashed).

$H_{t-1} \setminus h_t$	1	2	3	4
1	0.975	0.025	0.000	0.000
2	0.051	0.746	0.090	0.113
3	0.275	0.000	0.682	0.042
4	0.000	0.070	0.099	0.831

Table 4.6: Transition matrix for discovered VF states.

The table and the diagram show that there appear to be three relatively stable states: 1, 2 and 4. State 4 coincides with the starting healthy state, 1 and 2 appear to represent relatively stable end states and 3 a transitory state. These states are further explored by calculating the expected values of the variables associated with each state (see Figure

4.9). The clustering values of the variables discovered using k-Means clustering (see Figure 4.10) and compared with the mean values for normal and glaucomatous data in Figure 4.8. All values were standardised to have a $mean = 0$ and $sd = 1$, to void one of large data dominating the model. Expected state 4 in Figure 4.9 shows a normal rim width and VF sensitivity (similar values to the control in Figure 4.8 with high NFB sensitivity and low rim-associated variables), whereas state 1 shows marked diffuse rim narrowing (high values for the rim-associated variables), and moderate loss of retinal sensitivity (low values for the NFB sensitivities) similar to the glaucomatous in Figure 4.7.

This is what would be expected, based on known anatomical relationships. State 3 (an apparently transitory state) displays narrowing of the rim but little reduction of retinal sensitivity, whilst state 2 (a relatively stable state) shows some narrowing of rim, but no loss of retinal sensitivity. This characteristic progression in the field but not in the optic disc (as displayed in the HRT rim data) and vice versa is known to occur and it is interesting that these have been identified by the algorithm as precursors to full disease progression. More informative clusters are discovered using our approach. Furthermore the pseudo-temporal models will further assist in understanding the progression of Glaucoma.

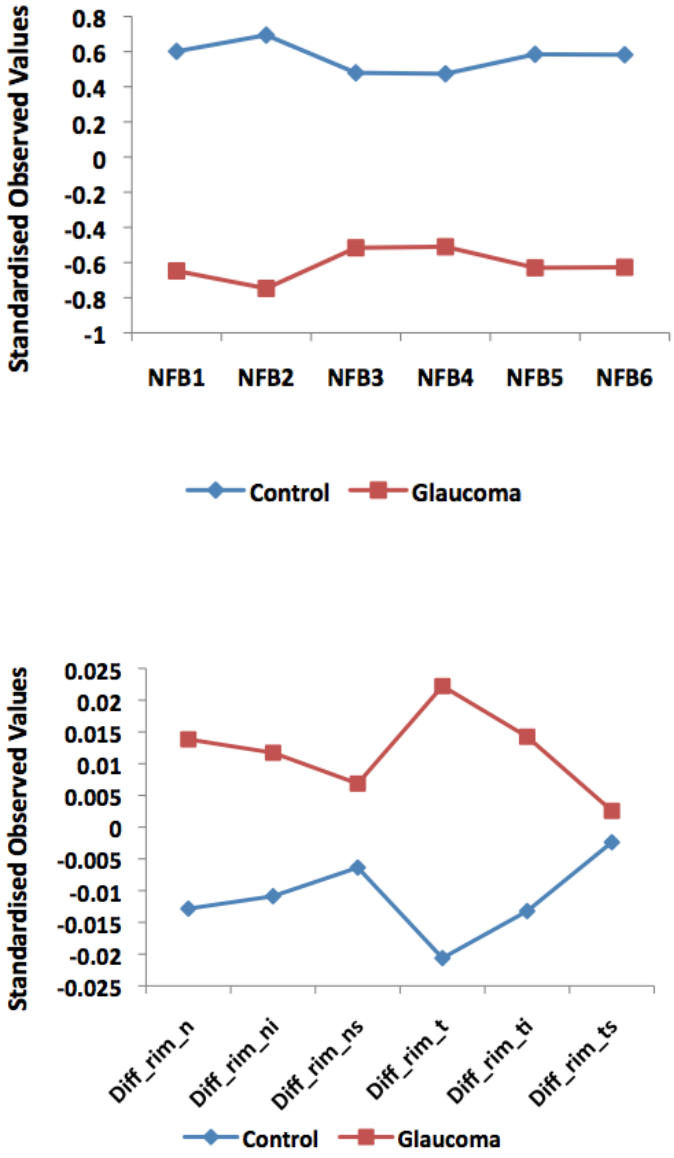

Figure 4.8: The mean data for the VF (top) and HRT (bottom) data as pre-classified using clinical analysis. NFB represent the sensitivity of a specifier Never Fibre Bundle with the VF, and Diff_rim represents the rime narrowing regions.

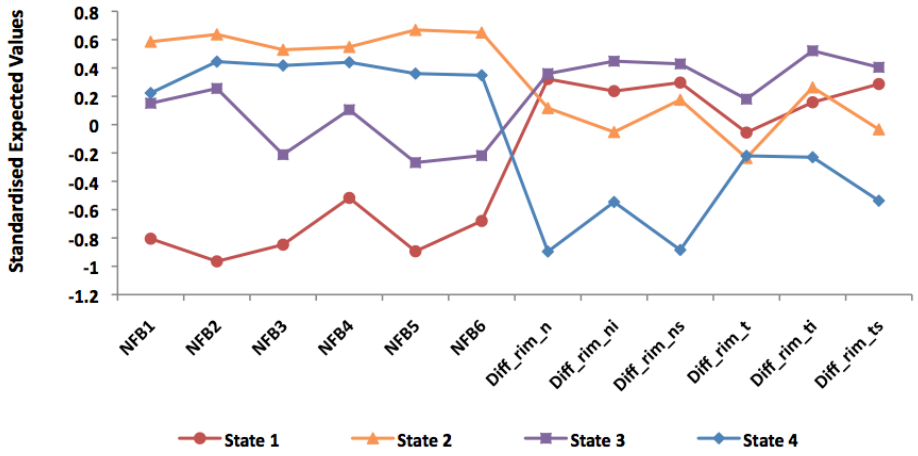

Figure 4.9: The expected data for VF and HRT discovered using Algorithm 2.

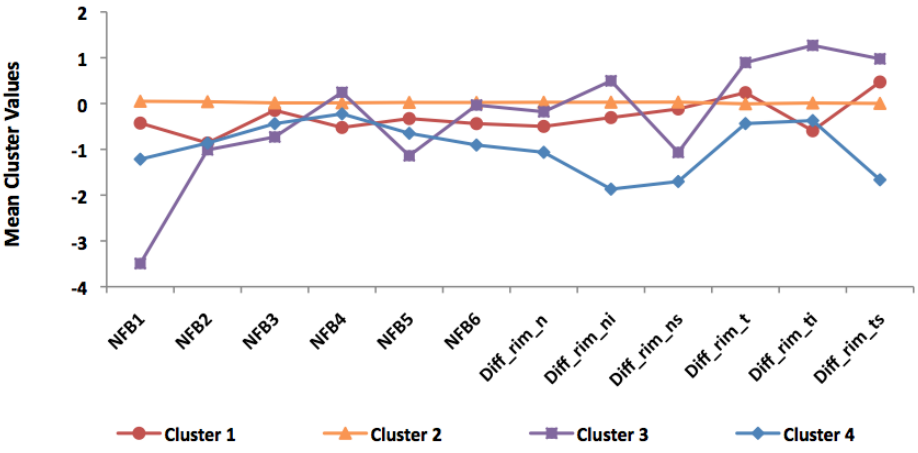

Figure 4.10: The mean cluster profiles for VF and HRT discovered using k-means clustering.

4.4.2 Breast Cancer

4.4.2.1 The Trajectories - Breast Cancer

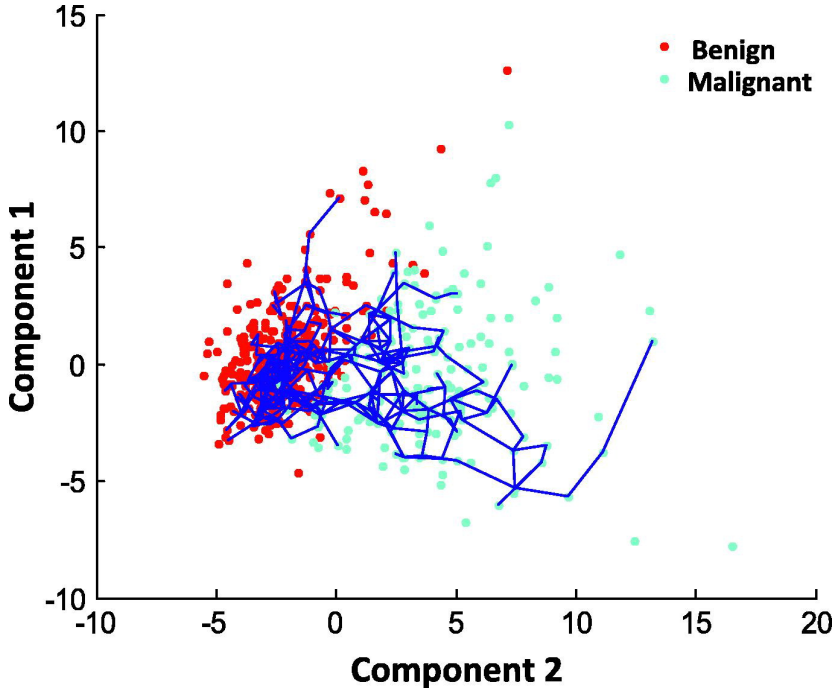

Figure 4.11: Typical trajectories learnt from the BC data. Benign are marked in red and Malignant in blue.

Figure 4.11 shows sample trajectories for the BC data. There is a clear cluster of benign tumours in red and those classified as malignant in blue. We use the relabelling scheme to see if we can identify the state transitions from healthy/benign to malignant.

4.4.2.2 End-State analysis - Breast Cancer

The BC state transition diagram generated from the transition matrix of the learnt H-MM is illustrated in Figure 4.12 (the figure shows transitions with a $p > 0.15$ as solid lines and $p > 0.05$ as dashed). The full transition matrix is given in Table 4.7. This table and diagram show that there appear to be four relatively stable states: 2, 3, 4 and

5. State 3 appears to coincide with the starting benign state, whilst state 2 represents a relatively stable malignant state. States 1, 4 and 5 appear to be transitory states, with state 5 being a key stage in the progression to advanced malignant tumour. The states are further explored by comparing the mean data for the pre-classified benign and malignant states (Figure 4.13) with the expected values of the variables, associated with each discovered state as shown in Figure 4.14. The clustering values of the variables discovered for the BC data using k-means clustering, as shown in Figure 4.15.

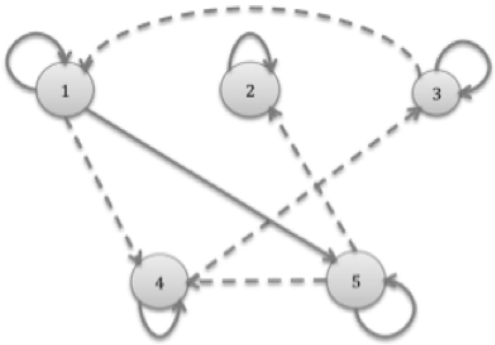

Figure 4.12: State transitions for the BC data. State 3 appears to coincide with the starting benign state, whilst 2 appears to represent a relatively stable malignant state, and 1, 4 and 5 to be transitory states, with state 5 being a key stage in the progression to advanced malignant tumour (a $p > 0.15$ is shown as a solid line and $p > 0.05$ as dashed).

$H_{t-1} \setminus h_t$	1	2	3	4	5
1	0.6963	0.0000	0.0403	0.0541	0.2093
2	0.0000	1.0000	0.0000	0.0000	0.0000
3	0.0959	0.0000	0.8929	0.0112	0.0000
4	0.0417	0.0377	0.1021	0.7764	0.0422
5	0.0000	0.1464	0.0000	0.0629	0.7907

Table 4.7: Transition matrix for discovered BC states.

State 3 does indeed seem to represent people with benign tumours (with generally low values for all metrics as seen in the benign cases in Figure 4.13), whereas state 2 represents cases of malignant BC (with generally high values as seen in the malignant cases in Figure 4.13). Interestingly, state 5, which looks like an intermediate stage in a trajectory will ultimately be in a state characterized as a malignant tumour. This shows characteristics of a malignant tumour but only in some variables such as high values for radius, perimeter and concave points. Other variables such as fractal complexity appear normal.

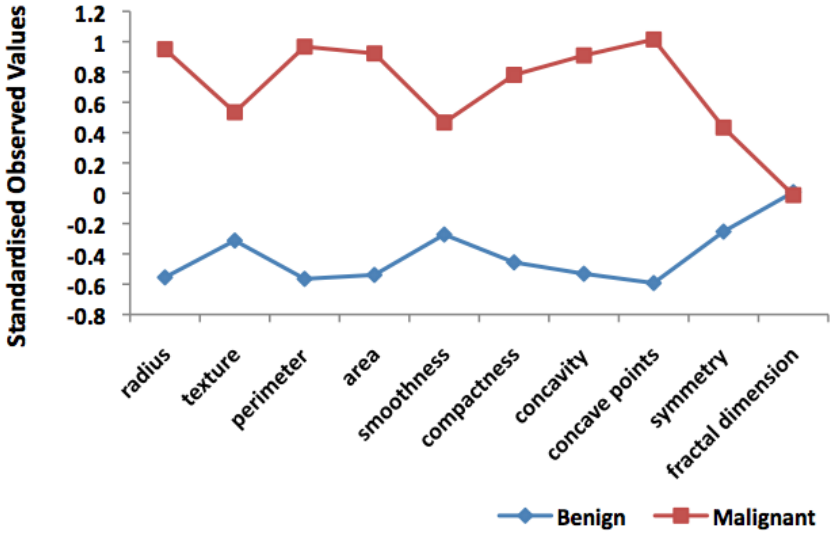

Figure 4.13: The mean data values for the pre-classified benign and malignant cases.

Here, the expected values from our approach are not really more informative than k-Means. However, our approach still offers the advantage of building transition models which help to understand the progression of the disease.

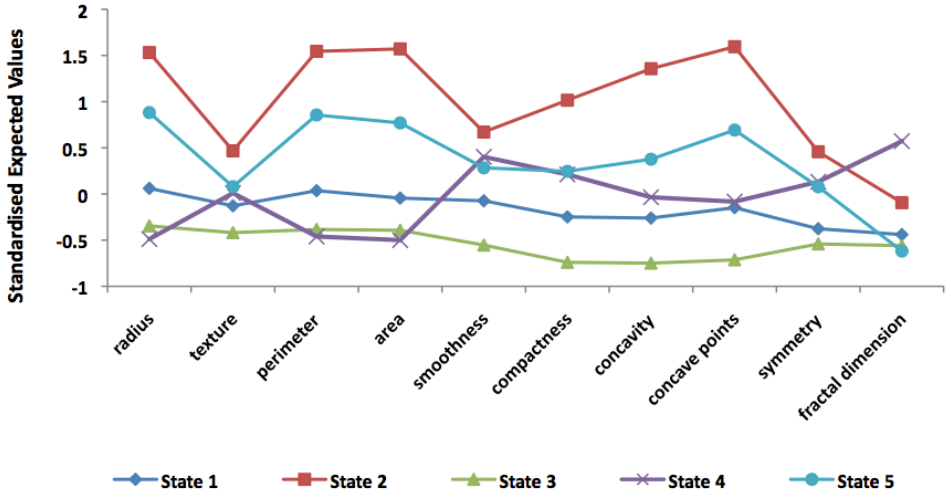

Figure 4.14: The expected values of data for each state discovered from the relabelling scheme on the BC data.

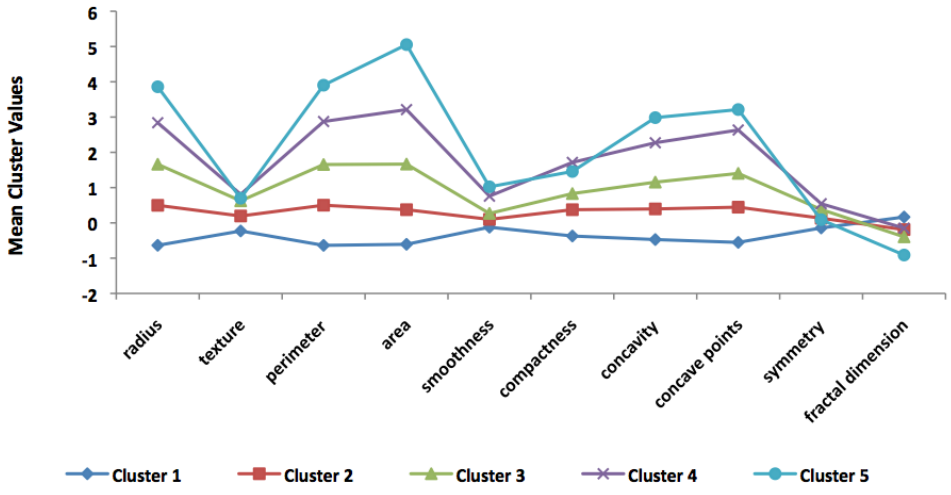

Figure 4.15: The mean cluster profiles for the BC data using k-Means clustering.

4.4.3 Parkinson's Disease

4.4.3.1 The Trajectories - Parkinson's Disease

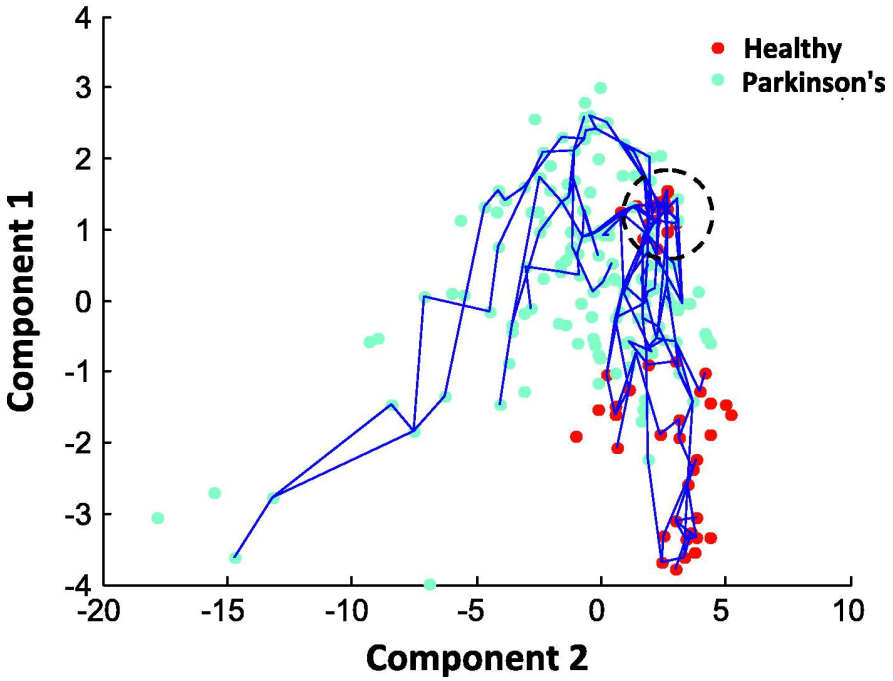

Figure 4.16: Typical trajectories learnt from the PD data. Healthy are marked in red and Parkinsonism in blue.

Figure 4.16 shows a curved trajectory from healthy to Parkinsonism. Note that the small cluster of healthy patients that appear to sit halfway through this trajectory after many other patients have been classified as sufferers (marked with a dashed circle). This is due to demonstrating a subset of symptoms and we explore this hypothesis by using the relabelling scheme.

4.4.3.2 End-State analysis - Parkinson's Disease

The PD state transition diagram generated from the transition matrix of the learnt HM-M is shown in Figure 4.17 (The figure shows transitions with a $p > 0.15$ as solid lines

and $p > 0.05$ as dashed).The full transition matrix is given in Table 4.8. This table and diagram show that there appear to be two relatively stable states: 1 and 2. State 1 coincides with the starting healthy state and 2 appears to represent a stable end state with 3 being a transitory state.

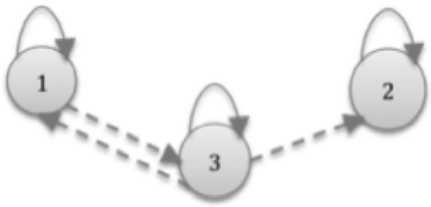

Figure 4.17: State transitions for the PD data. State 1 coincides with the starting healthy state, state 2 appears to represent a stable end state, with 3 representing a transitory state (a $p > 0.15$ is shown as a solid line and $p > 0.05$ as dashed).

$H_{t-1} \setminus h_t$	1	2	3
1	0.8860	0.0000	0.1140
2	0.0000	1.0000	0.0000
3	0.0771	0.0482	0.8748

Table 4.8: Transition matrix for discovered VF states.

The expected values of the variables associated with each state are shown in Figure 4.19, the clustering values using k-Means is shown in Figure 4.20 and the mean values for the pre-classified Parkinsonism patients and controls are shown in Figure 4.18. State 1 in Figure 4.19 shows people with a healthy profile very similar to the control profile in Figure 4.17. State 2 appears to resemble people with PD, especially in the measurements of HNR (higher) and the three MDVP-related metrics (higher). HNR represents Harmonics to Noise Ratio and is one of the most popular approaches

to measuring voice function [PT07]. The three MDVP metrics capture the average, maximum and minimum vocal fundamental frequency. In addition, DFA is one of the parameters currently used to distinguish healthy people and people from PD sufferers, since the scaling exponent of DFA is larger in PD than healthy people. It is positive to see that the scale of this for state 2 is larger than the state 1 and 3. PEE is a new measure of PD dysphonia, F0 is the natural pitch of healthy voices. State 2 seems to be characterised by a much larger vibrato and micro tremor than State 1. Jitter is another feature of PD and control measure values should be close to 0, as we see in the State 1. State3 appears to be a transition state somewhere between the control and PD stages, where certain key features seem to show characteristics of PD (such as RPDE and DFA), whilst others resemble those of the controls. This state could account for the cluster of apparently healthy individuals that look as if they are part way along the trajectory in Figure 4.16. The expected results using our approach have more informative clusters than static clustering of the original cross-sectional data, due to the explicit modeling of the dynamics of the disease.

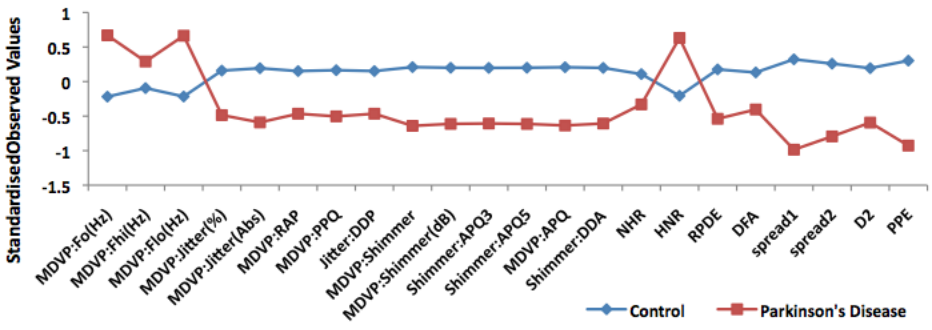

Figure 4.18: Mean PD data for pre-classified control and Parkinsons Disease.

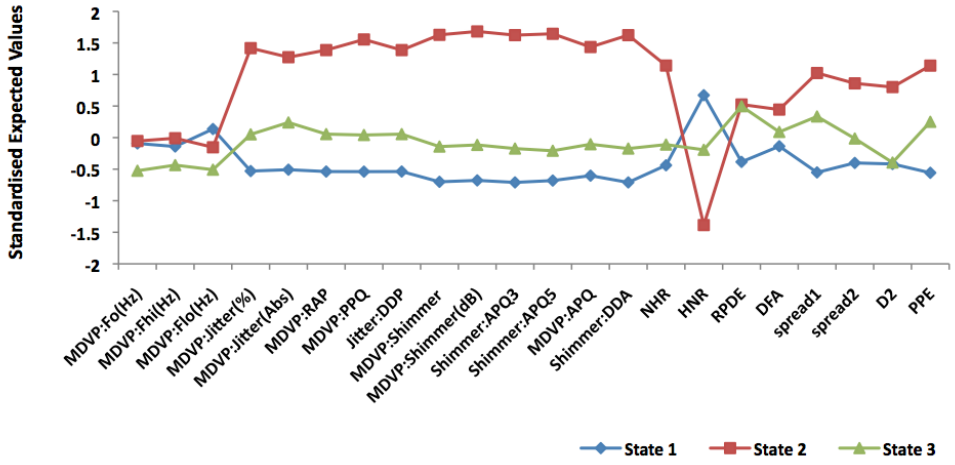

Figure 4.19: The expected data for each discovered state in the Parkinsonism Data.

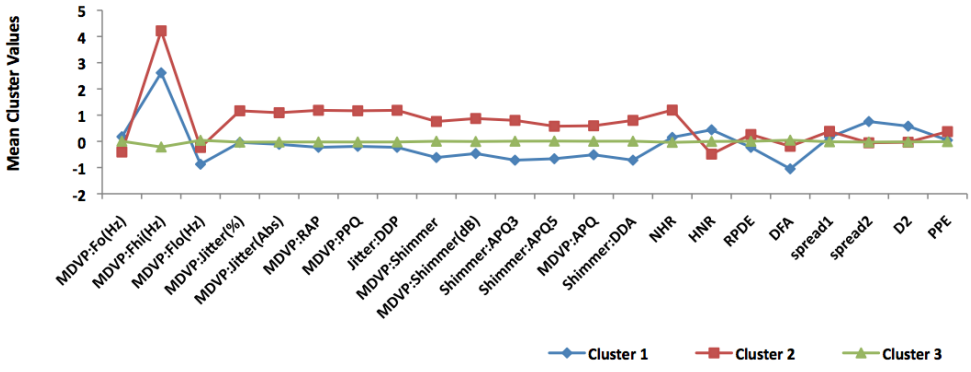

Figure 4.20: Mean cluster profiles using k-Means.

4.5 Summary

The results of the experiments utilising the approaches and techniques are discussed in this chapter. For the real world clinical data, the true underlying state transitions along a disease process are unknown. However, the discovered transitions can be explored in the medical context. This chapter has empirically demonstrated the advantages of the relabelling algorithm when applied to pseudo-time-series. The details are given to show how key intermediate stages of disease can be identified and evaluated using simulated data with complex multiple disease stages endpoints, as well as real clinical data from three very different diseases.

The next chapter will address how to overcome some of the current weaknesses of the pseudo time-series approach, namely the lack of genuine temporal information as the trajectories are only built as sequences through the data without any time stamps. It will explore how pseudo time-series can be calibrated using a small number of longitudinal datasets in order to inject real time stamps into the models of disease progression.

Chapter 5

Calibrating Pseudo Time-Series

In chapter 3, a resampling approach known as the Pseudo Temporal Bootstrap (TBS) [TGH10] was introduced. It aims to build multiple trajectories through cross sectional data, in order to approximate genuine longitudinal data. These Pseudo Time-Series (PTS) can then be used to build approximate temporal models for prediction. This approach has been extended to cluster important stages in disease progression using hidden Markov models and the EM algorithm [B$^+$98] as discussed in Section 3.4.1. However, the use of cross-sectional data to build these models will always be limited by the fact that no genuine time stamps have been used to infer the models. This chapter investigates the effect of incorporating genuine longitudinal data into the pseudo temporal models in order to calibrate them. This part of the work explores how to best balance combining the cross-sectional data with longitudinal data in order to minimise the need for too many expensive longitudinal data samples whilst being able to learn genuine temporal models.

5.1 Integrating data

In this chapter, we discuss the integration of cross-sectional and longitudinal data. The process of data integration normally involves combining data from different sources and providing users with a consolidated data view. Many data integration techniques address representation heterogeneity where similar data is stored in many different forms, as commonly seen in bioinformatics data [AMMM07]. Data Warehousing [Inm96] is key to many data integration projects (though still rare in biology and medicine) as it involves restructuring multiple databases in order to allow rapid access for analysis and data-mining through multidimensional modelling [CNF$^+$07]. Meta Analysis is also popular for data integration, particularly in ecological research where data can be expensive to obtain [CGGW05]. It works by supplying a statistical framework for identifying significant results over a number of independent published studies, and calculating the significance of all of the studies when they are brought together. It can be prone to publication bias where positive results are more likely to be published and therefore skewing the statistics [JM02]. Here we are combining data that contains the same essential variables (the clinical tests) the difference being that some are generated over time from the same patient and others are only recorded once per patient. Both types of data offer essential information, for example, disease variation in the general population are normally from cross-sectional studies while genuine temporal information of the disease from longitudinal data.

5.2 Experiments and Results

For this study, we explore the effect of adding relatively small numbers of time series to pseudo time-series generated from cross-sectional data, to see if the resulting models can be improved. Essentially it is to see if the limitations of pseudo time-series can be overcome (due to the lack of time-element in the trajectories which are in reality sequences of data) by calibrating them with real time-series. We explore this calibration on both simulated and real VF data.

5.2.1 Calibrating PTS on Simulated Data

For the simulated data, we generate time-series of length 30 from an autoregressive hidden Markov model (ARHMM (original) in Figure 5.1) to mimic typical biomedical longitudinal data (MTS). We then randomly sample a single point from these series to mimic the cross-sectional sampling of a population (CS DATA) but reserve 50 for the calibration.

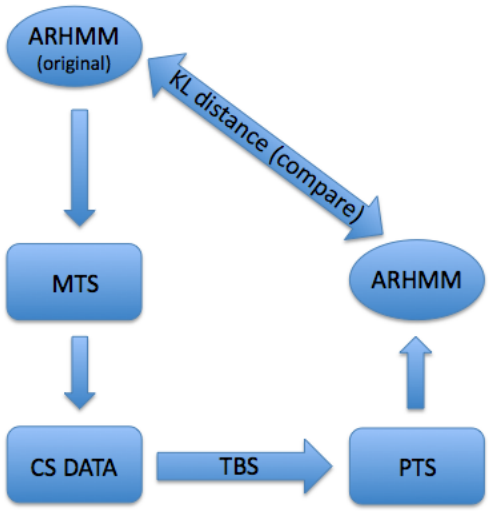

Figure 5.1: The scheme of the experiments: non calibrated.

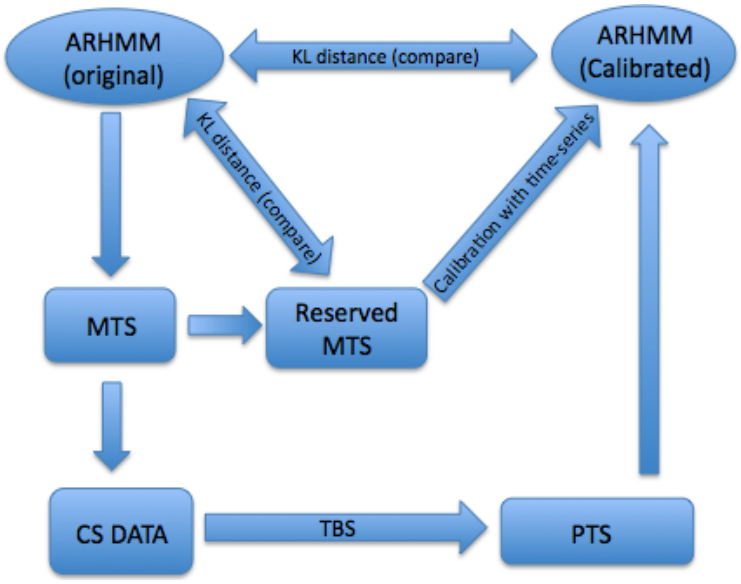

Figure 5.2: The scheme of the experiments: calibrated.

The experiment starts with 500 cross-sectional samples, as this was found to be a suitably large size to generate good pseudo time-series and models in [TGH10], and increment by 100 up to 1500 (the size of some increasingly large biomedical studies such as [GHLB$^+$12]). The Kulbaeck Leibler (KL) distance [Kul87] is used to explore how close a model learnt from the cross-sectional data using the temporal bootstrap is to the original generating model. Figure 5.1 illustrates the general scheme.

A number of the reserved original time-series (Reserved MTS) generated by the same ARHMM were then added to the pseudo time-series, and calibrated models are learnt from this (Callibrated ARHMM in Figure 5.2). We explore how close these new calibrated models are to the original model. Increments of 10 time-series were used as these seem to differentiate between the KL distances significantly. We also include how good the model is when learnt solely from the time-series used to calibrate the models to check that it is not simply using these small longitudinal samples to parameterise

(see Figure 5.2)

A similar experiment was performed using real VF longitudinal of 91 patient time-series. The data taken from a study of patients with Ocular Hypertension (often a precursor to Glaucoma). Each patient undergoes VF examination on a 4 monthly basis, hence the obtained time-series can be used to explore progression of the disease. The full study is documented in [KGHR$^+$03] and 91 patients were selected in this study based upon a minimum number of visits. One VF test is sampled from each of these selected patient's time-series to generate a cross-sectional sample and generate PTS data for learning models from. We then compare these models, as well as those learnt from a combination of PTS and real time-series to see how close the resulting models are to the one that is generated from the original longitudinal study. Figure 5.3 shows the results for all experiments, including learning PTS from cross-sectional samples of varying sizes and either not calibrating, calibrating with 10 time-series, or calibrating with 20 time-series.

The first obvious characteristic of these graphs is that calibrating does indeed improve the quality of the models with the KL distances that are closer to the original generating ARHMM. This is not surprising seeing that there is no genuine "time" in the PTS generated from the cross-sectional data. What is surprising, is that only a relatively small number of time-series are needed to improve these models, especially when there are lots of samples used from the cross-sectional data. This well supports the results from previous studies (Chapter 4) that the PTS does find good-but-not-perfect models (limited by the lack of real time-series) and that a small number

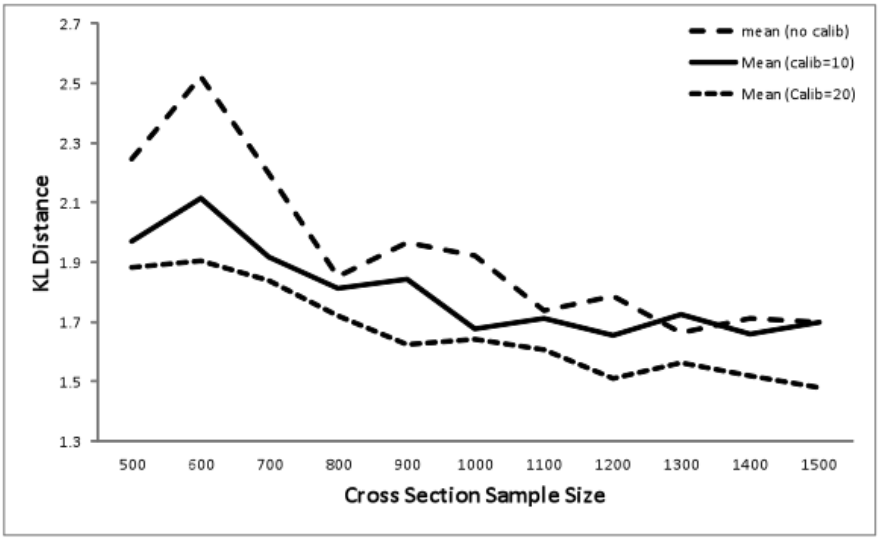

Figure 5.3: KL distance for varying cross-sectional study sample sizes with increasing number of longitudinal data for calibration.

of genuine time-series can calibrate these models. These findings provides good supporting evidence that expensive longitudinal studies can be relatively small in size if combined with larger cross-sectional studies to capture the general trajectories and the variability of disease progression within a population.

Figure 5.4 shows the confidence intervals of the KL distances generated from the study. From the one with no calibration and with calibration from 10 time-series, it can be seen that there is a steady decrease in KL distance as cross-sectional sample size increases where more and more reliable PTS are constructed. When the sample size is 1500, the KL distance mean becomes 1.70 +- 0.16. Note that when 10 time-series alone are used to learn the model, a mean KL distance of 2.08 +- 0.26 is obtained. This shows that the PTS generated from the cross-sectional data improves on models learnt from the time-series only by incorporating the variability within a larger population captured in the cross-sectional data. With calibration from 20 time-series, we see a

similar story, where increasing the cross-sectional sample size, builds better PTS and results in models that are closer to the original. For 1500 in the cross-sectional sample we see a KL distance of 1.48 +- 0.12. Note that when 20 time-series alone are used to learn the model we get a mean KL distance of 1.78 +- 0.15. Again, it proves that the PTS improves on time-series alone, but that the integration of both seems to generate the models that best reflect the underlying model.

We also explored the statistical significance of the differences between these KL distances using the Wilcoxon Rank comparison. Table 5.1 shows the Wicoxon Rank statistics comparing the KL distance between different models learnt using the different approaches. An asterisk is used to denote significant p values. First of all we notice that there are many significant values implying that the difference between models learnt using the two different approaches are significant. The most important statistics are those that show the models learnt with no calibration and only 500 cross-sectional data points are significantly different from most other models (row 1), but when 1500 cross-sectional data points are used the resulting model becomes much closer, only being significantly different from the model learnt from 50 full time-series (row 4). However, by calibrating these models we see a little improvement for 500 cross-sectional (CS) data points but for 1500 datapoints calibrated with 20 time-series, there is no significant difference between the models learnt from the full time-series. This implies that when the CS sample size is large enough and the resulting PTS models are calibrated with a relatively small number of real time-series, then a model can be learnt that is as good as one learnt from all time-series data.

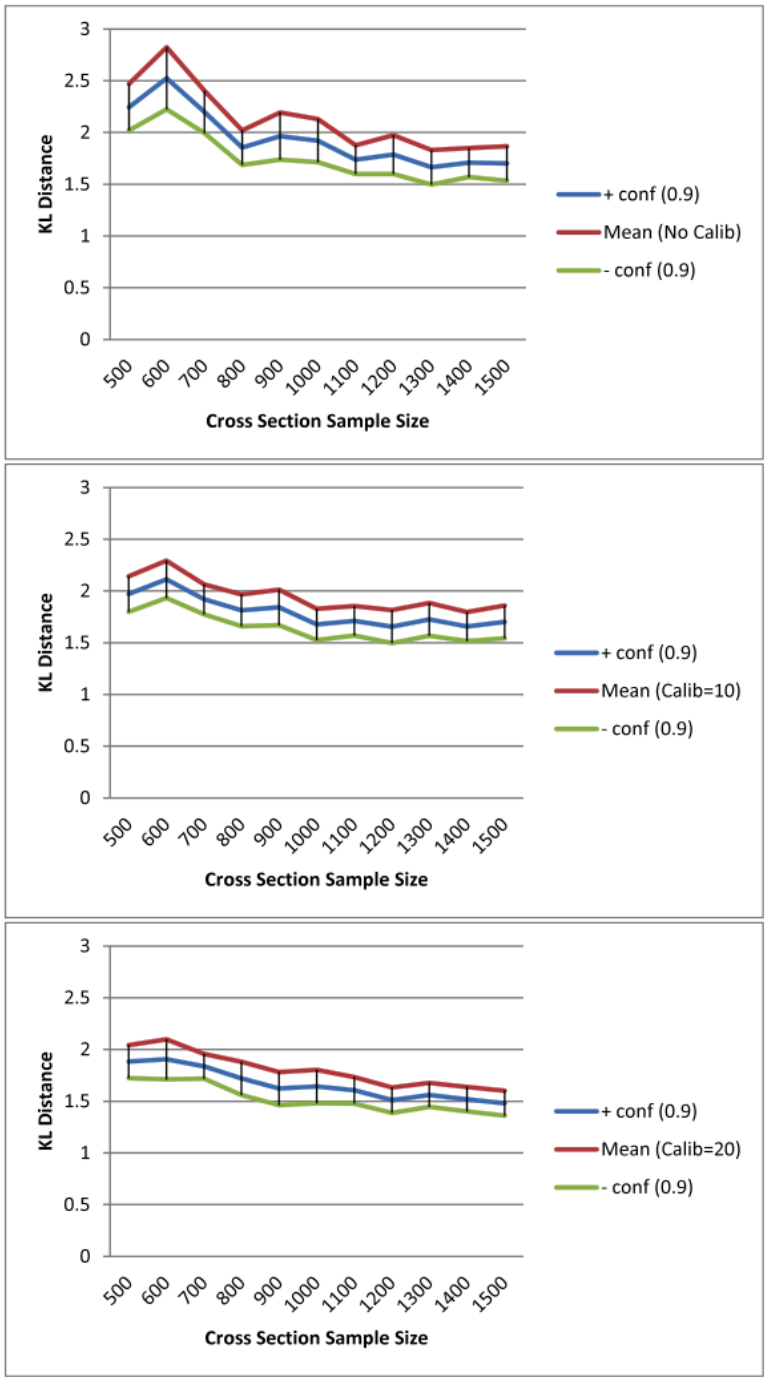

Figure 5.4: Confidence Intervals for the KL Distance to the original model generating the data for increasing sample sizes of cross-sectional data. i) with the non-calibrated model (top) ii) with the model calibrated with 10 time-series (middle) and iii) with the model calibrated with 20 time-series (bottom).

Wilcoxon Rank	cs500calib10	cs500calib20	cs1500nocalib	cs1500calib10	cs1500calib20	csfull30	csfull50
cs500nocalib	0.196	**0.047***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
cs500calib10	-	0.455	**0.062***	**0.036***	**0.000***	**0.010***	**0.000***
cs500calib20	-	-	**0.077***	0.130	**0.001***	**0.023***	**0.001***
cs1500nocalib	-	-	-	0.947	0.119	0.395	**0.064***
cs1500calib10	-	-	-	-	**0.052***	0.277	**0.047***
cs1500calib20	-	-	-	-	-	0.395	0.728
csfull30	-	-	-	-	-	-	0.291
csfull50	-	-	-	-	-	-	-

Table 5.1: Wilcoxon Rank Comparison between KL distances to original (significant p values are marked with an asterisk).

5.2.2 Calibrating PTS on Real Visual-Field Data

We now explore the effect of calibrating PTS using the real Visual Field time-series data described earlier. As there is no knowledge of the true underlying model, we firstly compare the KL distance between models that are repeatedly learnt from the original 91 patient time-series (91 MTS VF DATA in Figure 5.5) in order to get an idea of the general variance between models (MEAN VARIANCE) and to use this as a base-line. Essentially, if a model can be generated using PTS approaches with a KL distance that is not significantly greater than the general variance between different builds of the model on the full data, then we can be confident that the PTS models are of a suitably similar quality to those learnt from the full time-series.

Based on the consideration alone, the KL distance is calculated between a model learnt from the sampled cross-section using the PTS approach (PTS on 91 SAMPLED CS) and models learnt from the original 91 time-series. We then incrementally add a number of randomly selected real time-series (RANDOM 10/20 MTS) to calibrate the PTS model to see if this improves the KL distance.

Finally the KL distance is calculated between learning models using only the calibrated time-series to confirm that the PTS are indeed improving the resulting models.

The experiments are repeated 100 times to derive confidence intervals on the KL distances. Figure 5.5 shows the overall scheme for all experiments and Figure 5.6 shows the results.

One observation is that the KL distance between models that have been learnt on the full 91 time-series are in the region of 50 with a small confidence interval denoting a relatively small variance from one model learning to the next. The models that are

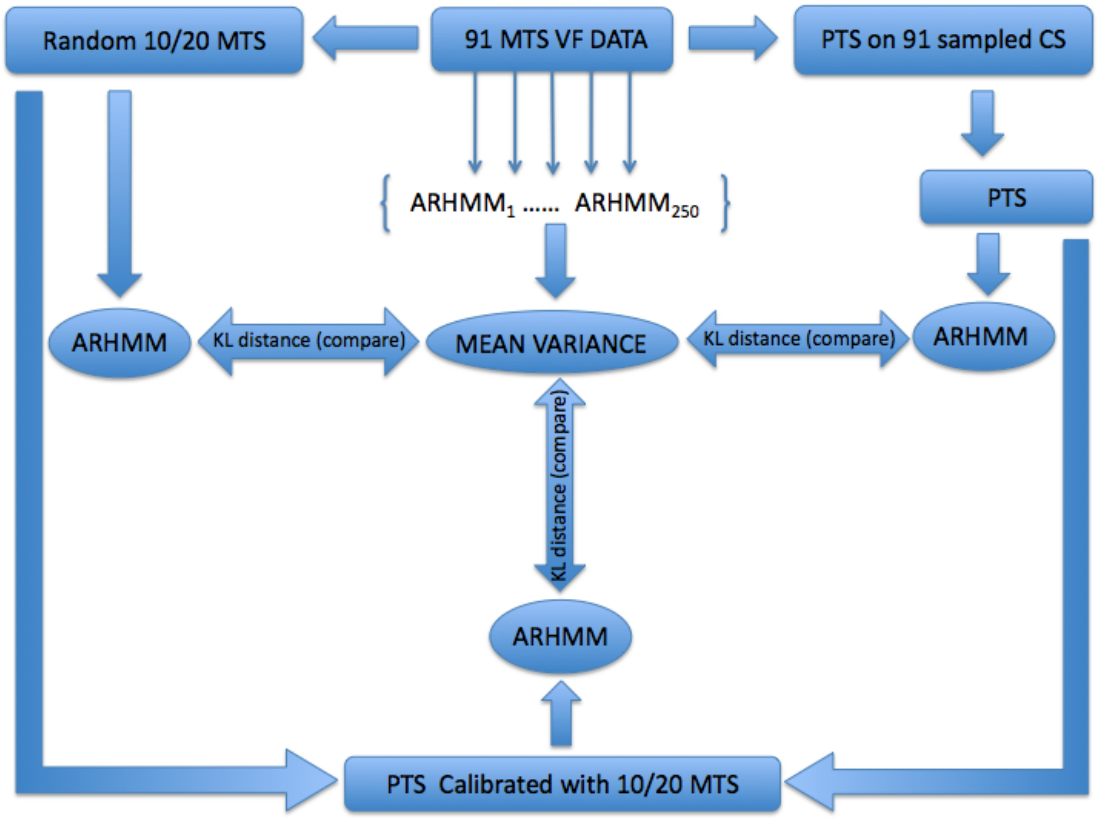

Figure 5.5: The overall scheme for all calibration experiments.

learnt from the sampled cross-section using the PTS approach are impressively close to the time-series models but distinctly higher in KL distance - approximate distance of 120 - (likely to be because we are lacking real temporal information). When 10 and 20 real time-series are used to calibrate the model, however, we see further improvement in the KL distance resulting in models that are demonstrably closer to the models learnt from all 91 time-series (mean distances of 100 and 80 for the models calibrated with 10 and 20 series respectively). Finally, models that are learnt from using the relatively small number of calibrating time-series only are clearly worse with much higher distance and very large confidence intervals (mean distances of 290 and 210 for

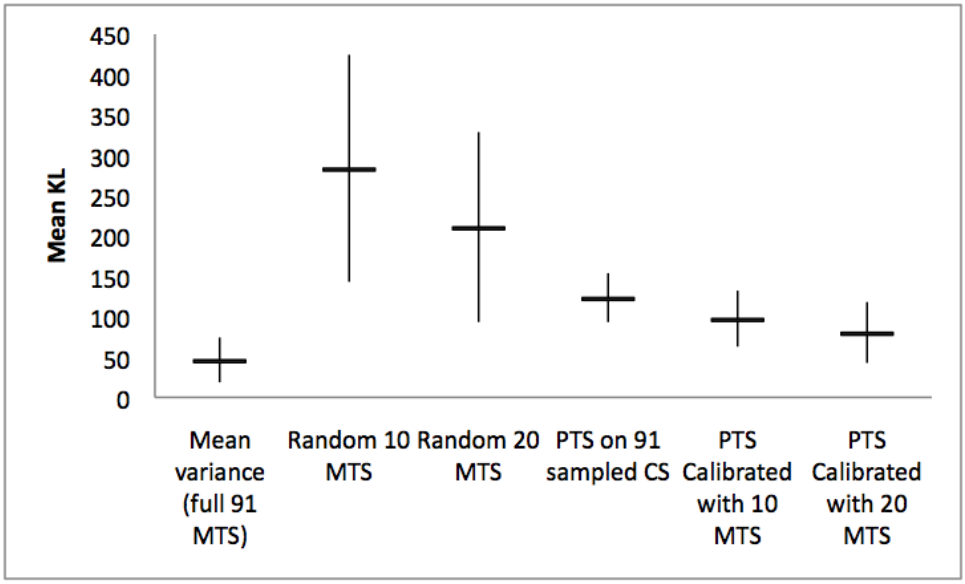

Figure 5.6: KL results for VF data with confidence intervals.

the 10 series and 20 series respectively).

To find out if these distances between models are significantly different, the Wilcoxon rank test is used again. Table 5.2 shows the result of applying this test to all combinations of models from Figure 5.4. An important thing to notice here is that nearly all of the models are indeed significantly worse than the variation between models learnt on the full longitudinal dataset (significant differences are marked with an asterisk) except for the PTS model calibrated with 20 real time-series. This shows that we can learn models that are as good as the natural variation between model building on the full longitudinal dataset by building PTS and calibrating with only 20 real longitudinal samples. It can also be seen that many of the inferior models are similar in terms of their distances except for the very worst models (learnt from only 10 time-series) which are different from the superior models which are both PTS models that have been calibrated.

	Mean variance	Rand 10	Rand 20	PTS	PTS Cal(10)	PTS Cal(20)
Mean variance (full 91 MTS)	-	**0.000***	**0.001***	**0.001***	**0.005***	0.011
Random 10 MTS	-	-	0.975	0.023	**0.002***	**0.001***
Random 20 MTS	-	-	-	0.042	0.014	0.010
PTS on 91 sampled CS	-	-	-	-	0.452	0.327
PTS Calibrated with 10 MTS	-	-	-	-	-	0.773
PTS Calibrated with 20 MTS	-	-	-	-	-	-

Table 5.2: Wilcoxon Rank significance.

5.3 Summary

In this chapter, we explored combining cross-sectional and longitudinal studies to build more robust models by simply aggregating PTS with real time-series. Although the PT-S approach alone does indeed learn very good models, by adding a small number of real time-series it is possible to get models that are considerably closer to the models learnt using all the time-series data that is available. We have shown that this is the case with significance on both the simulated data (where we know the true underlying model) and on VF data where we compare the distances to the general variation between models that are learnt from different repeats of the same model-building process on the full longitudinal data.

Chapter 6

Conclusions and Future Work

This chapter draws together the conclusions reached based on the research presented in this thesis. First, the main contributions are summarised, followed by a discussion of the limitations of the research presented. Finally, potential avenues for further research are presented, which are based on addressing the research limitations discussed in the previous section, and extending the applicability of the techniques presented in this thesis.

6.1 Conclusions

This thesis documented the application of intelligent data analysis techniques for extracting information from time series generated by different diseases. The results presented in this work relate to two major issues in this area of research: how clinical variables interact as a disease progresses along the trajectories in the data; and how to automatically identify different disease states along these trajectories, as well as the transitions between them. A combination of simulated data and three real-world biomedical cross-sectional datasets (Visual Field, Breast Cancer and Parkinson's Disease) were used to demonstrate and validate the new approaches introduced here. Being able to model trajectories and the temporal aspect of disease from these datasets is not trivial.

This thesis started with a review of previous and current research concerning the application of machine learning for biomedical data analysis in Chapter 2. Clinical trials were briefly discussed. The advantages and disadvantages of cross-sectional studies and longitudinal studies were also analysed. In the literature review, four typical models of classification and clustering were particularly examined in a biomedical context. The research in this thesis focused on the use of sequence-building through cross-sectional data (including trajectories with multiple endpoints).

The major work and achievements of this PhD thesis were presented in a logical sequence. Firstly, a formalisation of the pseudo time-series introduced by Tucker [TGH10] was explored. In addition it was investigated how cross-sectional and longitudinal data can both be better used to build more reliable models.

Different approaches to modelling trajectories through clinical data were identified

and hidden Markov models implemented given that they are effective at dealing with uncertainty and noise.

Secondly, an extension of the pseudo time-series approach, previously introduced in [LST12], involving the implementation of relabelling the hidden state to identify intermediate stages in the disease process. The extended approach allows us to identify the temporal nature of diseases, which is one of the major achievements of this study. Other similar studies have previously explored the progression of disease through latent variables, using longitudinal data. For example, studies that model non-stationary time-series using state-space models have been developed that simultaneously fit dynamic time-series model parameters whilst identifying changes in the underlying state ([TL04], [RH10]). However, none of these attempt to exploit the smoothness of disease progression to fit trajectories through cross-sectional data, which is much more abundant in clinical applications, in order to build time-series models and understand progression. This is what we present here. It is clear that many of the longstanding approaches as discussed in Chapter 3 to modelling disease progression are proving inadequate to dealing with issues of uncertainty in the dynamic and measurement processes and the ability to integrate cross-sectional studies with longitudinal studies.

Thirdly, as earlier, to demonstrate the effectiveness of the proposed approach, a number of real disease studies were utilised. This allowed any characteristics of the disease process that were discovered to be placed in a real medical context. Utilised real disease studies included: glaucoma using visual field test data, breast cancer using tumour image data and parkinson's disease using speech data and discussed in Chapter 3. The rather promising results based on the approaches and techniques mentioned

above are documented in Chapter 4. Regarding the *glaucoma data*, this study was able to identify:

1. stable states with abnormal VF sensitivity and marked rim narrowing;
2. transitory states with moderate narrowing of rim;
3. subtle loss of retinal sensitivity in the central macula.

Those results fit well with current knowledge of the progression of glaucoma of which initial symptoms can appear in the rim but not the visual field and vice versa.

By using the *cancer data*, the proposed approach also successfully identified:

1. stable states that reflect the benign and malignant tumour states;
2. an intermediate state that is characterised by a subset of the symptoms of malignant tumours.

With the *Parkinson's disease data*, a transitory state was discovered which has certain characteristics of Parkinsonism despite being pre-classified as symptoms of controls.

Chapter 4 empirically demonstrated the advantages of applying the proposed relabelling algorithm, to pseudo time-series. It highlights how key intermediate stages of disease can be identified and evaluated using simulated data (with complex multiple disease stages endpoints), and real clinical data from three very different diseases.

Finally, due to identified limitations of using cross-sectional data alone to build these models (the fact that no genuine time-stamps have been used to infer the models), in Chapter 5, we investigated the effect of incorporating genuine longitudinal data into

the pseudo temporal models in order to calibrate the models. This work particularly examined how to best balance the utilisation of cross-sectional and longitudinal data in order to minimise the cost of longitudinal data collection, a process which is quite expensive. This is essentially a matter of integrating cross-sectional and longitudinal data. The overall aim is to exploit the advantages of both types of studies - population diversity of cross-sectional data and temporal information in longitudinal data. We explored to what degree pseudo time-series models, learnt from building trajectories through a cross-sectional study, can be calibrated by a relatively small number of real time-series data form a clinical longitudinal study. The results show that almost all models are significantly different from the general variance when learning the model from the full 91 time-series. The only model that is not significantly different at the 1% level is the model that is learnt from the PTS and calibrated with 20 time-series.

6.2 Caveats and Future Work

There are a number of limitations in the research presented in this thesis. Firstly, the techniques presented have only been tested with the structure imposed by the HMM architecture. The utilised models are intended for stationary processes meaning that parameters and relations between variables are considered stable over time. However, extension of the techniques for the stationary HMMs has provided a solid foundation from which they can be further extended for modelling changing structure within a clinical time-series.

Secondly, when calibrating the pseudo time-series models, a simple process of concatenating the datasets was used. This may cause some difficulties in balancing

the influence of the pseudo time-series and the calibrating time-series. In other words, if there are too many pseudo time-series, the relatively smaller number of calibrating time-series may have little effect. Though this was not the case in our experiments, another more structured approach to combining them could be explored such as taking the pseudo time-series model as a prior model and updating it within a Bayesian framework using the calibrating data. That is, a suitable balance could be identified between the two sources of data

Additionally, for the simulated data, the number of variables was kept very small. Therefore, the applicability of the proposed techniques to a broader range of types of data and larger networks has not been considered. However, the utilisation of models such as dynamic Bayesian networks for data with more variables would be worth exploring.

As part of further work, the modelling techniques could be extended in a number of ways. As discussed in chapter 4, the techniques presented in this thesis have only been used with HMMs, although modelling temporal behaviour plays an important role in identifying key stages in a disease process. In particular, to improve the directionality of the learnt interactions, temporal information may be incorporated through time nodes and DBNs. As already mentioned in Section 3.4.1, a limitation of HMMs, concerns the fact that parameters and relations between variables cannot change over time. In many medical contexts, however, dependency relations between variables can change over time. For example in glaucoma, the Optic Nerve Head, which carries the visual functional signal, structurally changes during the progression of the disease, resulting in non-stationary series [YD03]. To overcome the stationarity in time series modelling

with graphical models, non-stationary DBNs have been recently introduced ([GH09], [RH10], [TH05]), which attempt to learn when the changes in structure occur.

Furthermore, Chapter 5 of this thesis explored the combination of data from cross-sectional and longitudinal studies to build more robust models of disease progression, by simply aggregating PTS with real time-series. However, some of the methods described in this thesis can be further extended that could be explored would be to take a Bayesian approach that uses informative priors [CS00]. By adopting a Bayesian approach using informative priors to integration, cross-sectional studies can be used to learn prior models [CS00]. This can be performed either directly on the data, resulting in static Bayesian networks, or via the pseudo time-series approach, described in this thesis, to produce DBNs models.

These priors can then be updated with the real time-series from longitudinal studies in order to 'calibrate' the temporal models. For example, The PTS models could be built and used as priors which then could be updated with the real time-series. In this way, we could control the influence of each type of data more carefully. If we have more reasons to trust the fidelity of the cross-sectional data more, we could bias the Bayesian updating process to the prior models, whereas if we want the longitudinal data to have more influence we could weaken the effect of the priors. This approach overcomes some of the issues concerning the models generated from the sequence reconstruction models such as the lack of genuine temporal information.

Integrating both longitudinal and cross-sectional data offers the advantage of modelling diverse populations, which incorporate samples of all stages of disease, whilst also encoding the genuine temporal characteristics of disease processes. The models

explored and extended in this thesis make important steps in this integration through the use of pseudo-temporal resampling and the produced results demonstrate their ability to identify important stages in disease progression.

Bibliography

[AFS+04] K.V. Allen, B.M. Frier, M.M.J Strachan, et al. The relationship between type 2 diabetes and cognitive dysfunction: longitudinal studies and their methodological limitations. *European journal of pharmacology*, 490(1):169–176, 2004.

[A.K96] McCallum A.K. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[Alb99] P.S. Albert. Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in medicine*, 18(13):1707–1732, 1999.

[AMMM07] R. Alfieri, I. Merelli, E. Mosca, and L. Milanesi. A data integration approach for cell cycle analysis oriented to model simulation in systems biology. *BMC systems biology*, 1(1):35, 2007.

[APMP07] M. Aristophanous, B.C. Penney, M.K. Martel, and C.A. Pelizzari. A gaussian mixture model for definition of lung tumor volumes in positron emission tomography. *Medical physics*, 34:4223, 2007.

[AYM+02] M.N. Ahmed, S.M. Yamany, N. Mohamed, A.A. Farag, and T. Moriarty. A modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data. *Medical Imaging, IEEE Transactions on*, 21(3):193–199, 2002.

[AZP06] M.P. Amato, V. Zipoli, and E. Portaccio. Multiple sclerosis-related cognitive changes: a review of cross-sectional and longitudinal studies. *Journal of the neurological sciences*, 245(1):41–46, 2006.

[B+98] J.A. Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

[BA06] L. Budagyan and R. Abagyan. Weighted quality estimates in machine learning. *Bioinformatics*, 22(21):2597–2603, 2006.

[Bag01] P.M. Baggenstoss. A modified baum-welch algorithm for hidden markov models with multiple observation spaces. *Speech and Audio Processing, IEEE Transactions on*, 9(4):411–416, 2001.

[Bar12] D. Barber. *Bayesian Reasoning and Machine Learning*. cambridge university press, 2012.

[BB01] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, 2 edition, 2001.

[BBdC+09] M.P. Basgalupp, R.C. Barros, A.C.P.L.F. de Carvalho, A.A. Freitas, and D.D. Ruiz, editors. *LEGAL-tree: a lexicographic multi-objective genetic algorithm for decision tree induction*, SAC '09. ACM, 2009.

[Ber03] B.P. Bergeron. *Bioinformatics Computing*. Prentice Hall Professional, illustrated edition, 2003.

[BGP99] P.H. Bourlas, E. Giakoumakis, and G. Papakonstantinou. A knowledge acquisition and management system for ecg diagnosis. *Machine Learning and Applications: Machine Learning in Medical Applications. Chania, Greece*, pages 27–29, 1999.

[BH05] A.F. Bakr and H.S. Habib. Combining pulse oximetry and clinical examination in screening for congenital heart disease. *Pediatric cardiology*, 26(6):832–835, 2005.

[BHK] G. Bontempi and B. Haibe-Kains. Feature selection methods for mining bioinformatics data.

[BJR13] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time series analysis: forecasting and control*. Wiley. com, 2013.

[BLM+99] R. Bellazzi, C. Larizza, P. Magni, S. Montani, and G. De Nicolao. Intelligent analysis of clinical time series by combining structural filtering and temporal abstractions. In *Artificial Intelligence in Medicine*, pages 261–270. Springer, 1999.

[Bor09] S. Borman. The expectation maximization algorithm-a short tutorial, 2009.

[Bou93] R.R. Bouckaert. *Probabilistic network construction using the minimum description length principle*. Springer, 1993.

[BQW+08] A.T. Broman, H.A. Quigley, S.K. West, J. Katz, B. Munoz, K. Bandeen-Roche, J.M. Tielsch, D.S. Friedman, J. Crowston, H.R. Taylor, et al. Estimating the rate of progressive visual field damage in those with open-angle glaucoma, from cross-sectional data. *Investigative ophthalmology & visual science*, 49(1):66–76, 2008.

[BSDP01] M. Behari, A.K. Srivastava, R.R. Das, and R.M. Pandey. Risk factors of parkinson's disease in indian patients. *Journal of the neurological sciences*, 190(1):49–55, 2001.

[BWG07] P. Bushel, R. Wolfinger, and G. Gibson. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Systems Biology*, 1(1):15, 2007.

[BZ08] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77:81–97, 2008.

[CCC05] R. Cédric, S. Chérif, and François C. A dynamic bayesian network for handling uncertainty in a decision support system adapted to the monitoring of patients treated by hemodialysis. In *Tools with Artificial Intelligence, 2005. ICTAI 05. 17th IEEE International Conference on*, pages 5–pp. IEEE, 2005.

[CGGW05] I.M. Côté, J.A. Gill, T.A. Gardner, and A.R. Watkinson. Measuring coral reef decline through meta-analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1454):385–395, 2005.

[CGHCT12] S. Ceccon, D. Garway-Heath, D. Crabb, and A. Tucker. Non-stationary clustering bayesian networks for glaucoma. *Proceedings of the Workshop on machine Learning for Clincial Data Analysis, ICML 2012*, 2012.

[CH92] G.F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

[Cha96] C. Chatfield. *The analysis of time series*. London: Chapman and Hall, 5th edition, 1996.

[CK02] A. Clare and R.D. King. Machine learning of functional class from phenotype data. *Bioinformatics*, 18(1):160–166, 2002.

[CLKW97] C. Chen, W. Lin, T. Kuo, and C. Wang. Adaptive control of arterial blood pressure with a learning controller based on multilayer neural networks. *Biomedical Engineering, IEEE Transactions on*, 44(7):601–609, 1997.

[CMJ98] L. Cao, A. Mees, and K. Judd. Dynamics from multivariate time series. *Physica D: Nonlinear Phenomena*, 121(1):75–88, 1998.

[CNF+07] J.B. Cushing, N. Nadkarni, M. Finch, A. Fiala, E. Murphy-Hill, L. Delcambre, and D. Maier. Component-based end-user database design for ecologists. *Journal of Intelligent Information Systems*, 29(1):7–24, 2007.

[Coo99] G. Cooper. An overview of the representation and discovery of causal relationships using bayesian networks. *Glymour and Cooper [36]*, pages 4–62, 1999.

[CS00] R. Castelo and A. Siebes. Priors on network structures. biasing the search for bayesian networks. *International Journal of Approximate Reasoning*, 24(1):39–57, 2000.

[CTC+05] T. Chen, T. Tsai, Y. Chen, C. Lin, R. Chen, S. Li, and H. Chen. A combined k-means and hierarchical clustering method for improving the clustering efficiency of microarray. In *Intelligent Signal Processing and Communication Systems, 2005. ISPACS 2005. Proceedings of 2005 International Symposium on*, pages 405–408. IEEE, 2005.

[CVDGV+09] T. Charitos, L.C. Van Der Gaag, S. Visscher, K.A.M. Schurink, and P.J.F. Lucas. A dynamic bayesian network for diagnosing ventilator-associated pneumonia in icu patients. *Expert Systems with Applications*, 36(2):1249–1258, 2009.

[CX04] Y. Chen and D. Xu. Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics*, 21(5):575–581, 2004.

[CYYK05] J.K. Choi, U. Yu, O.J. Yoo, and S. Kim. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 21(24):4348–4355, 2005.

[Dig02] P. Diggle. *Analysis of longitudinal data*, volume 25. Oxford University Press, USA, 2002.

[DRK04] T. Doom, M. Raymer, and D. Krane. Bioinformatics. *Journal of IEEE*, pages 24–27, 2004.

[DS10] J.N. Doctor and G. Strylewicz. Detecting wrong blood in tubeerrors: Evaluation of a bayesian network approach. *Artificial intelligence in medicine*, 50(2):75–82, 2010.

[Edd98] S.R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

[ESK02] L. Ertoz, M. Steinbach, and V. Kumar. A new shared nearest neighbor clustering algorithm and its applications. In *Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining*, pages 105–115, 2002.

[ET94] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1994.

[FEST01] M. Frize, C.M. Ennett, M. Stevenson, and H.C.E. Trigg. Clinical decision support systems for intensive care units: using artificial neural networks. *Medical engineering & physics*, 23(3):217–225, 2001.

[FGW99] N. Friedman, M. Goldszmidt, and A. Wyner. On the application of the bootstrap for computing confidence measures on features of induced bayesian networks. *AI&STAT VII*, 1999.

[Flo62] R.W. Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345, 1962.

[FM99] Y. Freund and L. Mason. The alternating decision tree learning algorithm. In *Proceeding of the Sixteenth International Conference on Machine Learning*, pages 124–133, 1999.

[FMR98] N. Friedman, K.P. Murphy, and S.J. Russell. Learning the structure of dynamic probabilistic networks. In *Proceedings of the 14th Annual Conference on Uncertainty in AI*, pages 139–147, 1998.

[FSN+00] J.P. Feighner, L. Sverdlov, G. Nicolau, J.F. Noble, et al. Cluster analysis of clinical data to identify subtypes within a study population following treatment with a new pentapeptide antidepressant. *The International Journal of Neuropsychopharmacology*, 3(3):237–242, 2000.

[Fu11] T. Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[FWIB95] D.B. Fogel, E.C. Wasson III, and E.M. Boughton. Evolving neural networks for detecting breast cancer. *Cancer letters*, 96(1):49–53, 1995.

[GDST+06] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006.

[Gei92] D. Geiger. An entropy-based learning algorithm of bayesian conditional trees. In *Proceedings of the Eighth international conference on Uncertainty in artificial intelligence*, pages 92–97. Morgan Kaufmann Publishers Inc., 1992.

[GH09] M. Grzegorczyk and D. Husmeier. Non-stationary continuous dynamic bayesian networks, 2009.

[Gha98] Z. Ghahramani. Learning dynamic bayesian networks. In *Adaptive processing of sequences and data structures*, pages 168–197. Springer, 1998.

[GHFH00] D. Garway-Heath, F. Fitzke, and R. Hitchings. Mapping the visual field to the optic disc. *British Journal of Ophthalmology*, 107(10):1809–1815, Oct 2000.

[GHLB+12] D.F. Garway-Heath, G. Lascaratos, C. Bunce, D. Crabb, R. Russell, and A. Shah. The united kingdom glaucoma treatment study: A multicenter, randomized, placebo-controlled clinical trial: Design and methodology. *Ophthalmology*, 2012.

[GP95] John G. and Langley P., editors. *Estimating Continuous Distributions in Bayesian Classifiers*. Morgan Kaufmann, 1995.

[HDA01] V. Hatzivassiloglou, P.A. Duboue, and Rzhetsky A. Disambiguating proteins, genes, and rna in text: a machine learning approach. *Bioinformatics*, 17(1):97–106, 2001.

[Hel13] IBM SPSS Modeler Help. The neural netwokrs model, 2013. http://pic.dhe.ibm.com/infocenter/spssmodl/v15r0m0/index.jsp?topic=

[HGC95] D. Heckerman, D. Geiger, and D.M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[HTO07] J. Huang, G. Tzeng, and C. Ong. Marketing segmentation using support vector clustering. *Expert systems with applications*, 32(2):313–317, 2007.

[HW79] J.A. Hartigan and M.A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[HXW+10] X. Hu, P. Xu, S. Wu, S. Asgari, and M. Bergsneider. A data mining framework for time series estimation. *Journal of biomedical informatics*, 43(2):190–199, 2010.

[Inm96] W.H. Inmon. *In Building the Data Warehouse*. John Wiley and Sons, 2nd edition, 1996.

[JAGRRJ+03] J. Jerez-Aragonés, J.A. Gómez-Ruiz, G. Ramos-Jiménez, J. Muñoz-Pérez, and E. Alba-Conejo. A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial intelligence in medicine*, 27(1):45–63, 2003.

[Jan99] N. Jankowski. Approximation and classification in medicine with incnet neural networks. In *Machine Learning and Applications. Workshop on Machine Learning in Medical Applications*, pages 53–58, 1999.

[JKD13] B. Jan, C. Kambhampati, and D.N. Davis. Alternating decision tree applied to risk assessment of heart failure patients. *Journal of Information Technologies*, 6(2):25–33, 2013.

[JLTL81] L. Janzon, S. Lindell, E. Trell, and P. Larme. Smoking habits and carboxyhaemoglobin. a cross-sectional study of an urban population of middle-aged men. *Journal of epidemiology and community health*, 35(4):271–273, 1981.

[JM02] M.D. Jennions and A.P. MOeLLER. Publication bias in ecology and evolution: an empirical assessment using the trim and fillmethod. *Biological Reviews*, 77(2):211–222, 2002.

[KCO02] P.P. Khil and R.D. Camerini-Otero. Over 1000 genes are involved in the dna damage response of escherichia coli. *Molecular microbiology*, 44(1):89–105, 2002.

[KFD+10] N. Klöting, M. Fasshauer, A. Dietrich, P. Kovacs, M.R. Schön, M. Kern, M. Stumvoll, and M. Blüher. Insulin-sensitive obesity. *American Journal of Physiology-Endocrinology And Metabolism*, 299(3):E506–E515, 2010.

[KGHR+03] D. Kamal, D. Garway-Heath, S. Ruben, F. O'Sullivan, C. Bunce, A. Viswanathan, W. Franks, and R. Hitchings. Results of the betaxolol versus placebo treatment trial in ocular hypertension. *Graefe's archive for clinical and experimental ophthalmology*, 241(3):196–203, 2003.

[KHK99] G. Karypis, E. Han, and V. Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.

[KIM03] S.Y. Kim, S. Imoto, and S. Miyano. Inferring gene networks from time series microarray data using dynamic bayesian networks. *Briefings in bioinformatics*, 4(3):228–235, 2003.

[Kiy11] T. Kiyan. Breast cancer diagnosis using statistical neural networks. *IU-Journal of Electrical & Electronics Engineering*, 4(2), 2011.

[KJV83] S. Kirkpatrick, D.G. Jr., and M.P. Vecchi. Optimization by simmulated annealing. *science*, 220(4598):671–680, 1983.

[KKG+99] Matja K., Igor K., Ciril G., Katarina K., and Jure F. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial Intelligence in Medicine*, 16(1):25 – 50, 1999.

[Kon01] I. Kononenlo. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23:89–109, 2001.

[KT02] J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. *Journal of the ACM (JACM)*, 49(5):616–639, 2002.

[Kul87] S. Kullback. Letter to the editor: The kullback-leibler distance, 1987.

[Kum13] M.N. Kumar. Alternating decision trees for early diagnosis of dengue fever. *arXiv preprint arXiv:1305.7331*, 2013.

[LAR03] A. Linden, J.L. Adams, and N. Roberts. Evaluating disease management program effectiveness: an introduction to time-series analysis. *Disease Management*, 6(4):243–255, 2003.

[LB94] W. Lam and F. Bacchus. Learning bayesian belief networks: An approach based on the mdl principle. *Computational intelligence*, 10(3):269–293, 1994.

[LCS+06] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J.A. Lozano, R. Armananzas, G. Santafe, A. Perez, and V. Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 2006.

[LFS+98] S. Liang, S. Fuhrman, R. Somogyi, et al. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific symposium on biocomputing*, volume 3, pages 18–29, 1998.

[LKBJ08] T. Lin, N. Kaminski, and Z. Bar-Joseph. Alignment and classification of time series gene expression in clinical studies. *Bioinformatics*, 24(13):i147–i155, 2008.

[LKZ00] Nada L., E. Keravnou, and B. Zupan. Intelligent data analysis in medicine, 2000.

[LLCJ00] Y. Li, L. Liu, W. Chiu, and W. Jian. Neural network modeling for surgical decisions on traumatic brain injury patients. *International journal of medical informatics*, 57(1):1–9, 2000.

[LMH+88] S. Lillioja, D.M. Mott, B.V. Howard, P.H. Bennett, H. Yki-Järvinen, D. Freymond, B.L. Nyomba, F. Zurlo, B. Swinburn, C. Bogardus, et al. Impaired glucose tolerance as a disorder of insulin action. longitudinal and cross-sectional studies in pima indians. *The New England journal of medicine*, 318(19):1217, 1988.

[LMR+07] M.A. Little, P.E. McSharry, S.J. Roberts, D.A.E. Costello, and I.M. Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMed Eng Online*, 6(23), Jun 2007.

[Log05] R. Loganantharaj. Predicting a transcription start site: case study with different genomes. In *Computational Systems Bioinformatics Conference, 2005. Workshops and Poster Abstracts. IEEE*, pages 199–200. IEEE, 2005.

[LPY+96] P. Larrañaga, M. Poza, Y. Yurramendi, R.H. Murga, and C.M.H. Kuijpers. Structure learning of bayesian networks by genetic algorithms: A performance analysis of control parameters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(9):912–926, 1996.

[LST12] Y. Li, S. Swift, and A. Tucker. Modelling and analysing the dynamics of disease progression from cross-sectional studies. *Journal of biomedical informatics*, 2012.

[LT06] P.J. Lisboa and A.F.G. Taktak. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, 19(4):408–415, 2006.

[LT10] Y. Li and A. Tucker. Uncovering disease regions using pseudo time-series trajectories on clinical trial data. In *Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference on*, volume 6, pages 2356–2362. IEEE, 2010.

[Man03] C.J. Mann. Observational research methods. research design ii: cohort, cross sectional, and case-control studies. *Emergency Medicine Journal*, 20(1):54–60, 2003.

[MBK98] R.S. Michalski, I. Bratko, and M. Kubat. *Machine learning, data mining and knowledge discovery: methods and applications*. New York: Wiley, 1998.

[Mit97] T.M. Mitchell. *MACHINE LEARNING*. Computer Science. McGRAW-Hill, international editions edition, 1997.

[MLP+04] N.B. Mohan, J. Lawrence, B. Patrick, M.B. James, and J.E. Mark. A hierarchical bayesian meta-analysis of randomised clinical trials of drug-eluting stents. *THE LANCET*, 364(9434):583–591, August 2004.

[MMM+00] R.G. Marvin, B.A. McKinley, M. McQuiggan, C.S. Cocanour, and F.A. Moore. Nonocclusive bowel necrosis occurring in critically ill trauma patients receiving enteral nutrition manifests no reliable clinical signs for early detection. *The American journal of surgery*, 179(1):7–12, 2000.

[MP01a] GeorgeD. Magoulas and Andriana Prentza. Machine learning in medical applications. In G. Paliouras, V. Karkaletsis, and C.D. Spyropoulos, editors, *Machine Learning and Its Applications*, volume 2049 of *Lecture Notes in Computer Science*, pages 300–307. Springer Berlin Heidelberg, 2001.

[MP01b] V. Mihajlovic and M. Petkovic. Dynamic bayesian networks: A state of the art, 2001.

[MSP05] A.M. Molinaro, R. Simon, and R.M. Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.

[Nag01] S.B. Nagl. Can correlated mutations in protein domain families be used for protein design? *Briefings in Bioinformatics*, 2(3):279–288, 2001.

[NHBM] D.J. Newman, S. Hettich, C.L. Blacke, and C.J. Merz. Uci repository of machine learning databases.

[Org] World Health Organization. International statistical classification of diseases and related health problems.

[Org98] World Health Organization. Parkinsons disease a unique survey launched., 1998. http://www.who.int/inf-pr-1998/en/pr98-71.html.

[Org10] World Health Organization. Programmes and projects: Cancer; world cancer day 2010: Quick cancer facts, 2010. http://www.lib.monash.edu.au/tutorials/citing/vancouver.html.

[OSP+92] S.C. Odewahn, E.B. Stockwell, R.L. Pennington, R.M. Humphreys, and W.A. Zumach. Automated star/galaxy discrimination with neural networks. In *Digitised Optical Sky Surveys*, pages 215–224. Springer, 1992.

[PA08] S. Palaniappan and R. Awang. Intelligent heart disease prediction system using data mining techniques. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, pages 108–115. IEEE, 2008.

[PdKJ+10] L. Peelen, N.F. de Keizer, E. de Jonge, R.J. Bosman, A. Abu-Hanna, and N. Peek. Using hierarchical dynamic bayesian networks to investigate dynamics of organ failure in patients in the intensive care unit. *Journal of biomedical informatics*, 43(2):273–286, 2010.

[PLV02] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[PSS+09] V.L. Patel, E.H. Shortliffe, M. Stefanelli, P. Szolovits, M.R. Berthold, R. Bellazzi, and A. Abu-Hanna. The coming of age of artificial intelligence in medicine. *Artificial Intelligence in Medicine*, 46:5–17, 2009.

[PT07] E. Peeling and A. Tucker. Making time: pseudo time-series for the temporal analysis of cross section data. In *Advances in Intelligent Data Analysis VII*, pages 184–194. Springer, 2007.

[PWM+99] Philip P.B., Marina W., J. M.O., George S., and William M.T. Pilot study of a point-of-use decision support tool for cancer clinical trials eligibility. *JAMIA*, 6:466–477, 1999.

[Rab89] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[RH10] J.W. Robinson and A.J. Hartemink. Learning non-stationary dynamic bayesian networks. *The Journal of Machine Learning Research*, 9999:3647–3680, 2010.

[RKM09] B.Y. Reis, I.S. Kohane, and K.D. Mandl. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. *BMJ*, 339, 9 2009.

[RM03] B.Y. Reis and K.D. Mandl. Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making*, 3(1):2, 2003.

[RPK+03] S.M. Resnick, D.L. Pham, M.A. Kraut, A.B. Zonderman, and C. Davatzikos. Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. *The Journal of Neuroscience*, 23(8):3295–3301, 2003.

[RRL99] R. Ruseckaite, G. Raškinis, and R. Lukauskiene. Computer interactive system for ascertainment of visual perception disorders. *Machine Learning and Applications: Machine Learning in Medical Applications*, pages 27–29, 1999.

[RSMJ85] H. Roland, H.J.V. Solke, M. Marcel, and H.C.L.H. Jan. Histologic multifocality of tis, t1-2 breast carcinomas implication for clinical trials of breast-conserving surgery. *Cancer*, 56:979–990, 1985.

[Sav12] N. Savage. Better medicine through machine learning. *ACM*, 55(1):17–19, 2012.

[SBR07] J.C. Slaboda, J.R. Boston, and T.E. Rudy. Using hmms to identify groups in a patient population: A simulation. In *Statistical Signal Processing, 2007. SSP'07. IEEE/SP 14th Workshop on*, pages 355–357. IEEE, 2007.

[Sch78] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[SG93] L. Satish and B.I. Gururaj. Partial discharge pattern classification using multilayer neural networks. *IEE Proceedings A (Science, Measurement and Technology)*, 140(4):323–330, 1993.

[SL02] S. Swift and X. Liu. Predicting glaucomatous visual field deterioration through short multivariate time series modelling. *Artificial Intelligence in Medicine*, 24(1):5–24, 2002.

[Slo02] D.K. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nature genetics*, 32:502–508, 2002.

[SMB+03] M. Su, H.G. Mahvash, L. Brenda, A. Karl, and B. Julia. Representation of south asian people in randomised clinical trials:analysis of trials' data. *BMJ*, 326:1244, June 2003.

[SNA87] J.P. Stuart, L.G. Nancy, and A.T. Anastasios. The Analysis of Multiple Endpoints in clinical trials. *BIOMETRICS*, 43:487–498, 1987.

[SS11] N. Sut and O. Simsek. Comparison of regression tree data mining methods for prediction of mortality in head injury. *Expert Systems with Applications*, 38(12):15534–15539, 2011.

[SSPGH06] N.G. Strouthidis, A. Scott, N.M. Peter, and D.F. Garway-Heath. Optic disc and visual field progression in ocular hypertensive subjects: detection rates, specificity, and agreement. *Investigative ophthalmology & visual science*, 47(7):2904–2910, 2006.

[Suz93] J. Suzuki. A construction of bayesian networks from databases based on an mdl principle. In *Proceedings of the Ninth international conference on Uncertainty in artificial intelligence*, pages 266–273. Morgan Kaufmann Publishers Inc., 1993.

[TGH10] A. Tucker and D. Garway-Heath. The pseudotemporal bootstrap for predicting glaucoma from cross-sectional visual field data. *Information Technology in Biomedicine, IEEE Transactions on*, 14(1):79–85, 2010.

[TH05] M. Talih and N. Hengartner. Structural learning with time-varying components: tracking the cross-section of financial time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):321–341, 2005.

[TL04] A. Tucker and X. Liu. A bayesian network approach to explaining time series with changing structure. *Intelligent Data Analysis*, 8(5):469–480, 2004.

[TLOS01] A. Tucker, X. Liu, and A. Ogden-Swift. Evolutionary learning of dynamic probabilistic models with large time lags. *International Journal of Intelligent Systems*, 16(5):621–645, 2001.

[TSN+12] M. Takada, M. Sugimoto, Y. Naito, H. Moon, W. Han, D. Noh, M. Kondo, K. Kuroi, H. Sasano, T. Inamoto, et al. Prediction of axillary lymph node metastasis in primary breast cancer patients using a decision tree-based model. *BMC Medical Informatics and Decision Making*, 12(1):54, 2012.

[TVLGH05] A. Tucker, V. Vinciotti, X. Liu, and D. Garway-Heath. A spatio-temporal bayesian network classifier for understanding visual field deterioration. *Artificial intelligence in medicine*, 34(2):163–177, 2005.

[VAH+10] S. Visweswaran, D.C. Angus, M. Hsieh, L. Weissfeld, D. Yealy, and G.F. Cooper. Learning patient-specific predictive models from clinical data. *Journal of Biomedical Informatics*, 43:669–685, 2010.

[VCT+00] M.L. Vaughn, S.J. Cavill, S.J. Taylor, M.A. Foy, and A.J.B. Fogg. Direct explanations and knowledge extraction from a multilayer perceptron network that performs low back pain classification. In *Hybrid Neural Systems*, pages 270–285. Springer, 2000.

[VFH97] A.C. Viswanathan, F.W. Fitzke, and R.A. Hitchings. Early detection of visual field progression in glaucoma: a comparison of progressor and statpac 2. *British Journal of Ophthalmology*, 81(12):1037–1042, 1997.

[vGTL08] M.A.J. van Gerven, B.G. Taal, and P.J.F. Lucas. Dynamic bayesian networks as prognostic models for clinical patient management. *Journal of biomedical informatics*, 41(4):515–529, 2008.

[VPR+07] M. Verduijn, N. Peek, P.M.J. Rosseel, E. Jonge, and B.A.J.M. Mol. Prognostic bayesian networks i rationale, learning procedure, and clinical use. *Journal of Biomedical Informatics*, 40:609–618, 2007.

[vVDVDV+02] Laura J. van't V., H. Dai, M.J. Van De Vijver, Y. D He, A.AM Hart, M. Mao, H.L Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

[WB08] E.W. Watt and A.A.T. Bui. Evaluation of a dynamic bayesian belief network to predict osteoarthritic knee pain using data from the osteoarthritis initiative. In *AMIA Annual Symposium Proceedings*, volume 2008, page 788. American Medical Informatics Association, 2008.

[WDL+90] J.H. Ware, D.W. Dockery, T.A. Louis, X. Xu, B.G. Ferris, and F.E. Speizer. Longitudinal and cross-sectional estimates of pulmonary function decline in never-smoking adults. *American journal of epidemiology*, 132(4):685–700, 1990.

[WGO03] X. Wang, J. Garibaldi, and T. Ozen. Application of the fuzzy c-means clustering method on the analysis of non pre-processed ftir data for

cancer diagnosis. In *Internat. Conf. on Australian and New Zealand Intelligent Information Systems (ANZIIS)*, pages 233–238, 2003.

[Wik13] Wikipedia. Expectationmaximization algorithm — wikipedia, the free encyclopedia, 2013. [Online; accessed 10-September-2013].

[Wil07] D.J. Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Briefings in bioinformatics*, 8(2):109–116, 2007.

[WLH+09] C. Wen, W. Liao, T. Hsieh, D. Chen, J. Lan, and K. Li. Computer-aided image analysis aids early diagnosis of connective-tissue diseases. *SPIE Newsroom, Biomedical Optics & Medical Imaging*, 2009.

[WLL99] M.L. Wong, W. Lam, and K.S. Leung. Using evolutionary programming and minimum description length principle for data mining of bayesian networks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(2):174–178, 1999.

[Wor09] At Work. Cross sectional vs longitudinal studies, 2009. https://www.iwh.on.ca/at-work/at-work-55.

[WP02] P.C. Woodland and D. Povey. Large scale discriminative training of hidden markov models for speech recognition. *Computer Speech & Language*, 16(1):25–47, 2002.

[XM07] X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th international conference on Machine learning*, pages 1055–1062. ACM, 2007.

[YD03] M. Yanoff and J.S. Duker. Ophthalmology 2nd edition, 2003.

[YJK+07] M. Yousef, S. Jung, A.V. Kossenkov, L.C. Showe, and M.K. Showe. Naïve bayes for microrna target predictionsmachine learning for microrna targets. *Bioinformatics*, 23(22):2987–2992, 2007.

[YJZ+06] H. Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li. A multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert Systems with Applications*, 30(2):272–281, 2006.

[YSA13] Li Y., Swift S., and Tucker A. Modelling and analysing the dynamics of disease progression from cross-sectional studies. *Journal of Biomedical Informatics*, 46(2):266–274, 2013.

[ZB08] M. Zvelebil and J. Baum. *Understanding Bioinformatics*. New Yor: Garland Science, 2008.

[Zha04] H. Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004.

[ZL10] T. Zeng and J. Liu. Mixture classification model based on clinical markers for breast cancer prognosis. *Artificial Intelligence in Medicine*, 48(2):129–137, 2010.

[ZLNRP99] I. Zelič, N. Lavrač, P. Najdenov, and Z. Rener-Primec, editors. *Impact of machine learning to the diagnosis and prognosis of first cerebral paroxysm*, 1999.

[ZOM99] A. Zafar, J.M. Overhage, and C.J. McDonald. Continuous speech recognition for clinicians. *Journal of the American Medical Informatics Association*, 6(3):195–204, 1999.

[ZSAK+05] L. Zhang, D. Samaras, N. Alia-Klein, N. Volkow, and R. Goldstein. Modeling neuronal interactivity using dynamic bayesian networks. In *Advances in neural information processing systems*, pages 1593–1600, 2005.

Appendix A

A.1 Math Notation

C	defined classes labels
$C_i = 0$	sample i corresponds to a healthy case
$C_i = 1$	sample i corresponds to a disease case
D	a real valued cross-sectional data
$D_{(i)}$	ith row of matrix D
D_k	a distance matrix
d_{ij}	shortest path from i to j
G	weighted graph
h	hidden states
k	number of pseudo time-series
m	the number of samples (patients)
n	the number of variables in the clinical test data
P	a set of pseudo time pseudo time-series
$P_a(x_i)$	the parent set of a node X_i
Q	a continuous vector of unknown parameters
T	sample size
w_{ij}	weight matrix
X^t	the variables in the time series
X	observed data
Z	a set of unknown latent variables

A.2 Abbreviation

ARHMM	Auto-regressive hidden Markov model
ADTree	Alternating Decision Tree
BC	Breast Cancer
BN	Bayesian Network
CPD	Conditional Probability Distribution
DAG	Directed Acyclic Graph
DBN	Dynamic Bayesian Network
DTree	Decision Tree
EM	Expectation Maximisation
HMM	hidden Markov model
HRT	Heidelberg Retina Tomography
KL	Kulbaeck Leibler
ML	Machine Learning
MNN	Multilayer Neural Network
MPTS	Multivariate Pseudo Time-Series
MTS	Multivariate Time-Series
NBs	Naïve Bayes
NN	Neural Network
PD	Parkinson's disease
PTS	Pseudo Time-Series
TBS	Temporal Bootstrap
VF	Visual Field