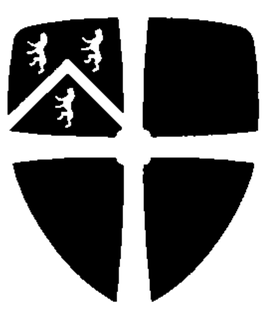
Clinical decision support

Cumming, Jonathan

Clinical Decision Support

Jonathan Cumming

A Thesis presented for the degree of
Doctor of Philosophy

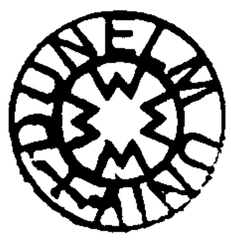


Department of Mathematical Sciences

Durham University

England

September 2006



Clinical Decision Support

Jonathan Cumming

Submitted for the degree of Doctor of Philosophy

September 2006

Abstract

Within orthopædics, clinicians routinely take multiple measurements on patients during the course of their treatment, often repeating the same measurements before and after operations, and subsequently at periodic follow-up consultations. This data combined with additional factors gives a wealth of information, resulting in a high-dimensional data set with a mixture of data types and a longitudinal aspect; all of which can be problematic in statistical analysis. Therefore, general statistical methods for the investigation and analysis of a generic medical data set are presented and developed.

Methods are proposed for supporting exploratory analysis of the data via novel visualisations of the patient's status over time across multiple variables, thus giving an easily interpretable overview of this evolution. To address the problem of high dimensionality of the data, a new approach to variable selection is proposed and developed using principal variables. The method is further extended by the use of temporal smoothing to tackle data with this repeated measures aspect allowing for the simultaneous reduction of the patient status variables over time. The ultimate goal of these analyses is to determine an appropriate model for the orthopædic data, with a focus on the modelling of the time series of patient progress. The techniques of graphical modelling and, in particular, those of chain graphs lend themselves to this problem. Additionally, they have the added benefit of a simple and intuitive visualisation which is of benefit to clinicians. All of these methods are illustrated via their application to two large-scale case study data sets concerning total joint replacement.

Declaration

The work in this thesis is based on research carried out at the Department of Mathematical Sciences, Durham University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Word count: 91601

Acknowledgements

I would like to thank my supervisor, David Wooff, for his expert advice and invaluable guidance throughout the course of my studies. I would also like to thank my industrial supervisor, John Egan, for his assistance in relating this work to orthopædics and communicating it to the orthopædic community, and I am also grateful for the financial support and laptop provided by E-Tech. I would like to thank Professor Paul Gregg, Jan van der Meulen, and the Royal College of Surgeons for providing the two orthopædic data sets that have been instrumental in developing the methods discussed in this thesis. As a Faraday Associate, I would also like to acknowledge the support provided by the Smith Institute. Finally, I would like to acknowledge the financial support of Engineering and Physical Sciences Research Council.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General Background

Clinical decision support is a term that is frequently used within the literature to describe a computer application which employs statistical methods or techniques from artificial intelligence to provide a clinician with important information or results that could inform their decision-making process, often with regard to a specific patient. These clinical decision support systems (CDSS) tackle a wide variety of different problems within the medical domain, ranging from the suggestion of diagnoses on the basis of certain symptoms, to time-sensitive monitoring of ventilator systems in intensive care, and further to using toxicological information to warn clinicians if they have prescribed a dangerous drug combination. It was Wyatt and Spiegelhalter [129] who gave the now commonly accepted definition of a clinical decision support system as:

> "an active knowledge system which uses two or more items of patient data to generate case-specific advice."

From such a definition, the role of statistics and statistical methods within a CDSS is to provide the mechanisms by which such advice is given to the clinician on the basis of the data.

It is the development and bringing together of such techniques and methods to form this framework of decision support that is a key goal of this thesis. The

setting of this decision-support framework is that of orthopædics and total joint replacement in particular. Within this area, clinicians routinely make several measurements on patients during the course of their treatment, often repeating the same measurements before and after treatment, and subsequently at periodic follow-up consultations. Consequently, a large wealth of information can soon accumulate and without detailed statistical investigation of these data sets the information they contain will typically be untapped. Thus one of the primary goals of this thesis is to provide the statistical support for an informative analysis of these data in order to inform the clinician and their decision-making.

This task is by no means straightforward as the data themselves are typically complex and varied in structure and content. Typically within orthopædics, a series of particular measurements will be repeatedly measured on a patient to assess and monitor their clinical status. These measurements are often the components of one of the many defined composite scores, such as the Oxford hip score [27], and are summed or averaged to give a single indicator of the patient's state. Whilst this method is a simple mechanism for assessing the patient, the combination of these various diverse measurements into a single quantity loses the detailed information present in the individual variables. Thus it becomes impossible to learn, for example, that a patient's status has worsened because of an increase in the individual pain measurements. This aggregation of these individual informative measurements into a single number is a severe over-simplification of the data and sacrifices a great deal of information on the patient in the process. Therefore, rather than perpetuate this simplification of the data it would be advantageous to retain the individual elements of these scores and analyse them together to gain a richer impression of the patient's condition.

As mentioned above, a patient is usually seen on several occasions during treatment - before and after their operation and possibly at subsequent follow-up consultations. Consequently, the dimensionality of the data becomes large as the constituent measurements of the composite scores are repeatedly observed at each time point. This compounds the size of the data set and introduces a longitudinal or repeated measures element to the data. However, these patient status variables are

not the only elements present in the data set. A large number of other variables are also recorded, such as patient demographic characteristics, details of diagnosis and treatment, and complications due to surgery. These additional factor variables all record information that may or may not impact on the patient's condition and the manner in which this condition evolves over time. Furthermore, there is no restriction on the nature of any of the variables, be they patient status measurements or other factors. In other words, the variables in the data can be a mixture of continuous, categorical and ordinal quantities.

The ultimate goal of this thesis is to attempt to provide mechanisms for the intelligent analysis of such data sets. However, a key restriction on the development of these methods is that they should apply to a general orthopædic data set, and that they should not be tailored to a specific problem area. This is a formidable challenge to develop or apply statistical methods to an unseen arbitrary data set that may typically be high dimensional, have a longitudinal component, and contain variables of mixed types.

A key feature of the analyses in this decision support framework is that the results should be easily interpretable to the clinician. Using methods that are intuitive in their understanding will enable clinicians to make full use of the statistical support being presented to them. Giving results that require interpretation by a statistician would be of no use in informing the clinician's decisions and so are inappropriate here.

Some general techniques that could be applied to these data would include an investigation of the variables representing the patient's medical status. The manner in which these measurements change over time would be an area of significant interest and the modelling of the patient status over time may allow for the prediction of patient state at a later time given their current position. Additionally, the relationships between these measurements would also be informative and could suggest measurements which were effectively redundant or uninformative - this would have direct implications for the calculation and interpretation of the composite scores. The effects and the interplay of the additional factor variables are also areas of possible interest since they could potentially have strong effects on the patient's status

and its evolutions over time. The large size and scale of the data set can be a daunting prospect, and hence the distillation of these large complex data structures into formats that are clinically useful and understandable would be of significant value, such as through appropriate plots and visual methods.

1.2 Overview

The remainder of this thesis is organised in a chronological fashion beginning with exploratory investigations of available orthopædic data sets presented before moving onto aspects of modelling these data, the associated problems and their solutions. As this thesis covers a broad range of subject areas within applied statistics, we have refrained from presenting a single literature review in favour of presenting such material at the appropriate location within the text.

Chapter 2 begins by introducing the types of orthopædic data to be studied. An abstract framework for such data and their analysis in generality is proposed and the two major orthopædic data sets are introduced. Chapter 3 follows on by presenting an exploratory analysis of these data in order to provide an understanding of the nature and the behaviour of the data as well as to highlight potentially problematic features thereof. Chapter 4 continues the exploratory theme by proposing a series of graphical methods for displaying complex high-dimensional data sets in a compact and intuitive form.

Having explored the data, Chapter 5 moves on to consider application of the graphical modelling methodology to the data with a specific view to developing a single model for each data set. Problems due to the high dimensionality of the orthopædic data necessitated examining the possibilities of applying data or variable reduction strategies. Therefore, Chapter 6 discusses such strategies and develops a novel method for selecting important variables from a data set that is replicated over time. This method is tested via simulation studies and compared to other variable selection methods available in the literature and is shown to perform very well. Chapter 7 then takes this variable selection procedure and applies it to the orthopædic data. Once we have identified a suitable subset of the original variables

from the data, we return to the graphical modelling of the data. Chapter 8 introduces a specialisation of graphical models, the chain graph model, and applies these techniques to the data to develop suitable final models of these data. The models are then used to provide predictions and the adequacy of these models and their goodness of fit is assessed and discussed. All of the methods and techniques proposed and presented in this thesis are discussed in Chapter 9, along with a thorough review of the problems and limitations that have been encountered in the course of the analysis. Finally, Chapter 10 contains a summary of the conclusions, both in general and those specific to orthopædics and medicine.

Chapter 2

Data Generalisation and an Introduction to the Data

The primary aim of this thesis is to investigate general sets of orthopædic data, rather than focusing on a single data set. Consequently, there is a need to develop a general structure for such data sets so that methods and techniques can be developed and applied to any data which conform to this abstract pattern. Using such an overarching framework for the types of data sets within this domain, we can envisage the types of research questions that are likely to be asked and can select or develop appropriate statistical techniques to answer them. This chapter therefore begins with the development of a general data framework in Section 2.1. The uses of such an abstraction would be limited without real data to examine. Two significant orthopædic data sets have been made available for study, and these data sets will be used throughout this thesis as extended case studies. Consequently, in Section 2.2 an overview of these data is presented and their origins, notable features and relationship to the general data model are discussed.

2.1 Data Generalisation

To generalise the data, we seek to develop an abstract framework encompassing the key features of the data's structure. In this setting, the word 'structure' is not used to describe relationships within the data, as in the term 'correlation struc-

ture'. Instead, it is used to describe fundamental relationships that exist *a priori* between the quantities being recorded in the data. For example, if a variable $\boldsymbol{x}$ is recorded chronologically before $\boldsymbol{y}$ then this ordering of the variables is a form of this structuring of the data. In this sense, we are specifying the *meta-structure* of the data.

Ordinarily, one would develop specific solutions and models for specific data sets. However, it may then be difficult or infeasible to extend these solutions to different data, which thereby limits the scope of their application. By working within a general framework, we allow for the development of techniques and methods that are applicable to any data which conform to our structural pattern. This is especially useful since it allows for the potential generalisation of the methodology to areas beyond the current specific domain of research.

Many abstractions of medical data have been developed in recent years, primarily motivated by the need to develop a computerised patient records system. These systems include the HL7 Reference Information Model [5], the Good European Health Record [63] and the European Health Record Architecture[17]. However, these generalisations are typically too technical in nature and too detailed in structure to be useful in these circumstances. Therefore a more abstract structure must be created.

2.1.1 Towards a General Structure

By using prior knowledge about the general forms of data within the orthopædic domain we can begin to develop a skeletal framework for an arbitrary data set. Firstly, the data are typically chronological in nature, with data being recorded at several different time points. This imposes a clear temporal structure on the variables within the data. This has the consequent implications that data observed at any one time point can depend only on observations that are either contemporaneous or prior in time.

The time points or *events* at which the data are recorded are:

- An *Initial Consultation*,
- resulting in a *Clinical Assessment*,

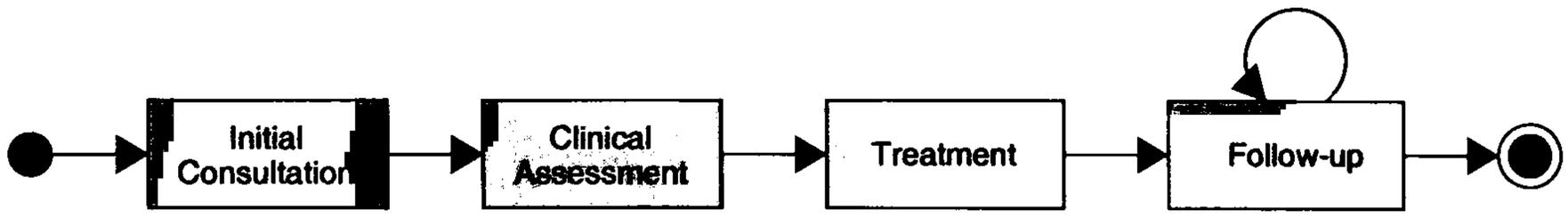


Figure 2.1: Process diagram of the basic 4-event structure.

- leading to *Treatment*,
- which is followed by one or more *Follow-up Consultations.*

This simple general structure is displayed in a standard UML process diagram [111] in Figure 2.1. By considering each event in the process and the data obtained at each point, we can refine this basic structure into a more usable framework.

At the first event, the Initial Consultation, we observe and record several sets of data. The first of which are the patient demographics. These consist of variables directly associated with the patient such as their age and sex as well as other potentially interesting or significant factors such as whether the patient is being treated in the public or private sectors and the length of time they have been on a waiting list. These factors are all pre-determined and non-random. These variables therefore represent potentially important covariates in any model of the data, so it is logical to place these variables first in the temporal ordering.

The goal of the Initial Consultation is to attempt to learn as much as possible about the unknown condition of the patient in order to determine the severity of their condition. This is achieved by measuring another key set of variables such as measures of the patient's pain and mobility. These variables can then be re-observed at the Follow-up Consultations and will allow the tracking and monitoring of the patient's condition. It is common in orthopædics for such variables to form the components of a standard scoring system, which are summed to provide a single numerical score for each patient.

The next stage is the Clinical Assessment. This could take the form of a diagnosis of the patient's condition. However, with rheumatoid and osteoarthritis in the hips and knees data the differences between the two conditions are considered to be such that the diagnosis can be made with certainty and is therefore treated as

a demographic characteristic of the patient. Another form of clinical assessment would be the decision on the type of treatment the patient was to receive (if any). However, if the patient is not treated it is likely that no data will have been recorded.

The subsequent stage in the process is the Treatment itself and at this stage there are two sets of data recorded. The first is composed of factors such as, most importantly, the type of treatment as well as any other characteristics, such as the type of anaesthetic the patient received or the grade of the operating surgeon. The second set of data recorded is the immediate outcomes, such as whether the operation was successful and details of any complications that may have arisen.

The final stage is the Follow-up Consultation at which point the key status variables are re-observed and recorded. Again there could be more potentially informative one-off variables recorded. Since this could also feasibly be the final stage in the process, it is likely that a measure of success is recorded, such as the patient's or clinician's satisfaction. The Follow-up Consultation could then be subsequently repeated, allowing for a further replication of these variables.

One general feature of the data is likely to be the mixture of data types. Many variables such as patient pathology and operation type are categorical, and the data could also include binary variables such as whether the patient was readmitted to hospital. However, variables such as age, weight and other clinical measurements can be regarded as continuous, and more subjective measurements such as pain levels are typically measured on an ordinal Likert-scale.

The complete version of the process and data collected is given in Figure 2.2. The reduced model for the data only is presented in Figure 2.3.

2.1.2 Discussion

It is clear that the most informative data to the clinician would be the set of key variables measuring the patient's status. In fact, these replicated observations of the same variables fall into the framework for repeated measures [24, 31, 86] or longitudinal data [32, 123]. However these frameworks would struggle with the mixture of data types and the inclusion of large amounts of extraneous information that is not repeatedly observed.

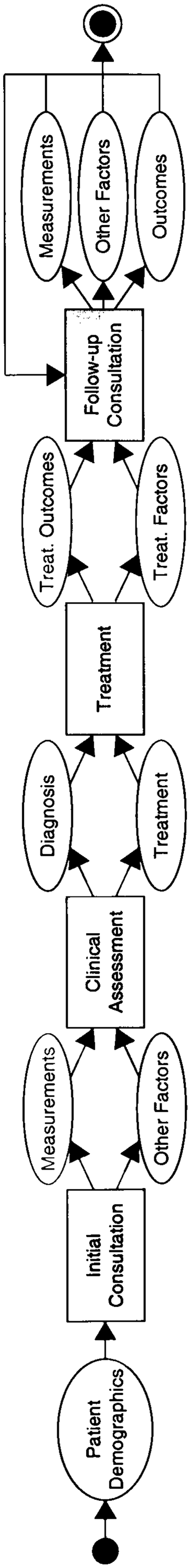


Figure 2.2: Temporal ordering of events and data for general orthopædic data.

Figure 2.3: Temporal structure of general orthopædic data.

Understanding the relationships between this group of key observations would be highly beneficial, as well as providing insights into how these relationships evolve over time. The effects of different treatment and other factor variables could be assessed by their effects on this core group of repeated observations. Prediction of the future status of the patient given certain treatment options would also be of interest to the clinician. Another interesting avenue of investigation would be to determine which other variables affect the final measures of success.

Having generated this general structure of the data, we can now observe that the framework is not specific to data on total joint replacement, nor is it even specific to orthopædics. This generalisation would cover any similar longitudinal study data within medicine or even in other fields.

2.2 Introduction to the Data

2.2.1 The Knees Data

The first data set to be examined (hereafter referred to as the 'knees data') consists of data on 599 patients who underwent a total knee replacement procedure between 1987 and 1997 [88, 72]. The data contains a total of observations on up to 124 variables for each patient. These data were collected through a series of up to four consultations between the patient and clinician and occurred pre-operatively and then at one, five and ten years post-operatively.

A total of 23 measurements were recorded at each consultation to monitor patient progress. These measurements form the components of the Nottingham knee scoring system [118] and thus compose a small repeated measures data set, corresponding to the *Measurements* in the general data structure of Figure 2.3. The 23 measurements are composed of a mixture of both ordinal and continuous data. The ordinal data represent measurements of subjective quantities such as pain levels and walking ability and are all measured on a 5-point Likert-scale where 1 represents the worst and 5 the best possible state. The continuous variables however, correspond to measurements of anatomical angles. It could be argued that angles should be treated as circular rather than continuous measurements, however since they represent the

range of motion of a limb the values of these measurements will only cover a fraction of the circle.

In addition to these repeated measurements the data also contain several variables recording patient details such as age, sex, and weight - these are again a mixture of continuous and categorical data. These variables correspond to the *Patient Demographics* in the data abstraction. The data also contains information on the patient's diagnosis, which was one of four possible conditions, and operation type, which was also a four-state discrete variable corresponding to the use of cement during the procedure. These variables are known as *Diagnosis* and *Treatment* in the general framework. In this case the type of treatment the patient received was randomised, where the patients either had an operation where either cement was or was not used during the procedure. Additionally, there are several binary variables corresponding to whether the patient experienced specific complications, these variables would be considered *Treatment Outcomes* under the data generalisation. A summary of the majority of the variables in the knees data set and their relationship to the general model is given in Table 2.1.

As the data is collected through a series of consultations over a 10-year period, there is variability in the sample size at each time point. Due to the long time-scale of the data collection, there is consequent attrition in the sample size at later time points. This is likely due to patients leaving the study, leaving the area or patient death. This results in a drop in the sample size from 599 cases pre-operatively to 559, 239 and 86 at one, five and ten years post-operatively. This dramatic drop in sample size would mean that statistical methods requiring complete cases for analysis would be inappropriate as we would only have 86 such cases to work with, and the remainder would either have to be imputed or discarded.

A closer examination of the data revealed that two of the repeated measurements - *Range of Motion* and *Other Knee Range of Motion* - were linear combinations of other variables in the data. These variables were removed from the data since they did not contribute any novel information to the data. Additionally, there was a pair of repeated measurements where only one of the pair was ever observed. These variables were *Coronal Tibio-Femoral Varus* and *Coronal Tibio-Femoral Valgus*,

Variable Name	Scale	Levels	Component in Data Generalisation
Age	Categorical	2	Patient Demographic
Sex	Categorical	2	Patient Demographic
Weight	Continuous		Patient Demographic
Diagnosis	Categorical	4	Diagnosis
Operation	Categorical	4	Treatment
Pain Frequency	Ordinal	5	Repeated Measurement
Pain Severity	Ordinal	5	Repeated Measurement
Night Pain	Ordinal	5	Repeated Measurement
Walking Ability	Ordinal	5	Repeated Measurement
Walking Aids	Ordinal	5	Repeated Measurement
Sitting Down	Ordinal	5	Repeated Measurement
Rising Up	Ordinal	5	Repeated Measurement
Standing	Ordinal	5	Repeated Measurement
Going Up Stairs	Ordinal	5	Repeated Measurement
Going Down Stairs	Ordinal	5	Repeated Measurement
Coronal Tibio-Femoral Angle	Continuous		Repeated Measurement
Fixed Contracture	Continuous		Repeated Measurement
Flexion	Continuous		Repeated Measurement
Extension Lag	Continuous		Repeated Measurement
Hip Abduction	Continuous		Repeated Measurement
Other Knee Fixed Contracture	Continuous		Repeated Measurement
Other Knee Flexion	Continuous		Repeated Measurement
Other Hip Abduction	Continuous		Repeated Measurement
Infection	Categorical	2	Treatment Outcome
DVT	Categorical	2	Treatment Outcome
Satisfaction	Ordinal	5	Outcome

Table 2.1: Summary of several variables of the knees data set.

	Diagnosis			
Operation	OA	RA	Osteonecrosis	Other
Cemented	319	31	1	3
Uncemented	212	18	1	0
Cemented Femur, Uncemented Tibia	2	1	0	0
Uncemented Femur, Cemented Tibia	8	3	0	0

Table 2.2: Contingency table of Diagnosis and Operation for the knees data.

which measured an anatomical angle of the limb either towards the body (varus) or away from the body (valgus). In this case, the pair was combined into a single variable *Coronal Tibio-Femoral Angle* with a varus value being recorded as positive and a valgus value recorded as negative.

The knees data contained many missing observations; the overall proportion of missing observations was 1%, which was relatively small. These missing values were simply imputed with the mean over all the non-missing observations [84]. This is common practice with orthopædic data and is straightforward to apply, without introducing models for the data at this premature stage. However, imputation by the mean does have the negative effect of attenuation in the variance of the imputed variables.

The categorical variables Diagnosis and Operation record potentially valuable and informative information about the patient's pathology and the type of treatment they received. However, as is shown in Table 2.2 we can see that the contingency table is somewhat sparse outside of the first two diagnoses and operations. Since these variables are considered to be critical factors to any analysis of these data, attention will be restricted to the two larger categories of each variable. The number of cases with the other pathologies and treatment types are so small it would be difficult to analyse them in the same way as the other data - it is likely that they should be examined individually on a case-by-case basis.

2.2.2 The Hips Data

The second data set consists of data gathered by the Royal College of Surgeons during their National Total Hip-Replacement Outcome Survey (NTHROS) [94] (referred to as the 'hips data'). This data set is far larger than the knees data and contains information on 12 666 patients who received a total hip replacement in the United Kingdom between 1996 and 1997. The data set is also larger in terms of breadth as well as number of cases, with the number of variables recorded per patient being a maximum of 202. Unlike the knees data which were gathered at face-to-face consultations, these data were collected via six separate questionnaires.

The first questionnaire is completed by the surgeon and contains 50 variables. These variables comprise basic *Patient Demographics*, such as patient sex and date of birth (which was used with the date of surgery to generate a continuous variable for the patient's age at that time), as well as whether the patient was private or NHS. Details regarding the patient's *Diagnosis* are recorded in a 6-state categorical variable which specified the pathology as being one of a combination of osteoarthritis, rheumatoid arthritis or 'other'. This variable was split into three binary variables each indicating the presence or absence of each individual condition. The *Treatment* variables were identified to be two binary variables relating to the use of cement in the operation (for similarity with the knees data) and two categorical variables representing the designs of the components of the prostheses. The surgeon's questionnaire also contained a large number of binary or categorical variables which could be identified as *Treatment Factors* - these include variables such as whether the patient received a general anæsthetic or an epidural and the grade of the operating surgeon. The surgeon's questionnaire also contains two variables relating to operative difficulties or complications which could be identified as *Treatment Outcomes.*

The patients themselves completed three of the questionnaires in the hips data set. These questionnaires were collected pre-operatively and at three and twelve months post-operatively. These questionnaires contain variables that fall into the *Repeated Measurements* and *Other Factors* categories. Twelve ordinal variables are recorded to assess the patient's status - these variables form the components

of the Oxford Hip score [27]. All of the variables are measured on a 1–5 scale, however unlike the knees data lower values represent a better patient condition than higher values. To form the Oxford Hip Score, the values are simply summed. The remaining variables cover various areas such as the patient's medical history in the pre-operative questionnaire, to information on their convalescence in the 3-month survey.

A further questionnaire contained some 55 variables recording specific details of any operative complications encountered. The final questionnaire was far smaller than all the others, containing only 7 variables. This questionnaire was directed at the patient's GP and records details such as whether the patient was treated for deep-vein thrombosis, and whether they were admitted to hospital. A summary of some of the variables in the hips data set is given in Table 2.3.

As with the knees data, there are varying sample sizes in the hips data due to the level of response to the individual questionnaires. There were a total of 10 411 completed surgeon's questionnaires and 7150, 6163 and 5917 completed pre-op, 3-month and 12-month patient questionnaires respectively. However, if we consider only those cases where the surgeon questionnaire and all patient questionnaires are complete, then the sample size drops further to only 2474. For the complications and GP questionnaires the sample sizes were substantially less at 975 and 698 respectively.

Furthermore, there is an additional complication with the hips data in that there is a problem with invalid data. The invalid data arose due to the nature and design of the questionnaires. Since all the data is categorical in nature, the questionnaire made it possible for a respondent to choose more than one value in a group of mutually exclusive categories, such as ticking both the 'Yes' and the 'No' boxes. In these cases, the data for those variables was discarded and treated as if it were missing.

Missing data in the hips data is more of a prominent problem than with the knees data set. In fact, for the pre-operative questionnaire 43% of the 12666 questionnaires recorded were totally blank and so were immediately discarded. The remaining questionnaires, however, were well answered with 92% of this remainder

recording values for all of the *Repeated Measures* variables. This leaves us only a small proportion of cases requiring imputation of the missing values. A similar pattern is observed in the other questionnaires with the many cases being blank, but with only a small amount of missing data among the remainder.

While the knees data set was a mixture of continuous, categorical and ordinal data, the hips data set contains only discrete data. The only (nearly) continuous variable was that of the patient's age which was calculated from their date of birth and was exact to the day. The hips data is otherwise exclusively composed of ordinal and categorical data.

Variable Name	Scale	Levels	Component in Data Generalisation
Gender	Categorical	2	Patient Demographic
Age	Continuous		Patient Demographic
Private or NHS?	Categorical	2	Patient Demographic
Pathology includes OA	Categorical	2	Diagnosis
Pathology includes RA	Categorical	2	Diagnosis
Pathology includes other	Categorical	2	Diagnosis
Femoral prosthesis type	Categorical	2	Treatment
F.P. cemented?	Categorical	2	Treatment
Surgeon Grade	Categorical	7	Treatment Factor
Usual pain	Ordinal	5	Repeated Measurement
Trouble washing	Ordinal	5	Repeated Measurement
Using transport	Ordinal	5	Repeated Measurement
Putting on socks	Ordinal	5	Repeated Measurement
Shopping	Ordinal	5	Repeated Measurement
Walking without pain	Ordinal	5	Repeated Measurement
Climb stairs	Ordinal	5	Repeated Measurement
Stand up from a chair	Ordinal	5	Repeated Measurement
Limping when walking	Ordinal	5	Repeated Measurement
Sudden severe pain	Ordinal	5	Repeated Measurement
Pain interfered with usual work	Ordinal	5	Repeated Measurement
Pain in bed at night	Ordinal	5	Repeated Measurement
Time on waiting list	Ordinal	4	Pre-op Factor
History of stroke	Ordinal	2	Pre-op Factor
Patient readmitted?	Categorical	2	Post-op Factor
Patient satisfaction	Categorical	3	Outcome

Table 2.3: Summary of several variables of the hips data set.

Chapter 3

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a well-established statistical approach which stems from the work of Tukey [121] and is well documented in the literature. The goal of EDA is to discover patterns in, and develop an understanding of, the data whilst generating hypotheses for subsequent investigation. The focus is typically on the use of visualisations [19, 119] and robust methods rather than complex model building. Tukey described EDA as a 'foundation stone' [121] in the analytic process. Therefore, this chapter contains an exploratory analysis of the two key data sets as an introduction both to the data themselves and to the nature of the problems encountered when analysing them. The chapter begins with a discussion of some of the assumptions made in the analysis. Section 3.2 then proceeds to investigate and explore the knees data set, and this is followed in Section 3.3 by a similar analysis of the hips data set.

3.1 Initial Comments and Assumptions

One key feature of the general orthopædic data as defined in Section 2.1 is that it is composed of a mixture of data types. Many variables such as patient pathology and operation type are categorical, and there are also many binary variables such as whether the patient was readmitted to hospital. However variables such as age, weight and other clinical measurements can be regarded as continuous and the variables which form the components of the hip and knee scores are measured on an

ordinal scale. These different scales for the data necessitate the use of procedures that are not restricted to variables of one particular type.

To simplify this situation, the ordinal variables are assumed to be continuous since ordinal data could be viewed as being the discretisation of a latent continuous quantity [103, 56]. The reasons for this are twofold - the first being that standard statistical techniques could be applied to the data for the exploration to gain a basic insight into the data. The second reason was that the key sets of repeatedly observed variables measuring patient status contain in excess of 10 ordinal variables with approximately 5 levels each. To retain the ordinality of the data would introduce significant and prohibitive dimensionality problems into the analysis which would heavily complicate the problem and would be detrimental to interpretation. The most obvious evidence of this is that the contingency table for such variables would contain more than 9 million cells posing problems both in terms of computer storage and processing.

The data sets are both fairly large with many variables. Thus, the focus of the exploratory analysis will be restricted. The main component of the data sets are the sets of repeated measurements which record the patient status, therefore most attention will be paid to these variables. In particular the properties of these variables (both pre- and post-operatively) along with any relationships to the diagnostic and treatment variables will be investigated. Any potentially informative covariates will also be included.

For both data sets, an identical set of exploratory techniques is applied. Since a primary goal of this thesis is to develop methods that would provide statistical support to clinicians it is envisaged that a substantial portion of the exploratory analyses could be performed semi-automatically. However, in cases where the data or results are unusual this should be brought to the clinician's attention as it may suggest that the methods being applied were inappropriate. In such circumstances alternative analyses would need to be performed.

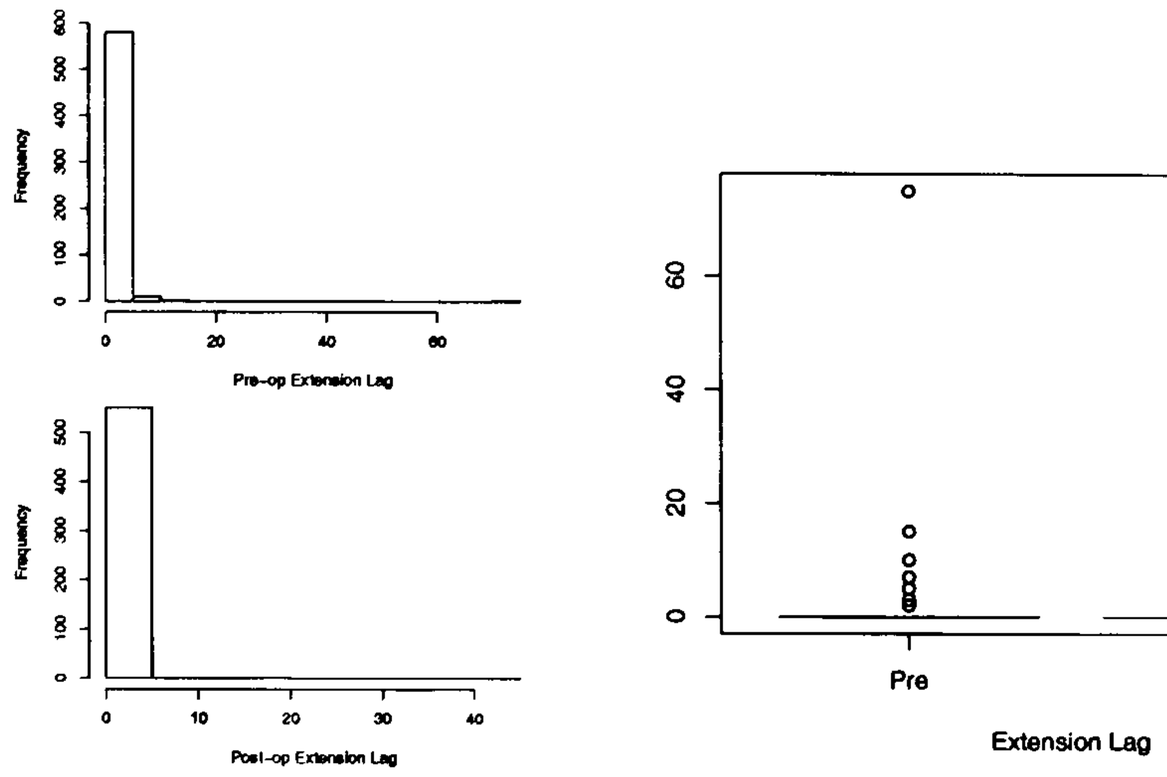


Figure 3.1: Histograms and boxplots of the Extension Lag variable in the knees data.

3.2 The Knees Data

3.2.1 Boxplots and Histograms

Boxplots [89] of the continuous pre-operative measurements show the majority of the variables to be reasonably symmetric with a few possible outliers. The boxplot further shows that the variable *Coronal Tibio-femoral Angle* displays some evidence of skewness, and furthermore the variable *Extension Lag* has a zero IQR and is heavily skewed. Histograms also corroborate the boxplots in identifying the variable *Extension Lag* as being exceptionally skewed giving rise to a number of potential outliers. This is illustrated in Figure 3.1. From this we can see that the majority of the pre-operative observations for Extension Lag are confined in the range 0–5, with a lesser proportion in the range 5–15 and then a single observation at 75. This final value could be due to recording error or a genuine outlier. A similar behaviour was observed post-operatively, with notable outliers at 40 and 44.

Boxplots of the continuous post-operative data show a change in location for *Coronal Tibio-femoral Angle* and the *Fixed Contractures* with the change being in the direction of a reduction. Otherwise, there is little difference between the pre- and post-operative values. The variable *Extension Lag* remains heavily skewed (Figure 3.1).

Plots of the ordinal pre-operative scores show evidence of asymmetry for several variables in the data such as *Going Up Stairs*. The results for three such pre-operative ordinal variables are shown by the histograms and plots of $\overline{x} \pm 2s$ in Figure 3.2, where $\overline{x}$ is the mean value and s is the sample standard deviation. The pre-operative values are shown on the left of that figure. Furthermore, several variables such as the pre-operative scores for *Sitting Down* and *Rising Up* in the knees data appear bimodal. The plots of the post-operative ordinal variables show a notable change in location in the direction corresponding to an improvement of the patient's condition, which suggests the intervening treatment is having a beneficial effect (as shown for the 3 variables to the right of Figure 3.2). Examination of histograms of the data shows that this change in location results in a strong skewness of the data in the direction of improvement.

3.2.2 Scatterplots

Scatterplots of the continuous variables of the knees data showed little evidence of associations among the variables, with the exception of the variables recording *Hip Abduction* and *Other Hip Abduction*. In order to better visualise the ordinal variables on a scatterplot a small amount of random noise or *jittering* was added to the original data [18]. This revealed that the ordinal knees variables typically exhibited a weak positive association. Some variable pairs such as *Going Up Stairs* and *Going Down Stairs*, and *Sitting Down* and *Rising Up* are particularly strongly associated presumably because the measurements are recording similar quantities.

Colouring the points by pathology and treatment indicated that there was no clear differentiation between the different factor levels. This apparent lack of distinction between the pathologies and treatments suggests firstly that the patient's pathology is not uniquely determined by the data recorded. Secondly, this would suggest the patient status, as recorded by these data, is not associated with the choice of treatment and also the different treatment types appear to have indistinguishable outcomes. The only exception here was that there was a possible association between *Weight* and the patient's pathology (see Figure 3.3). The scatterplot would appear to suggest that patients with rheumatoid arthritis (RA) weigh less

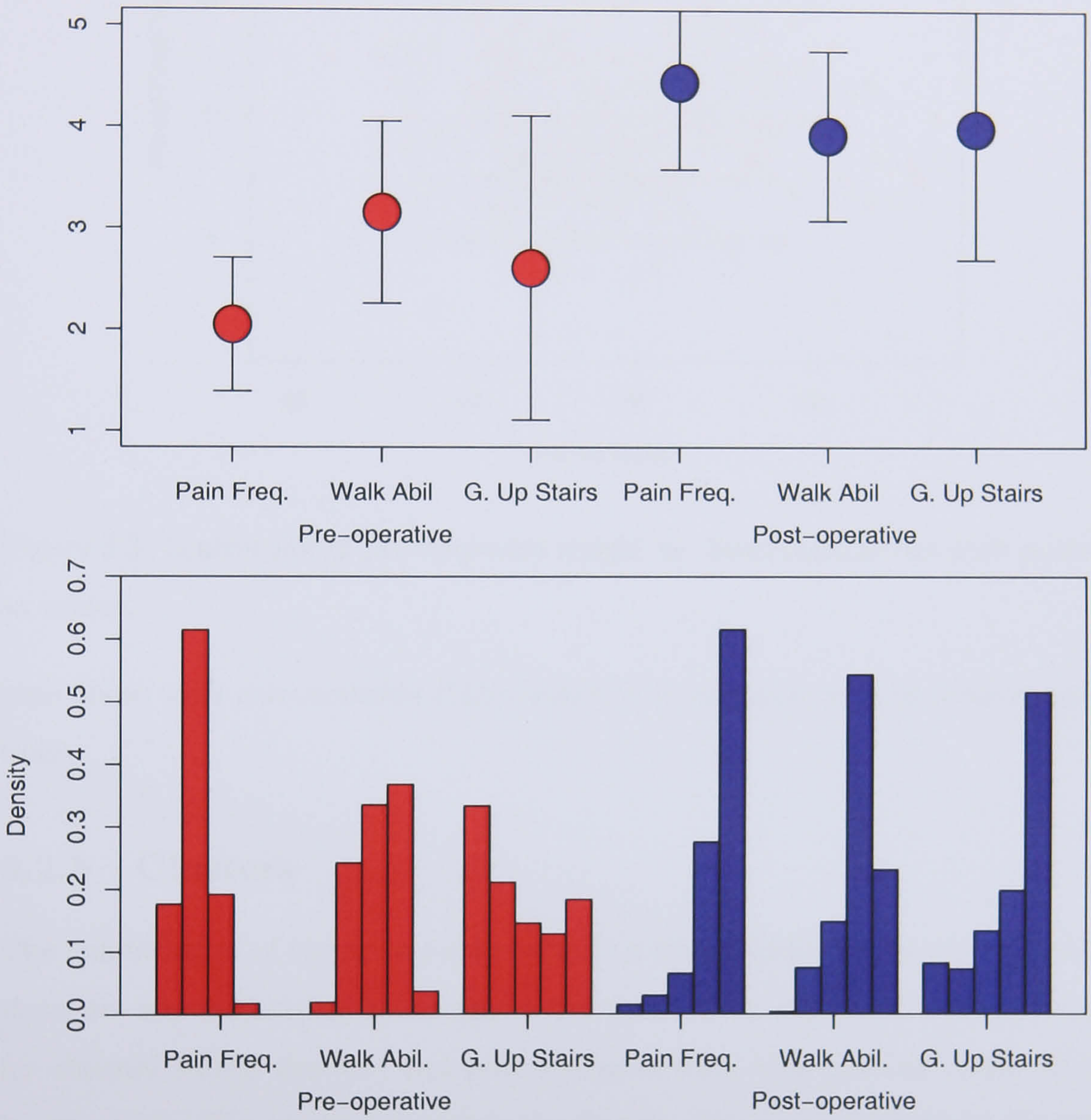


Figure 3.2: Mean and error bar plots and histograms for three ordinal variables in the knees data measured pre- and post-operatively.

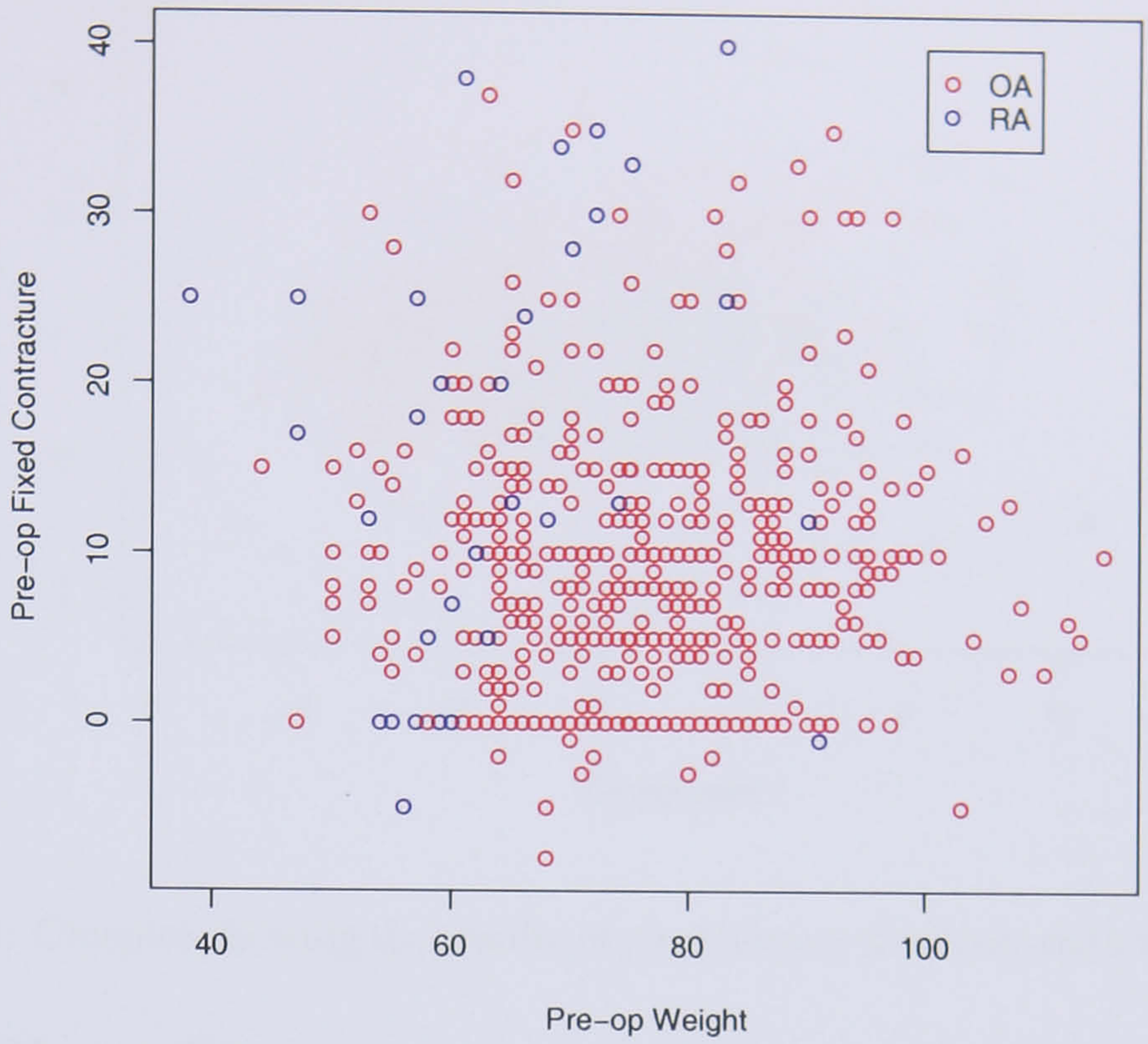


Figure 3.3: Scatterplot of pre-operative weight vs. fixed contracture with pathology as colour.

than those with osteoarthritis (OA), which is in accordance with clinical expectations.

3.2.3 Clusters

The examination of the data for clusters is a useful technique to explore whether there are any underlying groupings in the data. Both data sets were investigated for clusters within the pre- and post-operative data with missing values imputed by the mean. The method used was the Partitioning Around Medoids (PAM) [70] method which is more robust than the standard k-Means approach [60]. There was significant overlap in the clusters obtained for the knees data for values of k from 2 to 10 suggesting there was no evidence of cluster-type behaviour. The results for the clustering with $k = 2$ for the knees data is shown by a Clusplot [99] in Figure 3.4. We can see from this figure that there is a continuum of values, and there is no separation into clusters.

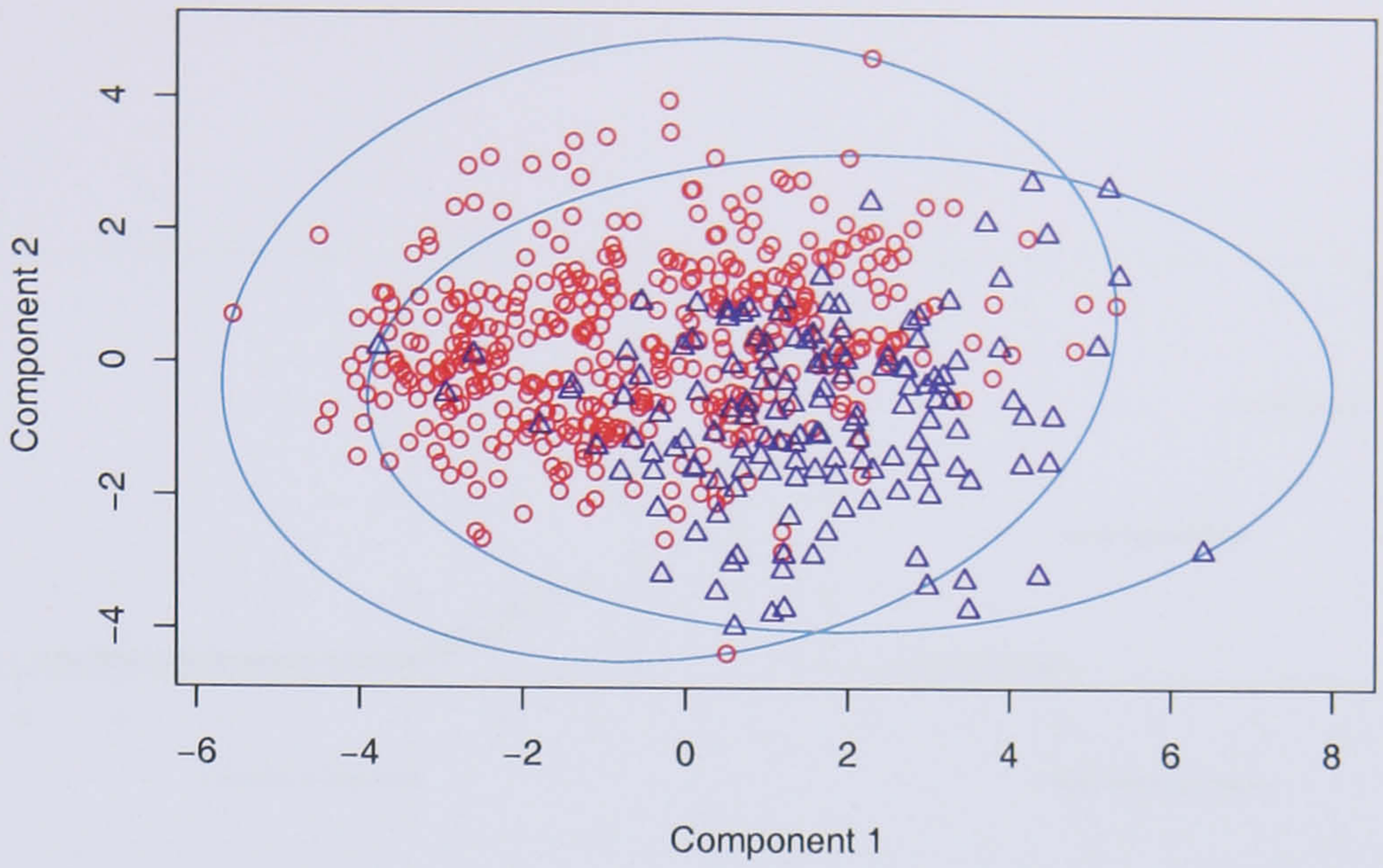
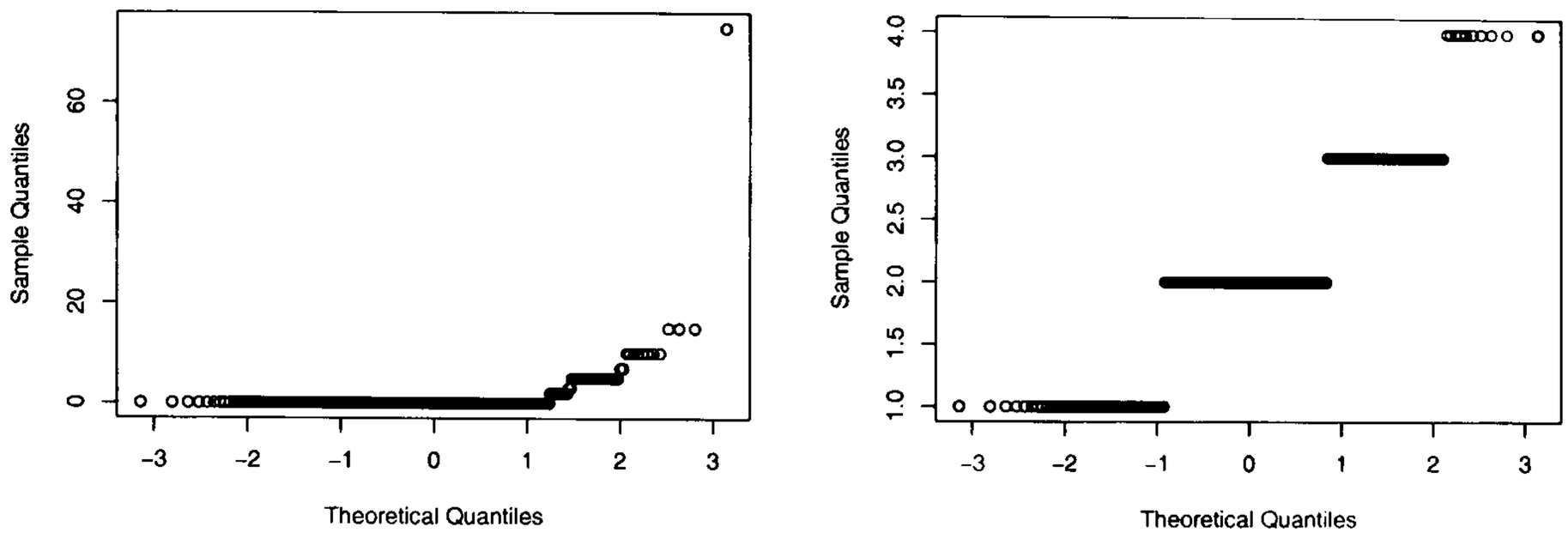


Figure 3.4: Clusplot showing the results of partitioning the knees data into 2 clusters.

3.2.4 Normality

Inspection of Normal quantile plots of the data confirm the lack of Normality highlighted by the boxplots and histograms. Whilst most of the continuous variables such as *Age* and *Weight* appear approximately Normal, *Extension Lag* in the knees data is non-Normal (Figure 3.5(a)). The non-Normality of the ordinal variables is also visible with the granularity effect producing the pronounced step pattern in Figure 3.5(b). The extreme shapes of these quantile plots are to be expected due to the non-Normal nature of the variables. Applying such quantile plots in these circumstances is not a helpful approach to follow, however the results are shown here to illustrate the potential pitfalls of attempting to apply a common analysis to a data set, which contains variables which behave differently.

Transformations of the data provide little improvement. As would be expected, transformation of *Extension Lag* using the maximum likelihood estimate of the power of the Box-Cox transformation [11] did little to improve Normality. Furthermore, a likelihood ratio test rejects the hypothesis that the optimal power of the Box-Cox transformation for all variables in the data is equal to 1, which is strong evidence that the data do not follow a multivariate Normal distribution. The suggested power transformations are given in Table 3.1.



(a) Extension Lag (b) Pain Frequency

Figure 3.5: Normal quantile plots of Extension Lag and Pain Frequency.

Variable	Est.Power	Std.Err	Variable	Est.Power	Std.Err
Age	2.348	0.295	Go Up Stairs	0.248	0.064
Weight	0.418	0.197	Go Down Stairs	0.084	0.065
Pain Freq	0.706	0.099	CTF Angle	1.961	0.118
Pain Severity	0.840	0.101	Fixed Cont	0.154	0.063
Night Pain	0.921	0.092	Flexion	2.229	0.199
Stability	0.604	0.084	Extension Lag	−5.637	0.241
Walk Ability	1.146	0.114	Hip Abduction	1.260	0.077
Walking Aids	2.263	0.197	Oth Knee FCont	0.181	0.064
Sitting Down	0.875	0.119	Oth Knee Flex	2.413	0.213
Rising Up	0.578	0.103	Oth Hip Abd	1.167	0.081
Standing	0.297	0.078			

Table 3.1: Table of the maximum likelihood estimates of the power of the Box-Cox transformation for variables of the knees data.

3.2.5 Means and Standard Deviations

Calculation of the means of the ordinal pre-operative knees variables indicates that the scores are typically falling between 2 and 3 points indicating a relatively poor initial patient condition since 1 point is the worst possible state and 5 the best. Only a few variables such as *Walking Ability* and *Sitting Down* have a mean score over 3 points. The corresponding standard deviations appear to fall into two groups - *Pain Frequency*, *Pain Severity* and *Walking Ability* all have standard deviations between 0.6 and 0.7, whereas the values for the other variables fall between 1.2 and 1.5. This could suggest that pre-operatively the patients' pain levels and walking ability are more consistently poor whereas there is greater variation amongst patients when measured on other scores. Additionally, this could be interpreted as indicating that some of the scores are meaningful in measuring an underlying latent variable, whereas the others are noisier, being more vague concepts.

Investigation of the respective post-operative quantities showed that the mean values were now between 4 and 5 suggesting an increase of 1-2 points when compared with the pre-operative values. Standard deviations also appear to have fallen for the majority of variables and now generally fall between 0.7 and 0.9. This would suggest that an effect of the intervening treatment is to improve the condition and the consistency of the patients.

3.2.6 Correlations

Examination of the pairwise correlation coefficients between the pre-operative variables for the knees data (Table 3.3) indicates that the variables are typically weakly correlated. The variables *Sitting Down* and *Rising Up* are strongly correlated with a value of $r = 0.944$, which is intuitively reasonable since both require the patient to perform a similar task. A strong correlation is also present between *Going Up Stairs* and *Going Down Stairs* ($r = 0.919$). Having such strong relationships within the data may suggest that there could be redundancies within these groups since their information is essentially conveyed by the other variables. There also appear to be moderate correlations between variables that are measured on one knee or hip and

	Pain Sev	Night Pain	Sit Down	Rise Up	Going Up	Going Down	Hip Ab	Other Hip Ab
Pain Freq	0.390	0.251	0.100	0.130	0.084	0.096	0.083	0.077
Pain Sev		0.192	0.057	0.084	0.112	0.131	0.076	0.062
Night Pain			0.102	0.112	-0.003	0.015	0.002	-0.031
Sit Down				0.944	0.308	0.276	0.217	0.178
Rise Up					0.312	0.281	0.216	0.163
Going U						0.919	0.273	0.215
Going D							0.281	0.239
Hip Ab								0.722

Table 3.2: Upper triangle of the correlation matrix for eight of the pre-operative knee variables.

repeated on the other, such as between the variables *Hip Abduction* and *Other Hip Abduction*. Inspecting the post-operative correlations (see Table 3.3) shows that, in general, the variables are more correlated after treatment than before and the same strong relationships that existed pre-operatively persist after treatment. Additionally, the three pain scores *Pain Frequency*, *Pain Severity* and *Night Pain* are now fairly strongly correlated with r between 0.6 and 0.8.

Calculation of the correlations between the pre-operative and the post-operative variables for the knees data shows several strong relationships, especially between *Weight* and the measurements on the unaffected knee or hip. The strength of these relationships is likely due to the fact that the treatment has little direct impact on the values so they are largely determined by the prior values. Only the pain scores, *Stability*, *Sitting Down* and *Rising Up* have negligible correlations with their post-operative counterparts. This is presumably because these scores are the most affected by the treatment that the patient's status on these variables will bear little resemblance to the prior condition.

3.2.7 Discrete Associations

The relationships between discrete variables can also be examined by summarising counts into a contingency table and performing Pearson's χ^2 test for independence

	Pain Sev	Night Pain	Sit Down	Rise Up	Going Up	Going Down	Hip Ab	Other Hip Ab
Pain Freq	0.825	0.626	0.245	0.260	0.222	0.230	-0.007	-0.024
Pain Sev		0.601	0.241	0.260	0.231	0.251	0.032	-0.031
Night Pain			0.304	0.290	0.156	0.186	-0.023	-0.035
Sit Down				0.954	0.449	0.433	0.259	0.213
Rise Up					0.453	0.445	0.260	0.206
Going U						0.928	0.382	0.286
Going D							0.356	0.296
Hip Ab								0.718

Table 3.3: Upper triangle of the correlation matrix for eight of the post-operative knee variables.

		Diagnosis	
		OA	RA
Operation	Cemented	319	31
	Uncemented	211	18

Table 3.4: Contingency table for Diagnosis and Operation.

(Fisher's exact test is inappropriate here due to the large sample sizes). Considering the variables *Diagnosis* and *Operation* in the knees data we obtain the 2×2 contingency table in Table 3.4. Since the data were obtained as part of a randomised trial with operation being assigned at random, performing a test for independence here would only serve to 'verify' the independence that we would expect in a randomised trial. Nonetheless, using Yates' correction we obtain a value of $\chi^2 = 0.0722$ on 1 degree of freedom leading to a p-value of 0.7881. Therefore the independence hypothesis cannot be rejected and we conclude that the way the patient is treated is independent of their pathology as would be expected in such a randomised study.

3.2.8 Comparing Subgroups

The effects of different levels of the categorical variables can also be assessed by comparing the means for the continuous variables with those levels. The pre-operative

means of the measurements for patients who were diagnosed with osteoarthritis (OA) those who diagnosed with rheumatoid arthritis (RA) are given in the first two columns of Table 3.5. The mean values were then compared using an independent sample t-test, the results of which are displayed in the remaining columns of Table 3.5 where significant p-values at the 95% level are coloured red. This shows that the mean for patients with osteoarthritis is significantly different than the mean for patients with rheumatoid arthritis for many measurements. This difference is primarily on the general walking ability scores, but also includes *Weight* which was previously identified as a potentially discriminating variable in Section 3.2.2 and Figure 3.3. However, when repeating the tests with the post-operative data, we find that there are only three significant differences between diagnoses - those being in *Weight*, *Walking Ability* and *Flexion*. This may suggest that there is little distinction post-operatively between these two groups. However, it would be prudent to observe at this point that some results may be spurious. When performing so many hypothesis tests, we would expect some spurious findings.

Repeating the tests for the post-operative means of the measurements for patients who received the 'Cemented' treatment and those who received the 'Uncemented' treatment revealed no evidence of significant differences in the effects of the two treatments one year after the treatment. There were also no significant differences between the pre-operative means, however in this case the lack of distinction is likely due to the randomisation over treatment.

The post-operative means can also be compared to the pre-operative means by a paired-sample t-test as in Table 3.6. For the knees data this shows a highly significant change in the direction of improvement for all measurements except *Flexion*, *Extension Lag* and *Other Knee Flexion*. However, the validity of the result for *Extension Lag* is likely questionable due to its profound skewness.

The papers by Gregoire and Driver [56] and Rasmussen [103] suggest that the application of t-tests to ordinal data is a sensible approach in spite of the data not being continuous. As discussed in Section 3.1, to retain the categorical nature of these variables would render analyses of such as those performed above either infeasible or impossible due to dimension constraints. Therefore, these methods are

	Diagnosis			
	OA	RA	$\lvert t\rvert$	p
Weight	75.848	65.531	5.738	$< 10^{-4}$
Pain Frequency	2.064	1.898	1.690	0.0458
Pain Severity	2.136	2.041	1.026	0.1527
Night Pain	3.053	2.918	0.732	0.2324
Stability	2.801	2.993	1.109	0.1339
Walking Abililty	3.211	2.616	4.519	$< 10^{-4}$
Walking Aids	4.245	3.878	3.591	0.0002
Sitting Down	3.601	3.072	2.621	0.0045
Rising Up	3.522	3.071	2.239	0.0128
Standing	2.865	2.629	1.096	0.1368
Going Up Stairs	2.680	1.931	3.371	0.0004
Going Down Stairs	2.555	1.867	3.153	0.0008
Ctf Varus	7.227	−1.704	5.433	$< 10^{-4}$
Fixed Contracture	8.833	13.673	4.079	$< 10^{-4}$
Flexion	106.826	105.490	0.489	0.3126
Extension Lag	0.502	2.005	2.817	0.0025
Hip Abduction	29.591	26.415	2.497	0.0064
Other Knee Fixed Cont.	4.254	4.168	0.108	0.4570
Other Knee Flexion	116.193	115.309	0.338	0.3677
Other Hip Abduction	31.544	30.353	0.971	0.1660

Table 3.5: Comparison of the pre-operative means for the different levels of Diagnosis.

	$\overline{D}$	$s_{\overline{D}}$	$\lvert t\rvert$	p
Weight	1.321	0.184	7.171	$< 10^{-4}$
Pain Frequency	2.399	0.043	56.421	$< 10^{-4}$
Pain Severity	2.415	0.041	59.104	$< 10^{-4}$
Night Pain	1.703	0.054	31.279	$< 10^{-4}$
Stability	1.869	0.056	33.281	$< 10^{-4}$
Walking Abililty	0.745	0.037	20.124	$< 10^{-4}$
Walking Aids	0.146	0.031	4.704	$< 10^{-4}$
Sitting Down	1.206	0.061	19.733	$< 10^{-4}$
Rising Up	1.251	0.061	20.400	$< 10^{-4}$
Standing	1.455	0.062	23.534	$< 10^{-4}$
Going Up Stairs	1.338	0.063	21.347	$< 10^{-4}$
Going Down Stairs	1.412	0.065	21.684	$< 10^{-4}$
Ctf Varus	−5.935	0.451	13.157	$< 10^{-4}$
Fixed Contracture	−6.399	0.317	20.211	$< 10^{-4}$
Flexion	−0.627	0.744	0.842	0.2000
Extension Lag	−0.185	0.205	0.901	0.1840
Hip Abduction	3.526	0.327	10.783	$< 10^{-4}$
Other Knee Fixed Cont.	−1.394	0.190	7.322	$< 10^{-4}$
Other Knee Flexion	−0.198	0.462	0.428	0.3343
Other Hip Abduction	1.315	0.317	4.156	$< 10^{-4}$

Table 3.6: Comparison of the pre-operative means to post-operative means for the knees data.

applied in a pragmatic fashion to the data assuming that it is continuous.

It is clear from the analyses performed and results presented in this section that there is a wealth of information in this data set. Consequently, there is a huge mass of results that one might calculate and those given in this section form only a small portion of that total. It is clear that what is required here is a set of efficient methods to summarise and present the key features of the data; such methods will be introduced in the next chapter on visualisations.

3.3 The Hips Data

3.3.1 Boxplots and Histograms

As the hips data is composed almost exclusively of discrete data, the use of boxplots are inappropriate. However, histograms and mean and error bar plots of the 12 repeated ordinal pre-operative measurements display a similar degree of asymmetry and lack of Normality as the knees data (see Figure 3.6). There also appears to be some evidence of bimodality for some variables. As with the knees data, the post-operative values display a change of location in the direction of an improvement in patient condition. Again, this results in a strong asymmetry and a skewness of the data in this direction.

3.3.2 Scatterplots

Scatterplots of the jittered pre-operative variables display some weak positive associations between the variables, though there were no tightly associated pairs or groups as with the knees data. Scatterplots of the post-operative variables against their pre-operative counterparts displayed no evidence of association for the ordinal variables.

3.3.3 Clusters

Clustering on the hips data by PAM also yielded little evidence for cluster-type behaviour. The results for $k = 2$ is shown in Figure 3.7 and, whilst the cluster

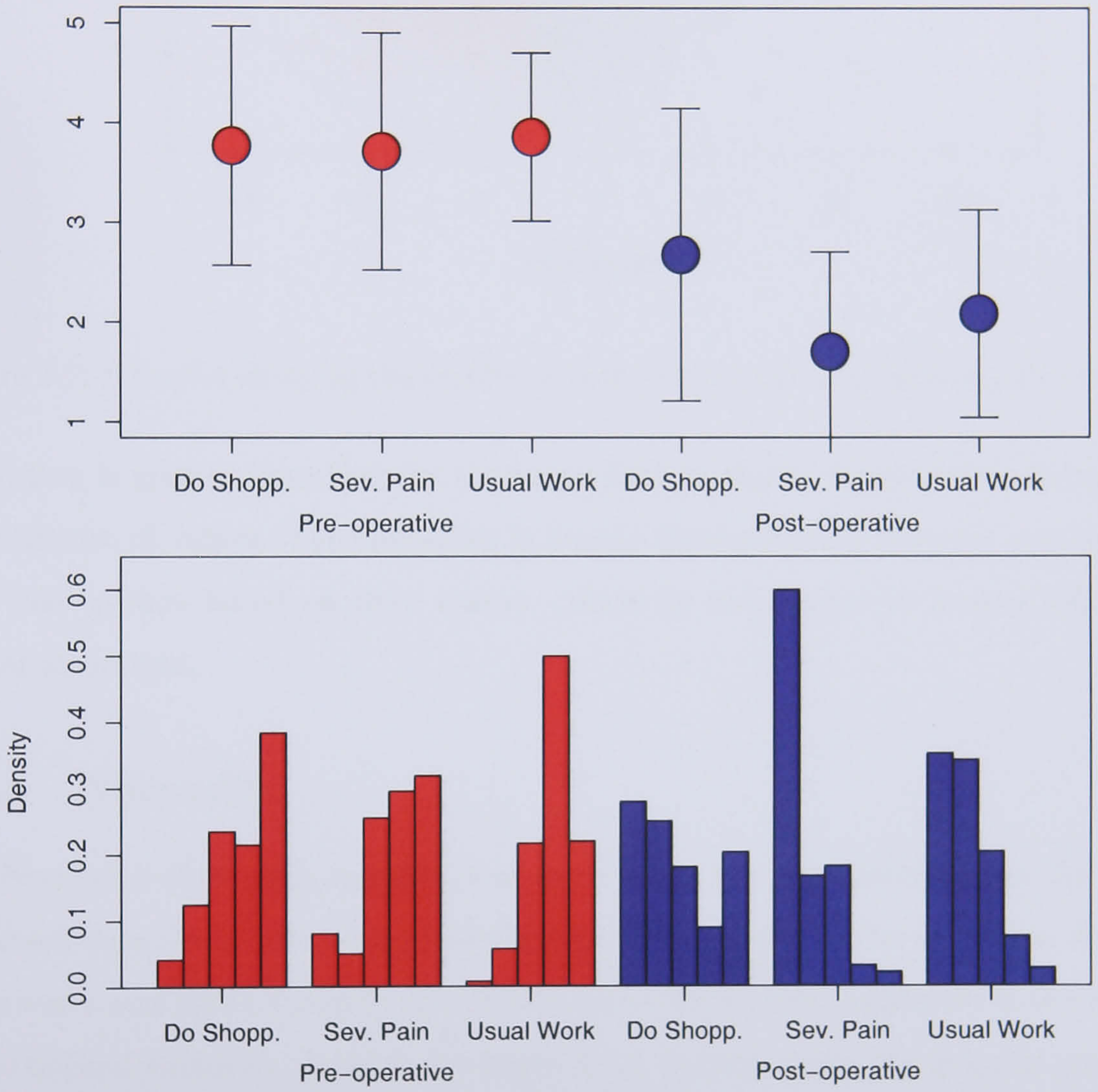


Figure 3.6: Mean and error bar plots and histograms for three ordinal variables in the hips data measured pre- and post-operatively.

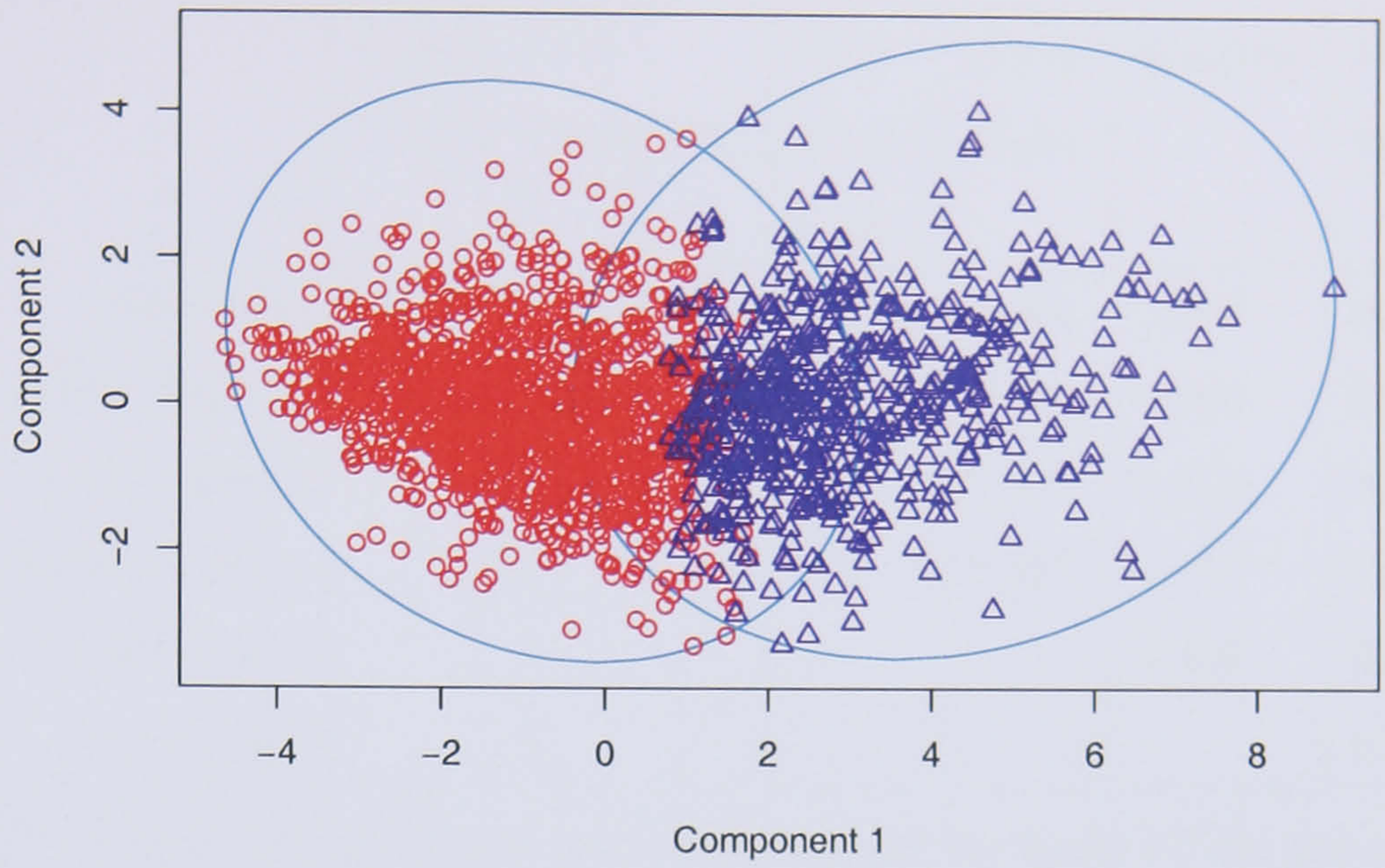


Figure 3.7: Clusplot showing the results of partitioning the hips data into 2 clusters.

separation is greater than that for the knees data, it shows that the data are again a continuum of values. The clustering is merely discriminating patients into 'good' and 'bad' groups based on their status - there do not appear to be any intrinsic cluster structures.

3.3.4 Normality

The Normality of the hips variables was assessed via Normal quantile plots. As with the knees data, these plots displayed the lack of Normality of the variables. Again, there was a pronounced step pattern to the plots due to the boundaries of the levels of the ordinal variables. As with the knees data, in these cases the quantile plots of the ordinal variables are not a useful tool since we know a priori that the variables are non-Normal. Application of Box-Cox transformations to the data made little improvement to the Normality of the variables.

3.3.5 Means and Standard Deviations

Examining the means for the hips data reveals that the mean scores are typically quite high with values from 3–5 indicating a poor patient condition. The corresponding standard deviations are between 0.7 and 1.2 which is relatively small and indeed

	Using Transp.	Do Shopping	Walk w/out Pain	Climb Stairs	Limping	Usual Work
Usual Pain	0.438	0.435	0.177	0.434	0.289	0.490
Using Transp.		0.527	0.222	0.552	0.290	0.498
Do Shopping			0.310	0.585	0.304	0.574
Walk w/out Pain				0.267	0.127	0.263
Climb Stairs					0.309	0.532
Limping						0.405

Table 3.7: Upper triangle of the correlation matrix for seven of the pre-operative hips variables.

smaller than the value for the knees data. This may suggest that pre-operatively the status of the patients is consistently poor. The respective post-operative means fall between 1 and 3 which, when compared with the pre-operative scores, shows a drop of between zero and four points. Post-operative standard deviations are largely similar to their pre-operative equivalents. This suggests that there is an improvement in patient condition, but there is still a similar degree of variation amongst the conditions of the patients themselves.

3.3.6 Correlations

The hips data appear to be much more strongly correlated than the knees data with correlations of around 0.37 (see Table 3.7). The correlation appears also to be more uniform than with the knees data with all variables having a similar degree of association. Consequently, there does not appear to be any evidence of the apparent correlation structures which are present in the knees data. The correlations for the post-operative data are also similar, with all variables being moderately correlated with one another. However, the level of correlation is slightly more than that pre-operatively with a typical value of around 0.43. This could suggest that the noise evident pre-operatively has been removed or reduced.

Unlike with the knees data, inspection of correlations for the hips reveal that the

post-operative variables are almost entirely uncorrelated to the pre-operative values with $r_{ij} < 0.07$ for all variable pairs. This suggests that the status of a patient after the operation has little connection to their status before, implying that the hip replacement affects patients in a way that is independent of their initial state.

As with Table 3.2, the correlation matrix here is hard to interpret on its own since it requires the simultaneous reading and comparison of multiple values in the table to determine the nature of the associations. Presenting information in the form of a table of numbers can be an inefficient method, especially when the goal is to communicate results to non-experts. It is clear here that a more efficient representation for these correlation matrices would be useful in this situation.

3.3.7 Discrete Associations

The hips data contains many discrete variables which could be tested for independence, however we shall focus on only a small subgroup. There are five binary variables to be examined: the first two represent the presence or absence of osteoarthritis or rheumatoid arthritis in the patient's primary pathology (and so approximately correspond to the *Diagnosis* variable in the knees data); the second two variables represent where cement was used on femoral or acetabular prostheses (corresponding to *Operation* in the knees data); and the final variable represents whether the patient is private or NHS.

Using Pearson's χ^2 test with Yates' continuity correction on the contingency tables we reject the independence hypothesis for the pathology variables ($p < 2.2 \times 10^{-16}$). Examining the contingency table suggests that the presence or absence of osteoarthritis is associated to the presence of absence of rheumatoid arthritis with more patients than expected having only one of the pathologies. We similarly reject the hypothesis of no association when examining the two cement variables suggesting that the use of cement on one prosthesis is not independent of the use of cement on the other ($p < 2.2 \times 10^{-16}$) with more patients than expected receiving cement on only one of the prostheses.

We fail to reject the independence hypothesis when considering the associations between the patient's private status and their pathology, as well as between their

private status and the use of cement. These results are not surprising, as we would expect the patient's pathology and treatment to be unassociated with their private status.

When investigating the interactions between pathologies and use of cement we find that we reject the independence hypothesis for osteoarthritis and the use of cement for both prostheses ($p \leq 1.2 \times 10^{-4}$). In this case we find fewer than expected people with osteoarthritis received cement on the acetabular prosthesis, and more than expected received cement on the femoral prosthesis. The same is not true for rheumatoid arthritis, however, and we fail to reject the hypothesis of no association for all variables.

3.3.8 Comparing Subgroups

As with the knees data we can perform independent sample t-tests to give us an initial impression of whether there appear to be any notable differences between subgroups of the data. The hips data has several categorical variables which define a number of groups. The first to be examined here is that which defines whether the patient's primary pathology includes osteoarthritis or not. The results are presented in Table 3.8 in the same style as for the knees data. We can see that majority of variables are significantly different between these two groups. This is with the exceptions of *Usual Pain* and *Standing Without Pain* for which there is insufficient evidence to conclude that there are differences between the groups. For the variables which display a significant difference between these two pathology groups, we can see that it is typically the group without osteoarthritis that exhibit lower mean values and hence appear to have a more favourable average state.

This procedure was repeated as before to compare the groups where the patient's pathology includes rheumatoid arthritis and those who do not. We obtain similar results which are not tabulated here for brevity, however they demonstrated significant differences on all variables with the group having rheumatoid arthritis being in a typically poorer state pre-operatively.

Another variable that was investigated for these pairwise differences was the variable which records whether the patient was treated privately or on the NHS. The

results for these comparisons are given in Table 3.9. Again we observe significant differences on all variables suggesting that there is a difference between the average patient in each group. Further examination show that, a priori, the patients who were treated privately have lower values on these variables indicating that their state was better than the NHS patients. This is likely due to waiting list for NHS treatment resulting in the patient's pathology being in a slightly more advanced (and hence poorer) state when they are seen by the consultant.

The post-operative means can also be compared to the pre-operative means by a paired-sample t-test as in Table 3.6. As with the knees data, we observe a highly significant change in the direction of improvement though in this case this improvement is evident on all of the patient status variables.

	No OA	OA	$\lvert t \rvert$	p
Usual Pain	1.428	1.494	1.795	0.0364
Washing	2.694	2.883	3.556	0.0002
Using Transport	2.482	2.566	2.085	0.0186
Put On Socks	2.219	2.367	2.719	0.0033
Do Shopping	1.906	2.264	5.351	$< 10^{-4}$
Walk W/out Pain	2.997	3.259	3.166	0.0008
Climb Stairs	2.368	2.757	4.127	$< 10^{-4}$
Stand W/out Pain	2.371	2.445	1.609	0.0539
Limping	1.278	1.469	4.573	$< 10^{-4}$
Severe Pain	2.122	2.298	2.643	0.0041
Usual Work	1.889	2.163	5.813	$< 10^{-4}$
Night Pain	1.791	1.944	2.519	0.0059

Table 3.8: Comparison of the pre-operative means for patients whose pathology did and did not include osteoarthritis.

	NHS	Private	$\lvert t \rvert$	p
Usual Pain	1.652	1.444	7.680	$< 10^{-4}$
Washing	3.146	2.791	9.039	$< 10^{-4}$
Using Transport	2.779	2.500	9.438	$< 10^{-4}$
Put On Socks	2.654	2.274	9.444	$< 10^{-4}$
Do Shopping	2.702	2.105	12.146	$< 10^{-4}$
Walk W/out Pain	3.377	3.195	2.941	0.0016
Climb Stairs	2.874	2.474	10.747	$< 10^{-4}$
Stand W/out Pain	2.703	2.369	9.795	$< 10^{-4}$
Limping	1.605	1.410	6.273	$< 10^{-4}$
Severe Pain	2.556	2.209	7.029	$< 10^{-4}$
Usual Work	2.405	2.065	9.748	$< 10^{-4}$
Night Pain	2.120	1.879	5.317	$< 10^{-4}$

Table 3.9: Comparison of the pre-operative means for NHS and private patients.

Chapter 4

Visualisations

The use of graphical methods to present data and to communicate the statistical features thereof has a long history with Tukey [121], Tufte [119, 120], Cleveland [18, 19], and more recently Wilkinson [126] having written extensively on the subject. The usefulness of such methods is undisputed; as Tufte says: 'Graphics *reveal* data'. By showing a great deal of information in a relatively small space, large and complex data sets can be made more 'coherent' and easier to understand and interpret. Since the orthopædic data sets we have examined thus far are typically high dimensional and also incorporate small time series elements, the application of appropriate visual methods to gain insight into the data would be a sensible and prudent course of action. In particular, the goal here is to present a large amount of information succinctly, but meaningfully to non-expert clinicians.

This chapter introduces three visual methods for displaying either summary statistics of the data or the results of statistical tests in a way to enhance their interpretability. Section 4.1 introduces a very simple visualisation of the results of multiple t-tests for the differences in the means of two groups of data. Following on from this, in Section 4.2 a method for illustrating the correlations amongst groups of variables is presented, which was originally proposed by Friendly [50]. Then in Section 4.3 a methodology inspired by profile analysis is introduced to depict the changes in the mean values of all variables over time, which incorporates a suitable standardisation for all variables and to facilitate comparisons between the profiles of multiple groups. Finally, the chapter ends with some comments and remarks on

the methods presented.

4.1 t-Test Plots

The comparison of the means of two groups of data is one of the most basic tests within statistics. The application of the standard independent and paired samples tests using Normal or t distributions is covered in many introductory texts on statistics [46, 104, 12]. As such, these tests should be familiar to any performing even the most basic of statistical analyses. However when performing many such tests on data sets containing several variables the results will take the form of tables such as Table 3.5 which can be both daunting and potentially difficult to interpret. In this section a simple visualisation of these results is presented and illustrated with examples.

4.1.1 Methodology and Results

A very simple visualisation of the results of a t-test, such as those in Table 3.5 could be presented by a simple bar chart where each bar represents the t value for the two-sample test for each variable. The reasons for plotting the raw t values and not the associated probabilities are twofold. Firstly, large differences between the groups result in large ts with small corresponding significance probabilities - this reversal would be detrimental to interpretation as the reader associates large changes with large values. Secondly, the t value itself provides information on the direction of the difference, whereas this information is unavailable when considering probabilities alone. For the orthopædic data the results will be comparable across variables since the sample sizes are the same for all measurements, which makes the degrees of freedom constant. The original t statistic is plotted rather than $|t|$, since the direction of the deviation from zero is also informative. Furthermore, those t values which exceed the bounds for significance can be coloured to draw attention to these significant deviations. However as an alternative to plotting t-values, one could plot a function of the associated p-values such as $1-p$, $-\log p$ or even $\text{sign}(t)|\log p|$. However, for the purposes of this graphic, we shall be considering only the t-values.

Subject to the usual practical difficulties of multiple significance testing such as assessment of normality, the independent sample t-test would be of use when analysing the orthopædic data as it will enable the comparison of means of measurements between two separate groups, such as different diagnoses and treatments. We assume a sample $X_1, \ldots, X_n$ is drawn from a normal distribution that has mean μ_X and variance σ^2, and an independent sample $Y_1, \ldots, Y_m$ is drawn from another normal distribution that has mean μ_Y and variance σ^2. To ascertain whether the two samples were drawn from the same population, we test the null hypothesis that the two samples have the same mean, i.e. $\mu_X = \mu_Y$ and hence $\mu_X - \mu_Y = 0$. Any evidence of deviation from this hypothesis is illustrated by significant values of the t-statistic:

$$t = \frac{(\overline{X} - \overline{Y})}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where s_p is the pooled sample variance:

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$$

and s_X^2 and s_Y^2 are the sample variances for X and Y respectively. The t statistic follows a t distribution with $m + n - 2$ degrees of freedom.

The t-tests in Table 3.5 assess the differences in the pre-operative means for the different levels of diagnosis for the knees data. These tests are illustrated via a t-test plot in Figure 4.1, where the difference is of the form rheumatoid arthritis patients minus osteoarthritis patients. Thus positive t values indicate that the mean for a patient with rheumatoid arthritis is larger than the corresponding value for a patient with osteoarthritis. The covariate *Weight* is included in this graph with the other response variables since we discovered in Chapter 3 that it could potentially discriminate *Diagnosis*. Since the sample sizes are fixed for both groups, all t values have the same number of degrees of freedom thus ensuring they are all directly comparable. It should also be noted that the majority of these variables are ordinal rather than discrete, which may pose a potential problem for the application of t tests which assume Normality of the data. However, [103] illustrated that application of the t test to such data is still reasonable since the ordinal variable can be viewed as a discretised version of a latent continuous quantity.

Significant t values are coloured red to add emphasis to those variables; the two horizontal lines correspond to the appropriate critical value of the t-distribution. Significant differences can be seen for the variables *Weight*, *Coronal Tibio-Femoral Varus* (CTFVAR) and *Fixed Contracture* (FCONT). In general, we can see that patients with osteoarthritis appear to be heavier and have better levels of general mobility. There appears to be little difference between the two sets of patients in terms of pain with *Pain Severity* (PAINS) and *Night Pain* (PAINN) non-significant and *Pain Frequency* (PAINF) just scraping past the significance threshold in favour of the osteoarthritis group.

Repeating the same procedure to compare differences between the two types of operations gives us the plot displayed in Figure 4.2. This plot illustrates that there is very little distinction between the two forms of treatment since all of the scores are non-significant. It should be noted that since this boundary is entirely arbitrary and given the problems with the validity of the t-test assumptions the true 'significance' of the difference in these cases is somewhat tenuous.

When the two samples are not independent, such as comparing pre- and post-operative measurements for the same patients, we must turn to the paired sample t test. In this case, each X_i is paired with a corresponding Y_i and we work with the differences $D_i = X_i - Y_i$, which we assume follow a normal distribution with $\mathrm{E}[D_i] = \mu_X - \mu_Y = \mu_D$ and $\mathrm{Var}[D_i] = \sigma_D^2$ where σ_D^2 is the variance of the population of differences. Since σ_D is generally unknown, the t statistic is given by:

$$t = \frac{\overline{D} - \mu_D}{s_{\overline{D}}}$$

where $\overline{D}$ is the sample mean of the differences, and $s_{\overline{D}}$ is the sample standard deviation of $\overline{D}$ which is defined as $s_{\overline{D}} = \sqrt{s_D^2/n}$ where s_D is the sample standard deviation of the D_i. This then follows a t distribution with $n-1$ degrees of freedom.

Table 3.6 displays the results of performing such a test to compare the post-operative means to the pre-operative means for the knees data by setting $\mu_D = 0$. The associated t-test plot is presented in Figure 4.3 where t values outside the interval $[-6, 6]$ have been truncated and are indicated by the symbols $\triangle$ or $\triangledown$. We can see immediately from this plot that there is overwhelming statistical evidence of a substantial improvement in the patient condition post-operatively for all but three

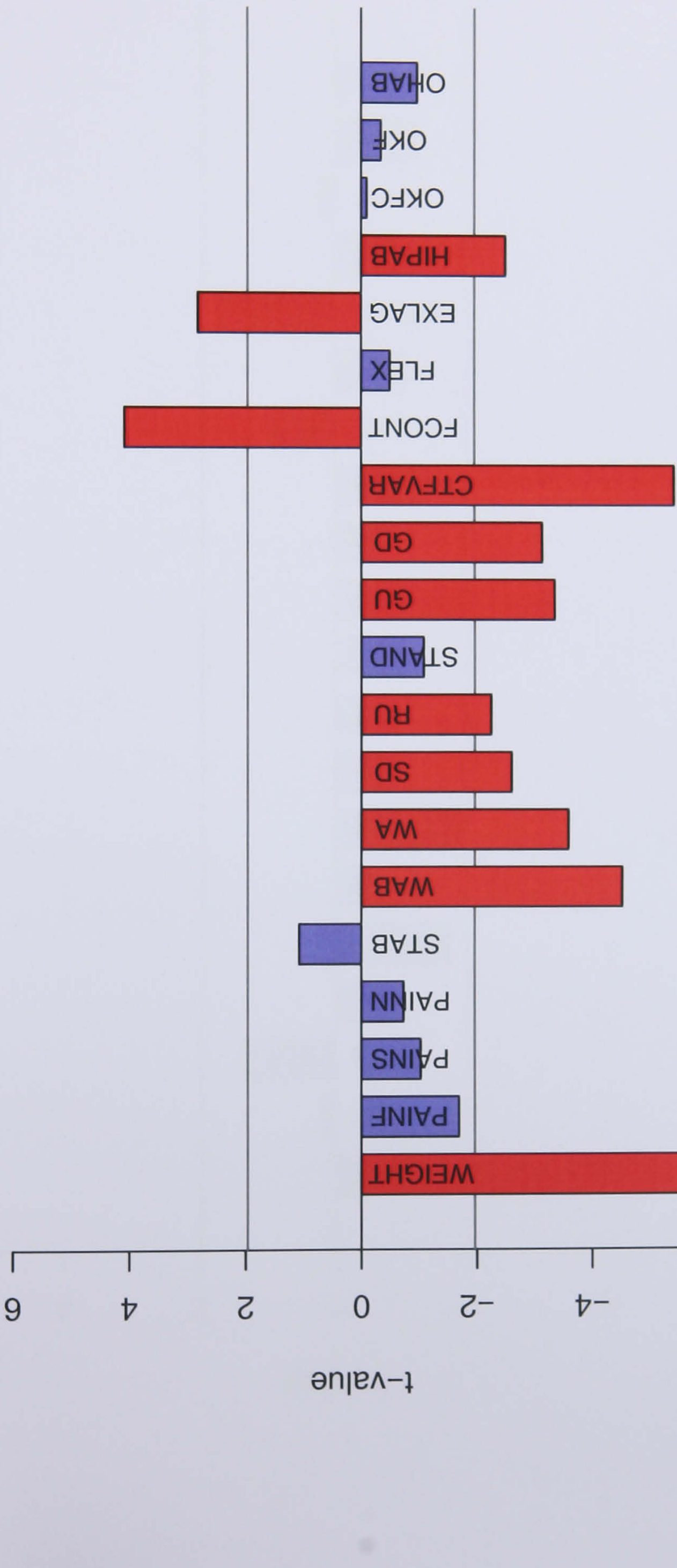


Figure 4.1: t-test plot assessing the differences in the pre-operative means the two diagnoses (RA - OA) in the knees data.

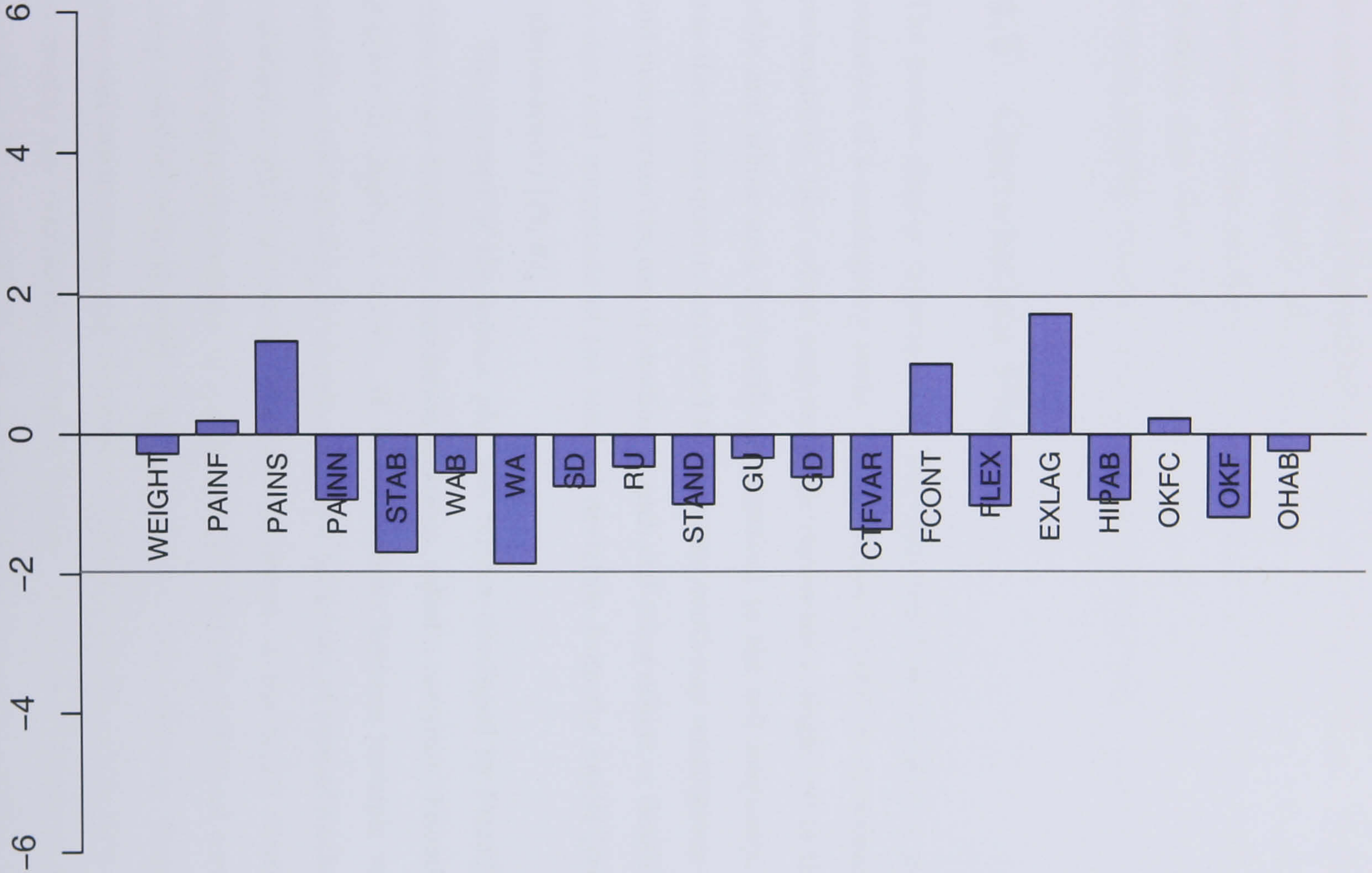


Figure 4.2: t-test plot assessing the differences in the pre-operative means the two operations (Cemented - Uncemented) in the knees data.

variables. Only *Flexion* (FLEX), *Extension Lag* (EXLAG) and *Other Knee Flexion* (OKF) display no evidence of change over the intervening period. However, it should be noted that whilst the statistical significance of the difference from before to after the operation is great, the corresponding practical change is typically only slightly more than a one-point increase on a scale of 1-5 with *Walking Ability* (WAB) and *Walking Aids* (WA) being somewhat less, and *Pain Frequency* (PAINF) and *Pain Severity* (PAINS) showing more than a two-point average increase.

4.2 Correlation Plots

The mosaic display, introduced by Hartigan and Kleiner [58] is a graphical representation of a contingency table. The contingency table is represented by a grid of rectangles or 'tiles' where each rectangle represents a single cell of the contingency table and whose area is directly proportional to the cell frequency. This graphic was then subsequently extended to illustrate multi-way contingency tables [59, 48] and incorporate the use of shading in each cell using colour or density proportional to sign and magnitude of the residual deviation from the model (typically that of independence) [49, 62].

The concept of the mosaic plot was further developed by Friendly [50] into an exploratory display for correlation matrices, called a *corrgram*. Friendly's corrgrams graphically display a matrix, $\boldsymbol{R}$, of correlations between multiple variables. Each pairwise correlation $\boldsymbol{r}_{ij}$ is visually depicted using one of several methods in a larger rectangular grid framework reflecting the layout of the larger correlation matrix. Friendly presents a variety of means of displaying the individual correlation values using coloured cells as with a mosaic plot, or using different shapes of different sizes and orientations and alternative shading methods. Friendly's use of ellipses to display the correlation matrix is similar to the glyph-based plots of Murdoch and Chow [91], though the use of colour is likely an easier method to interpret the correlations than elliptical symbols. For our purposes however, we shall focus on the corrgram which is a simple variation of the mosaic plot which we shall refer to as a 'correlation plot' to distinguish it from Friendly's other possible representations.

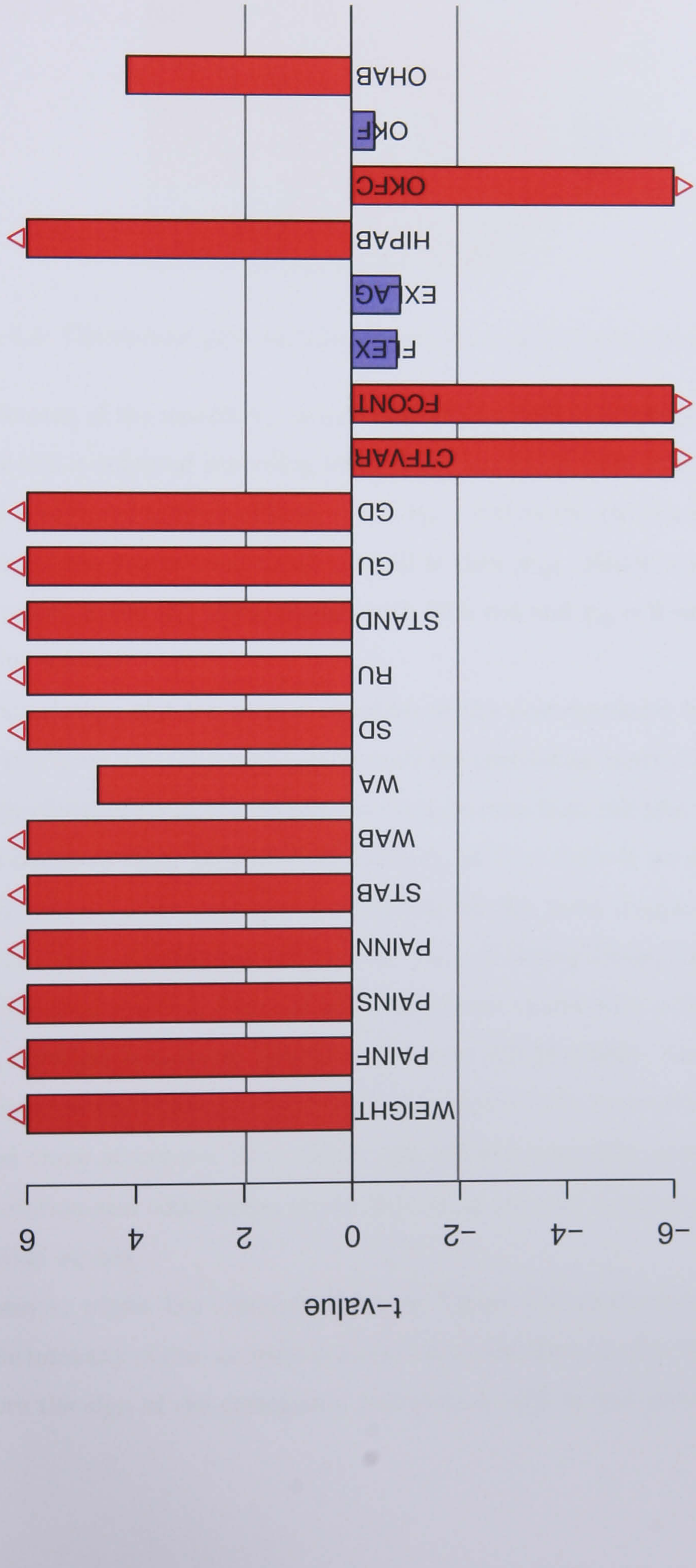


Figure 4.3: t-test plot assessing the differences between the pre- and post-operative means in the knees data.

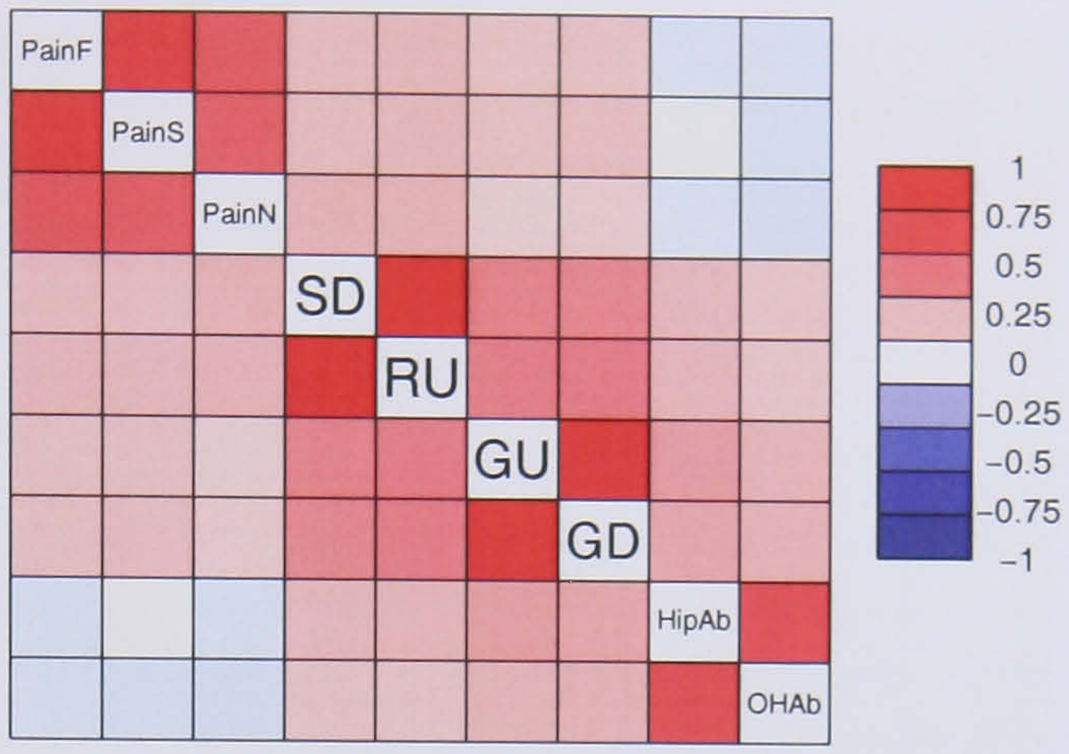


Figure 4.4: Correlation plot for selected variables of the post-operative knees data.

Each element of the matrix, $\boldsymbol{r}_{ij}$ is represented by a cell in the larger rectangular grid and the cell is coloured according to the sign and magnitude of $\boldsymbol{r}_{ij}$. If $\boldsymbol{r}_{ij} < 0$ then the cell is coloured a shade of blue, and if $\boldsymbol{r}_{ij} > 0$ then the cell is a shade of red. The intensity of the colour used to fill the cell is then $|\boldsymbol{r}_{ij}|$. Hence a value of $\boldsymbol{r}_{ij} = 0.5$ would result in the cell being shaded with 50% red and $\boldsymbol{r}_{ij} = 0$ would result in the cell being white.

A correlation plot for several variables of the post-operative knees data is presented in Figure 4.4 - this plot corresponds the correlation matrix in Table 3.3. The presence of structure in the correlations can be seen from the plot. There are slight modifications to those presented by Friendly, in that there is no diagonal hatching to each cell and that variables are labelled on the main diagonal of the matrix. The very strong correlations between the pairs of *Sitting Down/Rising Up* (`SD/RU`), *Going Up Stairs/Going Down Stairs* (`GU/GD`) are visible with a lesser, though still strong, correlation between the hip abductions (`HIPAB/OHAB`). The strong relationships between the three pain scores are also visible. The correlation plot has easily revealed these structures in a simple and efficient way that does not require the interpretation and comparison of the individual pairwise correlations in the overall correlation matrix.

However, whilst the correlation plot in Figure 4.4 may convey information on both the intensity of the correlation as well as its direction, it may be more reasonable to ignore the sign of the correlation and plot all cells in the same colour. The use

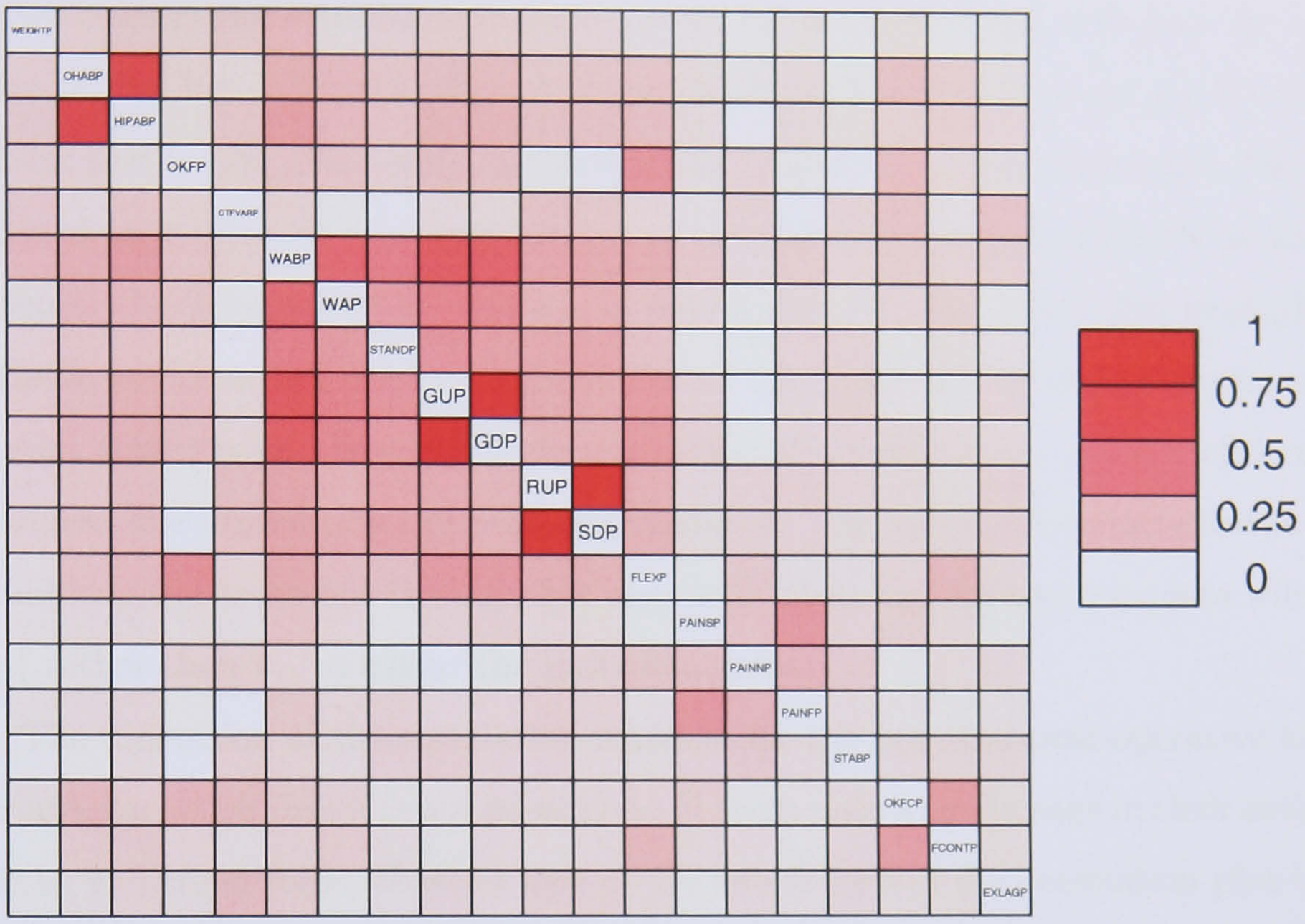


Figure 4.5: Absolute value correlation plot for the absolute values of the correlations the pre-operative knees data.

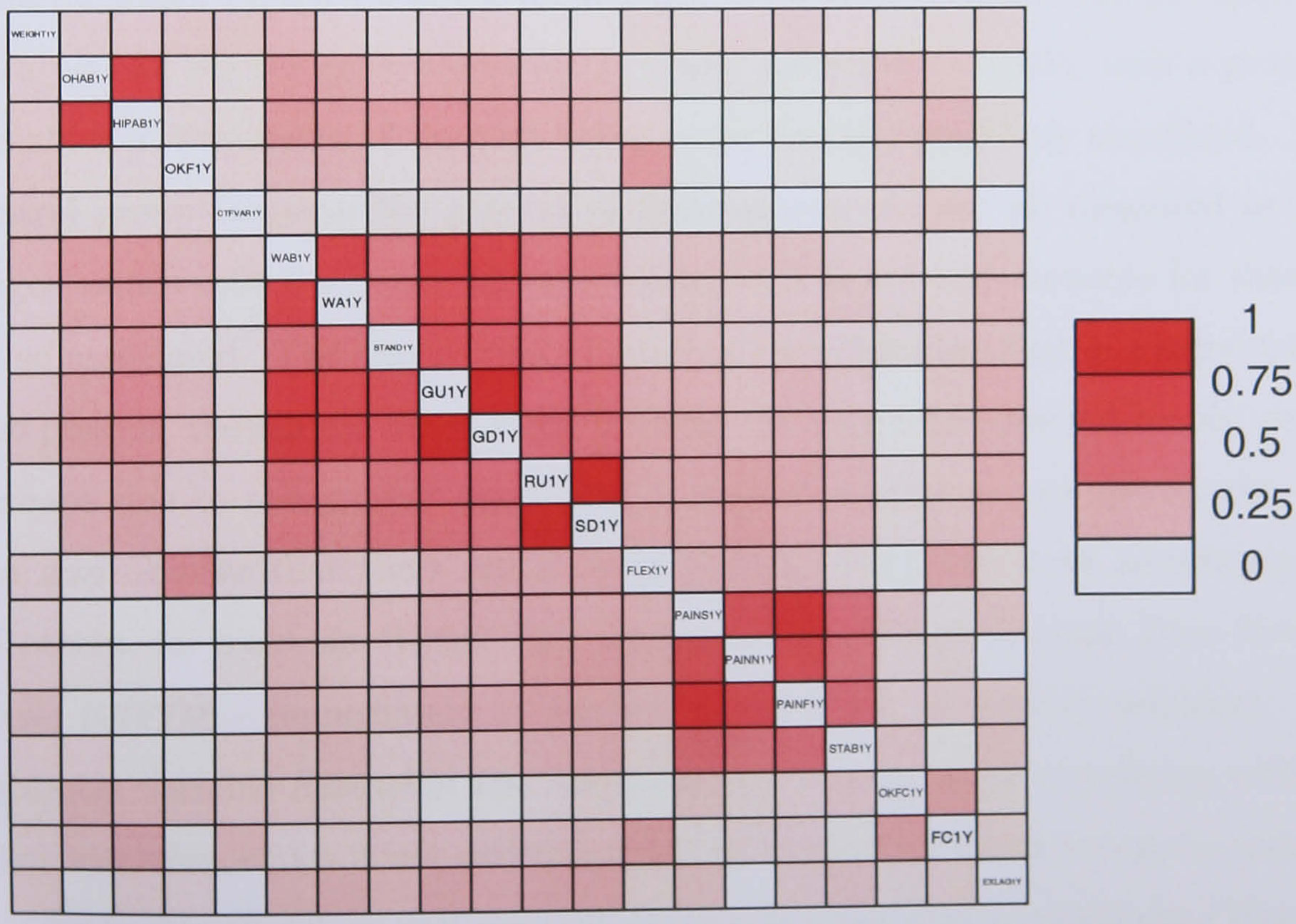


Figure 4.6: Absolute value correlation plot for the post-operative knees data.

of two colours can complicate interpretation, for example if two cells had the same value of $|r_{ij}|$ but differed in sign, it is hard to determine that they are equal or even similar due to the differences imposed by the colours. Furthermore, since the sign of the data is often an arbitrary artefact of the choice of the data collector or survey designer the information it contains is potentially of little value. For example, a variable could record a numeric indicator of the levels of the patient's pain on a 5-point Likert scale. Depending on the choice of the physician, a value of 5 could represent the highest level of pain or the lowest. The information contained in the variable is the same but the sign has reversed. Therefore subsequent plots will use $|r_{ij}|$ rather than r_{ij} to colour the individual cells.

The dimension of the correlation matrices for the pre- and post-operative knees variables are such that it is not possible to fit them onto a single page in their entirety and in numerical form. However one of the advantages of the correlation plot is its compactness, which enables the presentation of these correlation matrices albeit in an alternative form. Figure 4.5 shows the complete correlation matrix between the pre-operative knees variables, and Figure 4.6 shows the same for the post-operative data (of which Figure 4.4 was a submatrix). If we firstly consider the pre-operative data we can see the correlations are typically quite low (< 0.25), with a group of variables in the centre of the plot being more strongly positively correlated. This central group compose the general mobility scores and are all measured so that larger values reflect a better patient status, so it is not unreasonable for these to be so associated. The closely related pairs of variables identified in Figure 4.4 are also present, though the associations between the pain scores are noticeably weaker perhaps due to the greater variability in patient conditions pre-operatively. We can also observe that there are a few variables with fairly weak associations to all others, for example *Weight* (the leftmost column) and *Coronal Tibio-Femoral Angle* (`CTFVAR` - immediately to the left of the block of central variables). The rightmost variable *Extension Lag* also illustrates a fairly weak correlation with the other variables, which is not surprising due to most of its values being the same.

Comparing to 4.6 we can see that there is a great deal of similarity. However the intensity of the correlations post-operatively seems to have increased - the pain

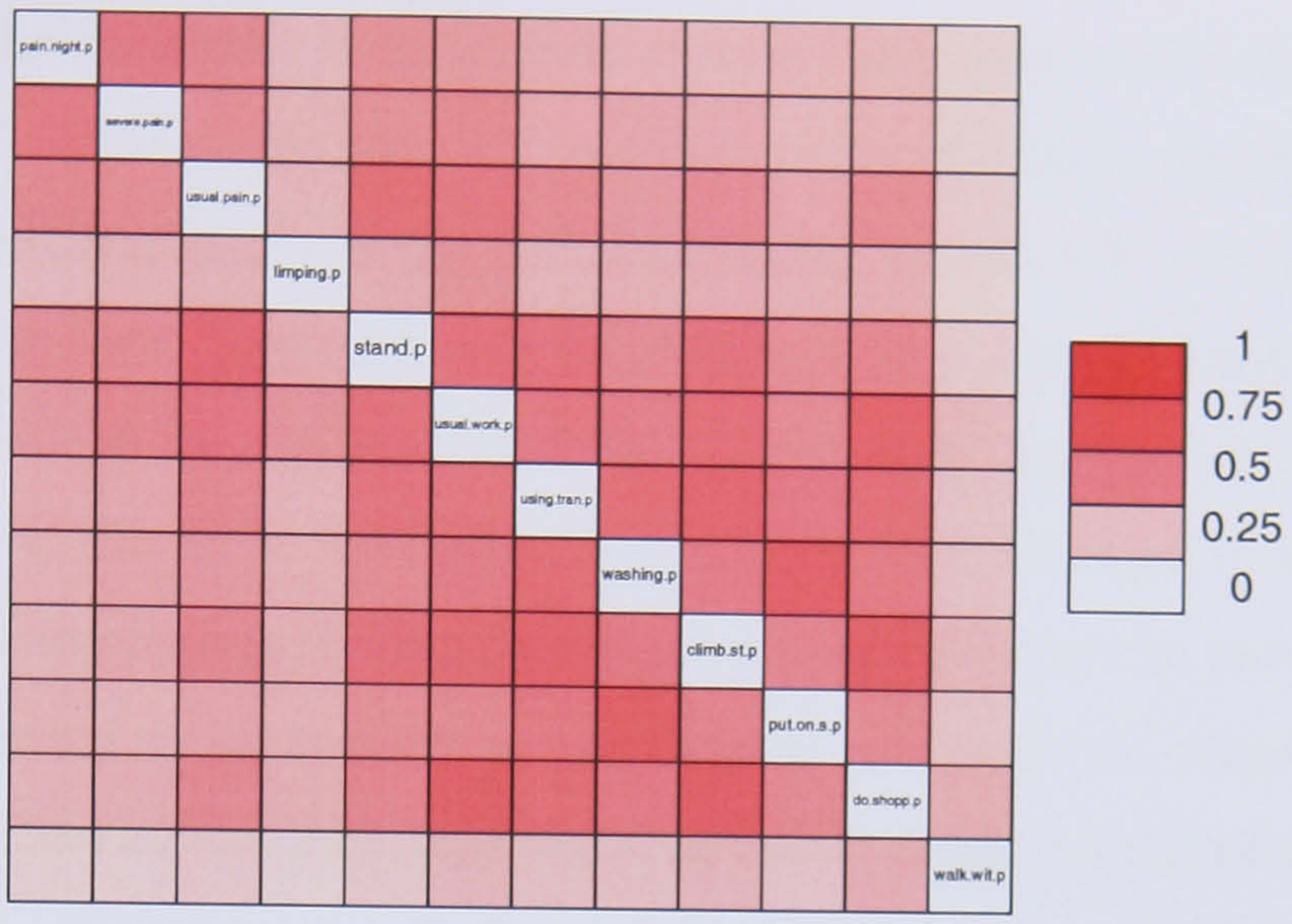


Figure 4.7: Correlation plot for the pre-operative hips data

scores in the bottom-right are now more tightly correlated and are now slightly more correlated to the other mobility measures. This is likely due to the effect of the treatment having some form of a unifying effect across patients, reducing the variability present pre-operatively. The other variables however seem to exhibit little change, though the variable measuring *Coronal Tibio-Femoral Angle* now appears to be slightly less correlated to other variables.

The variables in Figures 4.5, 4.6 and 4.7 have been ordered according to the scheme proposed by Friendly [50] where the variables are arranged in the angular order of the first two eigenvectors of the correlation matrix. The order of the variables is determined from the order of the angles α_i:

$$\alpha_i = \tan^{-1}(e_{i2}/e_{i1}),$$

where $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are the first two eigenvectors of $\boldsymbol{R}$. Highly correlated variables will have similar locations in the biplot of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ and so will have similar values of α_i. This ordering thereby seeks to arrange the variables in a manner that will expose the underlying correlation structure. For the correlation plots of the knees data the ordering over the post-operative variables was used to arrange both plots so that they would be directly comparable. To linearise the angles, the circle is split at the point where the separation between two adjacent α_i is greatest.

Repeating the plots for the hips variables provides slightly less insight into the

structure of the variables. A correlation plot for the pre-operative variables is given in Figure 4.7. This shows that all variables are moderately positively correlated and there is little evidence of the variables falling into tightly correlated groups as with the knees data. The only notable feature here is that the rightmost variable `walk.wit` (how far the patient can walk without pain) appears to be slightly less associated with the other variables. The same is true, thought to a lesser extent, with the variable `limping.p` with its visible pale bands across the plot. Examination of the post-operative plots show a very similar plot with all variables exhibiting moderate positive correlations, and the apparently weaker associations to `walk.wit` disappear.

4.3 Profile Plots

4.3.1 Introduction

One of the principal limitations of the t-test plots discussed in Section 4.1 is that the comparisons of variables over time can only be performed on pairs of time points. The visualisation of the continually changing mean values can only be attained through the comparison of pairs of sequential times. This restriction is somewhat limiting and there is a need for a mechanism to display the evolution of the means across all time points.

Various methods have been developed to show such information. Andrews *et al* [4] introduce the window and interval plots where the means of various variables are plotted directly accompanied by an appropriate error interval in order to determine significance of differences and perform multiple comparisons. They applied their results to illustrate the main effects of several ANOVA models.

Another near-identical approach can be found within *profile analysis* [57, 117, 114] which is a specialisation of multivariate analysis of variance (MANOVA) to a situation with multiple responses all on a similar scale. A further extension where there are several responses observed and at different times is sometimes called the *doubly-multivariate design.* If $Y_{t,i}$ is the ith observation made at time t, then we fit the univariate ANOVA model $Y_{t,i} = \mu + \tau_t + \varepsilon_{t,i}$ where μ is the grand mean,

and τ_t is the main effect at time t and errors, ε_{it} are assumed to follow the Normal distribution $\mathcal{N}(0, \sigma^2)$. The *profile* of Y is then simply the $\mu + \tau_t$ for all t for a particular variable, with a profile plot being the plot produced by directly graphing these quantities. This is, to all intents and purposes, simply a plot of the main effects obtained from an analysis of variance. Profile analysis is then based on the profile of the response variable(s) and all tests are performed in terms of the appearance of the profile rather than in the specifics of the results of the analysis of variance.

When investigating the effects of an additional categorical factor variable A, the ANOVA model would become $Y_{t,a,i} = \mu + \tau_t + \alpha_a + \beta_{t,a} + \varepsilon_{t,a,i}$ where α_a is the main effect for level a of A and β_a is the interaction term for time t and level a of A. To obtain the profile, we would then plot $\mu + \tau_t + \alpha_a + \beta_{t,a}$ for all times t and levels a of A. An example of two such profiles is given in Figure 4.8 and concerns the patient's walking ability for the two different diagnoses in the knees data set. It illustrates the changes in the mean of walking ability for the two groups over the four time points in the data and shows that the mean walking ability for the osteoarthritis group is better than the rheumatoid arthritis group at all time points. Whilst these plots are closely related to the ANOVA modelling of the data, the values plotted in Figure 4.8 are in fact simply the conditional means obtained directly from the data. Therefore these plots can be constructed directly without the need for the modelling itself and could thus be useful as exploratory graphics.

One shortcoming of profile analysis or the standard ANOVA modelling approach of these data is that they fail to allow for a dependency between the subsequent observations, which is inappropriate in the case of these repeated measures data. A solution to this problem is given in Crowder and Hand [24], which allows for unconstrained covariances between the response variables. However, both of these methods are for use in modelling the data and, for the purposes of finding a visual representation of the data, using such approaches in their entirety would introduce a complex data model at a premature stage. A further disadvantage of some of these methods is the requirements for complete cases in the data, in the sense that observations must be made at all time points and missing values must be imputed. This is not appropriate in the case of the orthopædic data under investigation where

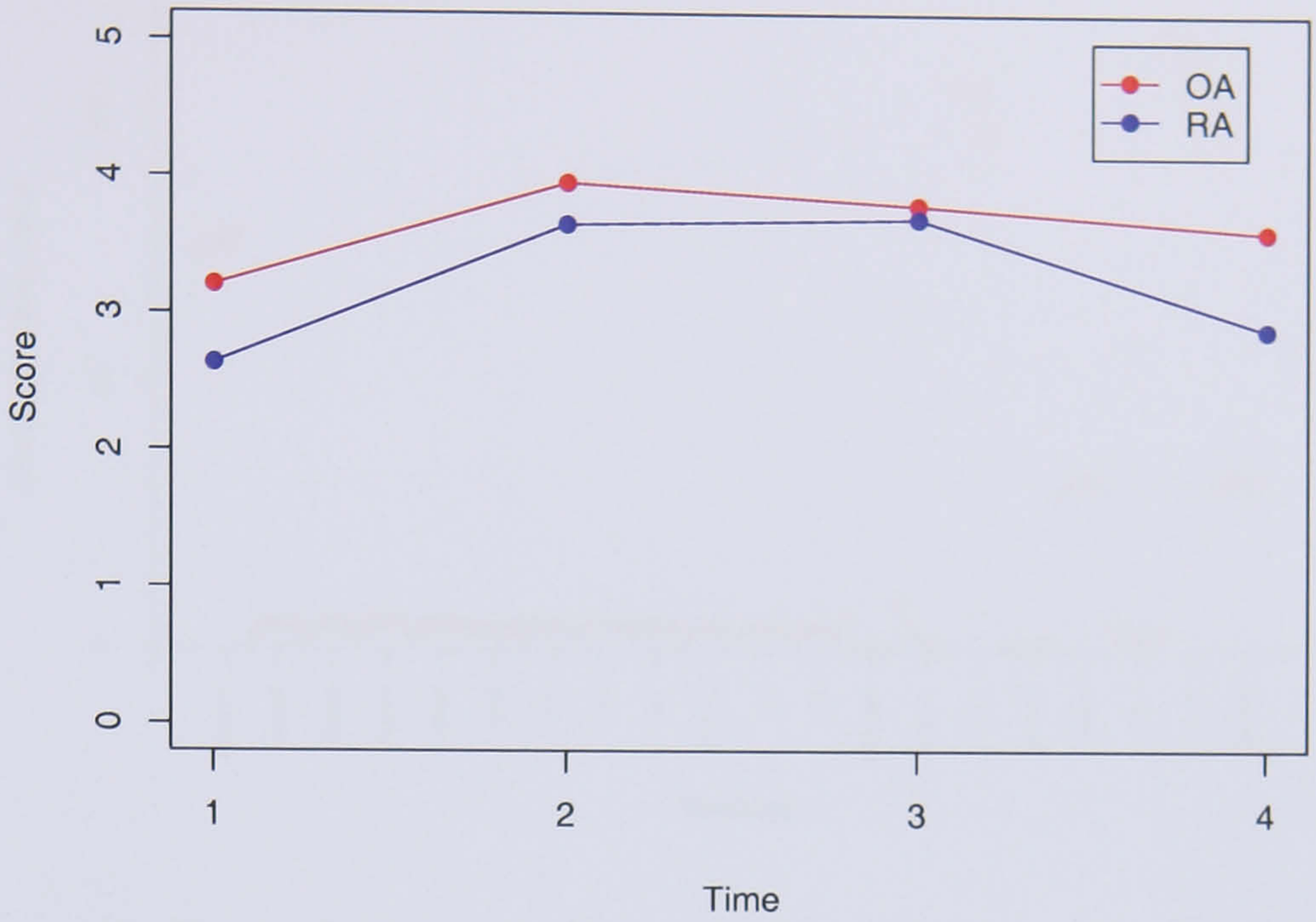


Figure 4.8: Profiles of the walking ability for the two diagnoses in the knees data.

the sample size substantially reduces over time. Therefore it would be sensible to focus only on the profile plots, which depict the changes in the conditional means over time for the different groups within the data. These mean values can simply be calculated from the raw data and do not require completeness of cases making them suitable for an exploratory graphic.

4.3.2 Standardised Profile Plots

The profile plot is a useful visualisation of the data and shows the evolution of the walking ability score over time. It also illustrates that there is only a small practical difference between the two groups based on diagnosis. However, this only focuses on a single variable and both the knees and hips data sets are highly multivariate which creates a problem. One approach would be to produce and examine separate plots for all key variables in the data set, however this would be impractical and would make it highly difficult to ascertain how pairs or groups of variables change over time.

An alternative method for tackling this dimensionality problem would be to display all the profiles simultaneously. However, as we can see from Figure 4.9,

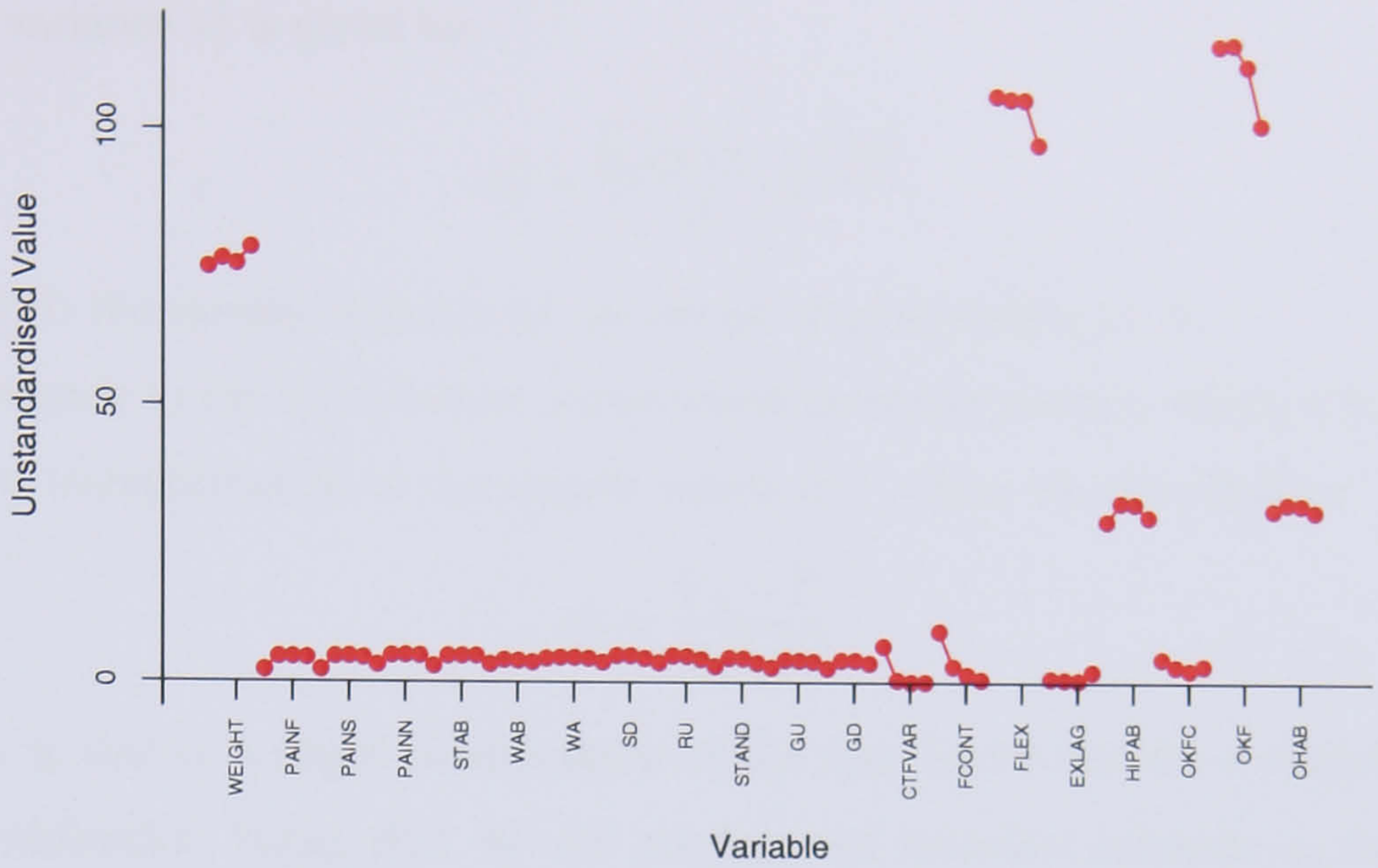


Figure 4.9: Unstandardised profiles of the key variables of the knees data.

when our response variables have substantially different scales the resulting plot is less than informative. Due to differences in both scales and locations of the variables, some trends that may be significant for variables with small means and variances are obscured by the effect of plotting them with variables with larger means and variances. For example in Figure 4.9, weight and the two flexion measures are obscuring any detail of the profiles of variables on smaller scales, such as the pain scores.

To compensate for this problem we can attempt to standardise the values so they are transformed onto a common scale. This eliminates the problems described above and will allow for an easy comparison between the different variables. A logical first step would be to use the standard transformation to t values performed in an independent sample t-test. So, let Y be a quantity of interest and let Y_t be random quantities denoting the value of that variable at different times $t = 1, \ldots, T$. Then the individual components of a profile are the values $\overline{Y}_t$, the sample means of a random sample sized n_t taken from Y_t. If we assume that $Y_t \sim \mathcal{N}(\mu_t, \sigma^2)$ then, trivially, $\overline{Y}_t \sim \mathcal{N}(\mu_t, \sigma^2/n_t)$. Since μ_t and σ^2 are unknown, we can estimate them using the $\overline{Y}_t$'s grand mean $\overline{Y} = \sum_t (n_t \overline{Y}_t)/N$ where $N = \sum_t n_t$, and the pooled

sample variance s_p^2 is given by:

$$s_p^2 = \frac{\sum_{t=1}^{T}(n_t - 1)s_t^2}{N - T}.$$

where s_t^2 is the sample variance of the sample corresponding to Y_t.

This leads to the first obvious transformation of the profiles, which is to use the standard transformation of the sample means to t values via the relation:

$$w_t = \frac{\overline{Y}_t - \overline{Y}}{s_p/\sqrt{n_t}}. \tag{4.1}$$

This is simply a slight modification of the standard t-test for a sample mean. The modification being that we are pooling our variance estimate s_p over all t samples rather than just 2. However this still corresponds to a significance test for a difference in means under the hypothesis $\mu_t = \mu$ and so could be interpreted as such. It should be noted that we have the case that the samples Y_{ti} are not at all independent and so there will be a resulting dependency among the t values which could have misleading results. Nonetheless, if we wish to produce a profile plot, then we have that the components of a single profile are the $\overline{Y}_t$ – the means of the corresponding variable at each time point. When we wish to consider multiple groups we can compute a separate profile for each group, where the individual group profiles are based on that group's conditional means.

The application of this method of standardisation to five of the knees variables is shown in Figure 4.10. We can see that the problems due to scale and location have now been eliminated and that we can now see how the standardised profiles of the variables change over time and with respect to one another. For example, it is clear that there is little notable change in the average weight, but pain frequency has a relatively sharp increase and then a subsequent decline over time.

There is a disadvantage with this scaling method and that is due to the fact that we do not divide each component of the profile by a constant. Since n_t is not constant for all t, this standardisation could result in unfortunate side effects such as the reversal of profile values. For example, if we have two groups with the following values: $\mu = 10$, $\overline{Y}_1 = 11$, $\overline{Y}_2 = 12$, $s_p = 4$, $n_1 = 9$ and $n_2 = 2$. We have the case that $\overline{Y}_1 < \overline{Y}_2$, however due to the differing sample sizes we get $t_1 = 0.75$ and

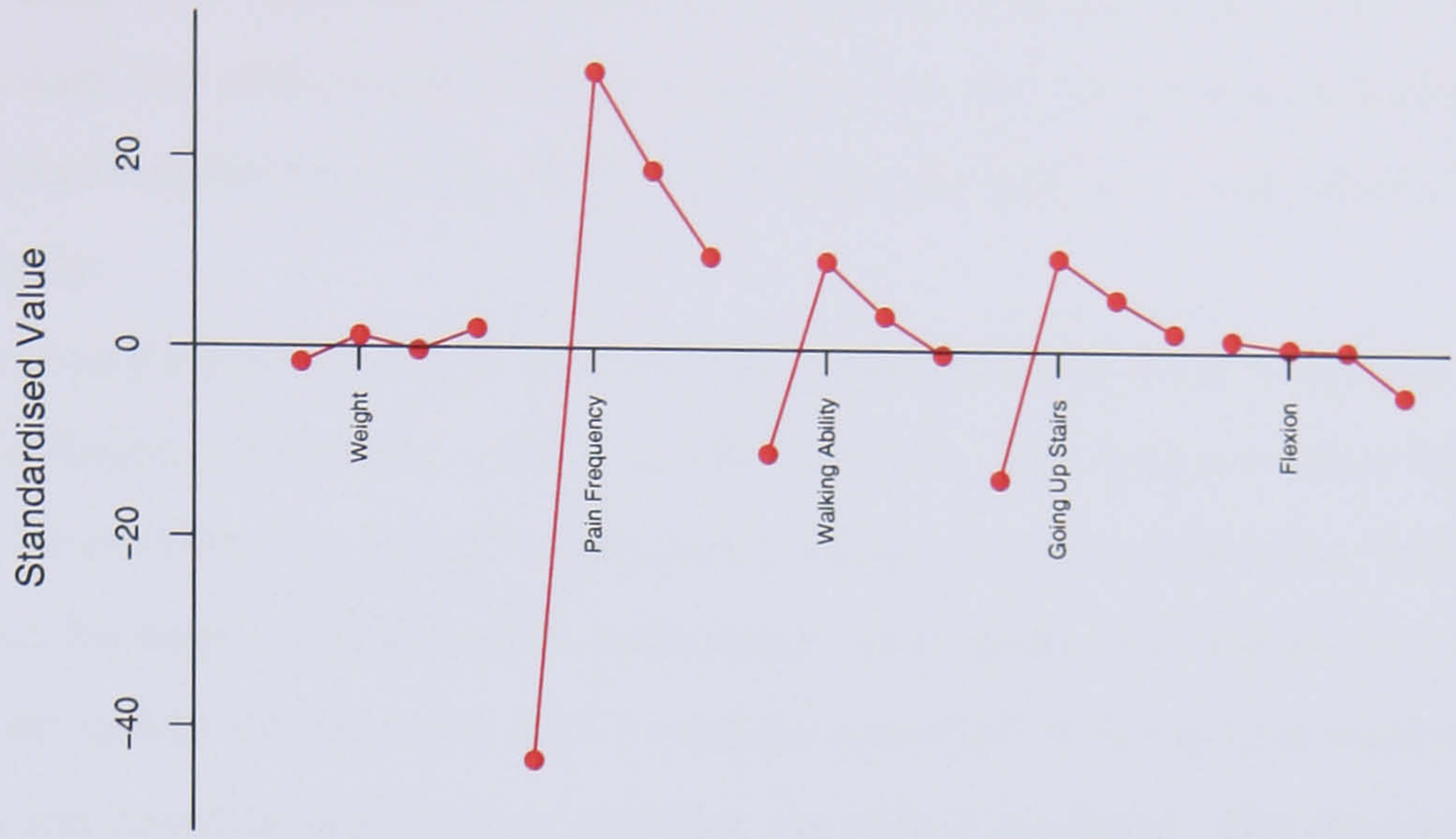


Figure 4.10: Standardised profiles for five variables of the knees data using method 4.1.

$t_2 = 0.71$ and so $t_1 > t_2$. Whilst a perfectly sensible result statistically, this artefact of the transformation does not preserve the original ordering of the profiles and so is a distinctly undesirable feature that would compromise the interpretability of the plot.

To address this problem, we can move from standardising to the usual t statistics as in (4.1), to consider *effect sizes* instead. Effect sizes are widely used in the social sciences [45] and are commonly to be found in meta-analysis studies. The notion of the effect size was introduced by Cohen [20]. Cohen describes the effect size attributed to a particular phenomenon, such as a treatment effect, as being "the *degree* to which the phenomenon is present in the population". Cohen constructs the effect size to be independent of sample size in order to prevent effects in larger populations with correspondingly lower errors from being inflated to reflect their greater statistical significance. Small or zero effect sizes still correspond to a failure to reject the null hypothesis of no effect, and large values correspond to the converse. The calculation of an effect size is thus simply a modification of the t-values in (4.1) with the omission of the term in n_t:

$$v_t = \frac{\overline{Y}_t - \overline{Y}}{s_p}. \tag{4.2}$$

In this case, this effect size measure is similar to Cohen's d [20] and Hedges's g [105] though the differences to these measures are the use of a pooled estimate of the standard deviation and the fact that we are pooling over more than 2 samples respectively.

This standardisation would preserve the ordering of the profiles but would eliminate the duality of the plot with a significance test. The interpretation would now have to be restricted to being standardised values and any association with a t-test could not be made. Additionally, information concerning the sample sizes at each time point would be lost since it is no longer included in the standardisation. This information could be depicted on the plot via colour intensity. The intensity of the points could correspond to their relative sample sizes using:

$$i_j = \frac{\sqrt{n_j - 1}}{\max_j(\sqrt{n_j - 1})} \tag{4.3}$$

where i_j is the colour intensity for the jth point in the plot and n_j is the corresponding sample size. By using $n_j - 1$ rather than n we ensure that $i \in [0, 1]$ with the boundary values occurring when n_j is 1 and $\max_j(\sqrt{n_j - 1})$ respectively. The lines joining two points on the plot can then be shaded using an intensity which equals the mean of the intensities of the two points being joined.

In addition to using colour intensity, one might choose to represent the sample size via the size of the plotting symbol. For example, using a circle or square with radius or side length equal to $\sqrt{n}$. Another possible alternative is to indicate sample size via the width of the line used to draw the line segments. This latter suggestion has a long history as documented in [119] where a 19th century graphic illustrates the terrible losses in Napoleon's army in Russia.

The results from applying this method of standardisation to the five knees variables are shown in Figure 4.11 for comparison with method 4.1. We can see that the profiles are suitably standardised but have a somewhat different shape than those in Figure 4.10 with the earlier profiles misrepresenting the relative sizes of the means for the five variables. The scale is also markedly less extreme due to the divisor no longer itself being divided by n_t. A further choice of value for a divisor would be to use the results of the profile analysis itself by using the estimate of the variance of

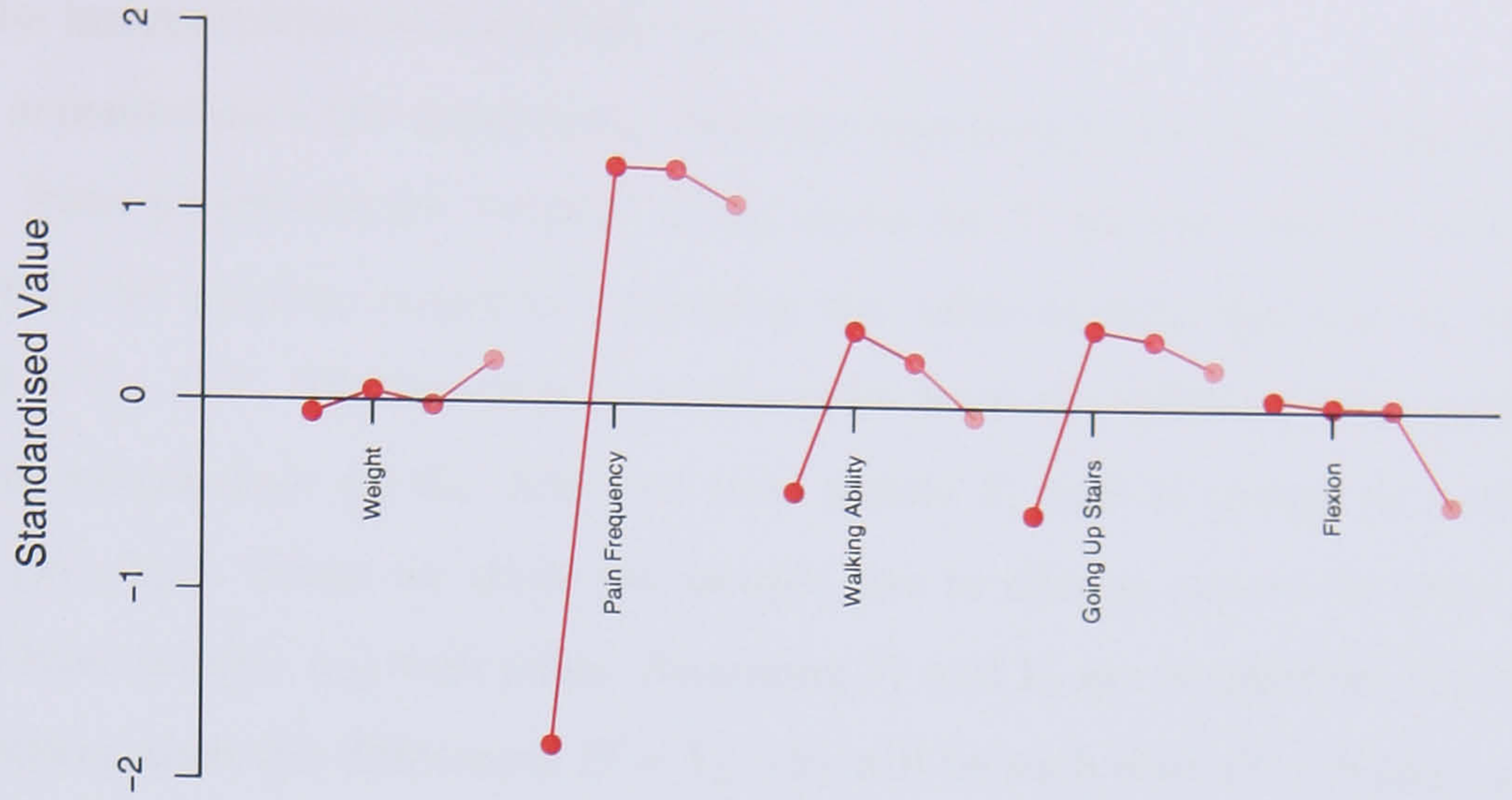


Figure 4.11: Standardised profiles for five variables of the knees data using method 4.2 showing relative sample size by colour intensity.

the random error in the ANOVA model, $\hat{\sigma}$. This would give a plot of profiles which preserves profile ordering but each profile would be scaled by a different constant to that in 4.2, giving profiles which are directly proportional to those in Figure 4.11.

4.3.3 Paired Profile Plots

One of the chief shortcomings of the standardised profile plots of the previous section is the assumption that the data from each time point are independent of one another. This assumption is violated as the condition of a patient at any time point is dependent on their previous state. Therefore, whilst the standard profile plots in the previous section give an appropriate display of the locations of the (conditional) means at the different time points, the information they convey relating to the change in patient condition does not take account of this dependency. For example in Figure 4.11, we can see that the weight profile is typically close to zero but appears to drop slightly at the third time point. This is due to the fact that the sample size decreases over time and some of the heavier patients had dropped out causing a corresponding drop in the average weight. If, instead, we were to compare the weights of the patients at the third time point with their previous weight values, we would observe that the patient's weights have, in fact, increased. This is at odds

with the interpretation of the profile plot.

To accommodate the dependency between time-points, we can use the standard theory from paired-sample t-tests. If we again let Y be the variable of interest and let Y_t be random quantities denoting the value of that variable at different times $t = 1, \ldots, T$. We can then form the pairs from the different time points. for example we can pair up the first two time points Y_1 and Y_2 giving us a series of values $(Y_{1,i}, Y_{2,i})$. Since we allow the sample size to change across the time points we will have $\min(n_1, n_2)$ such pairs. Assuming Y_1 and Y_2 are samples from a Normal distribution, then the differences $D = Y_2 - Y_1$ will be such that $D \sim \mathcal{N}(\mu_2 - \mu_1, \sigma_D^2)$ where σ_D is unknown and is estimated by s_D, the sample standard deviation of the differences. Under the null hypothesis that the mean difference is zero, the standard methodology is to base inferences on

$$t = \frac{\overline{D}}{s_D/\sqrt{n}} \tag{4.4}$$

which follows a t distribution with $n - 1$ degrees of freedom and where $\overline{D}$ is the mean of the D values.

It would be meaningless to plot these t values directly as the divisor of the fraction will differ from point-to-point leading us to encounter problems with reversals of profiles such as those encountered with using method (4.1) to standardise the profile plots in Section 4.3. This is due to both s_D and n varying from point to point. To counter these problems, we can use similar techniques as those employed with the standardised profiles. If we firstly assume that for a single variable, all the pairwise differences have the same variance we can then estimate this by the pooled sample variance:

$$s_p^2 = \frac{\sum_{j=1}^{T-1}(n_j - 1)s_{D_j}^2}{N - (T - 1)} \tag{4.5}$$

where we have sets of pairs D_j where $j = 1, \ldots, T - 1$ with associated sample sizes n_j and standard deviations s_{D_j}.

In order to prevent the reversals of the profiles, we must adjust the standardisation in a manner similar to that of (4.2). However as before, this will also remove the duality with the statistical t-test, though it is a necessary mechanism to display the data in a manner which enables the comparison across variables, time points

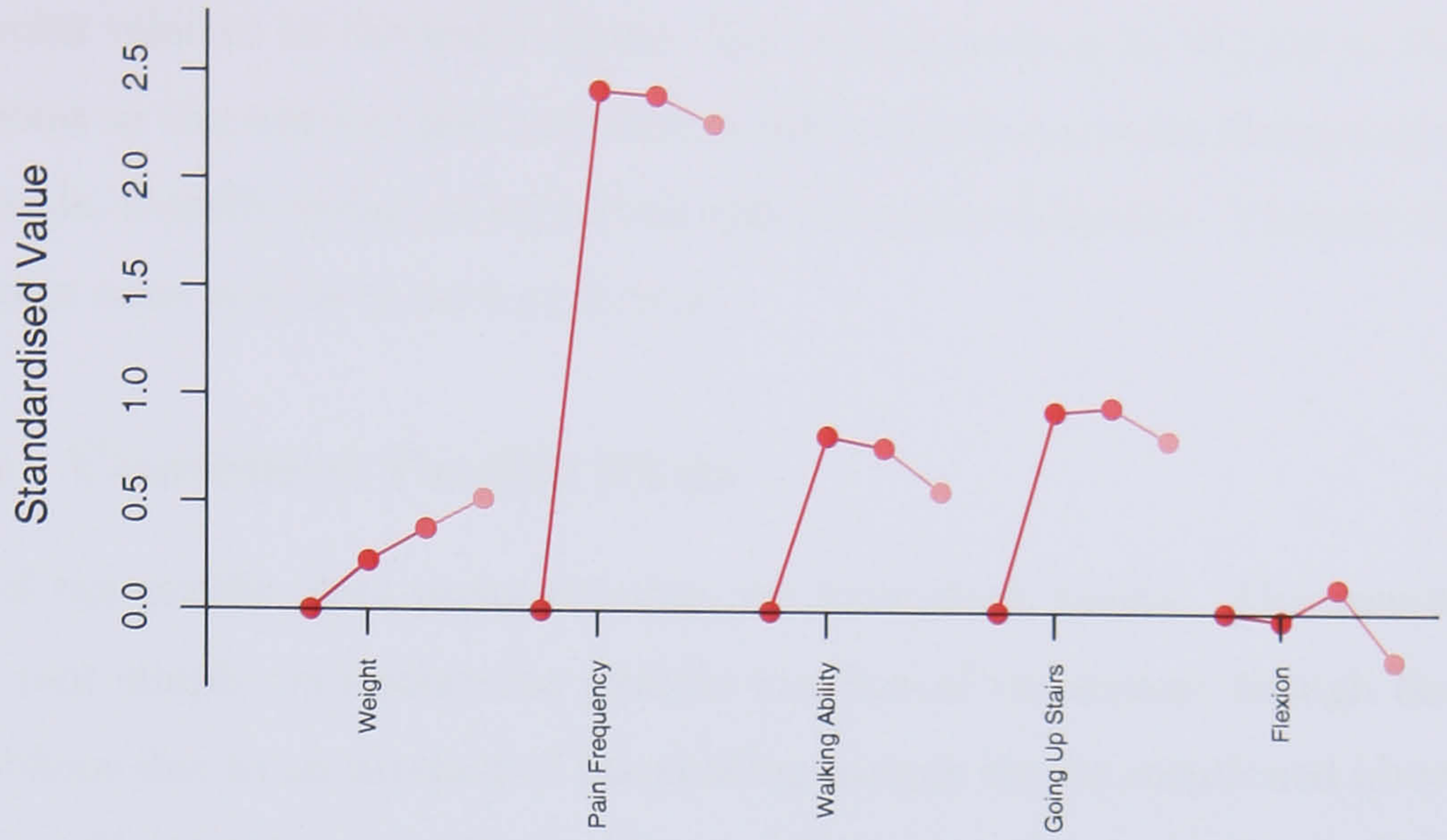


Figure 4.12: Paired profiles for five variables of the knees data using (4.6) and showing relative sample size by colour intensity.

and other groups. Therefore for each set of pairs we calculate:

$$w_j = \frac{\overline{D_j}}{s_p}. \tag{4.6}$$

For visualisation purposes on the paired profile plot, the initial value for each profile was set to be zero. The subsequent values were then the standardised mean difference values corresponding to the pairs (Y_1, Y_2), (Y_1, Y_3), etc. Thus a zero value for w_j would correspond to a mean value equal to $\overline{Y_1}$. This uses the initial time point as a baseline against which the subsequent data are compared and enables a sensible depiction of the changes over time. To display information on the sample sizes at each point, the colour intensity method used in Section 4.3.2 can be employed.

The results from applying this pairwise visualisation method to the same five knees variables is displayed in Figure 4.12 for comparison with Figure 4.11. We can see that the profiles are now all arranged to start at zero, corresponding to the baseline value against which all the paired differences for the later time points are compared. We can also see that by looking at the paired differences between the different time points some of the profiles now have a different shape. For example, the standardised profile for *Weight* in Figure 4.11 stays close to the mean value and then deviates slightly upwards at the final time point. Considering the paired version of this profile in Figure 4.12 we see that *Weight* actually increases at each

time point relative to the initial value. The apparent drop in *Weight* at the third time point in the original plot was due to some heavier patients disappearing from the sample, thereby giving an erroneous indication of a reduction. The paired profile method is more robust to such problems.

4.3.4 Combined Profile Plots

Both of the profile plots presented thus far have their merits. The standardised profile plot simply represents the relative location of the means, though there can be problems due to an artefact of the shifting sample size as mentioned above. The paired profile plot corrects this problem by displaying the mean change relative to a baseline. However, if we have two or more subgroups in the profile plot then the initial means and hence the baselines will be different for the different groups. However, the information about the relative locations of these starting points is lost as the paired profile plots will all start at zero. Thus we gain a more accurate insight into the changes of the mean over time with the paired profile, but we lose the information about relative positions which we easily represented in the standardised profile.

To address both problems simultaneously, we can use the paired profile plot and offset the baselines from the origin by an appropriate amount to re-introduce the notion of location that we had with the standardised plots. To this end, we can construct the plot using the following form:

$$x_j = \frac{\overline{Y}_1 - \overline{Y}}{s_p} + \frac{\overline{D}_j}{s_p}. \tag{4.7}$$

where $\overline{Y}_1$ is the sample mean at the first time-point, and $\overline{Y}$ is the grand sample mean taken over all time points (s_p and $\overline{D}_j$ are as previously defined). Thus we combine the forms of (4.2) and (4.6) to marry together the accurate depiction of the relative locations and evolutions over time. A summary algorithm for constructing this plot is given in Figure 4.13.

The results from applying this combined profile method to the five knees variables that have been graphed previously is presented in Figure 4.14. We can see from this plot that the paired profiles from Figure 4.12 have been translated to the

1. For each variable v_i where $i = 1, \ldots, p$:

   (a) For each time point t_j where $j = 2, \ldots, T$:

      i. Let $\boldsymbol{Y}_j^{(i)}$ be the vector of length n_j containing the observations on variable v_i at time point t_j.

      ii. Calculate the mean value of v_i at time t_j, $\overline{Y}_j^{(i)}$.

      iii. Calculate the differences $\boldsymbol{D}_j^{(i)} = \boldsymbol{Y}_j^{(i)} - \boldsymbol{Y}_1^{(i)}$. Calculate the mean $\overline{D}_j^{(i)}$ and standard deviation $s_{D_j^{(i)}}$ of the $\boldsymbol{D_j}^{(i)}$.

   (b) Find the pooled sample standard deviation for v_i, $s_p^{(i)}$ using (4.5) and calculate the mean value of v_i over all time points, $\overline{Y}_.^{(i)}$.

   (c) Calculate the paired difference values: $p_j^{(i)} = \overline{D}_j^{(i)} / s_p^{(i)}$, where $\overline{D}_1^{(i)} = 0$.

   (d) Calculate the offsets: $\omega_j^{(i)} = (\overline{Y}_j^{(i)} - \overline{Y}_.^{(i)}) / s_p^{(i)}$.

   (e) Plot the combined profile values, which are given by: $x_j^{(i)} = \omega_j^{(i)} + p_j^{(i)}$.

Figure 4.13: Algorithm for the construction of a combined profile plot.

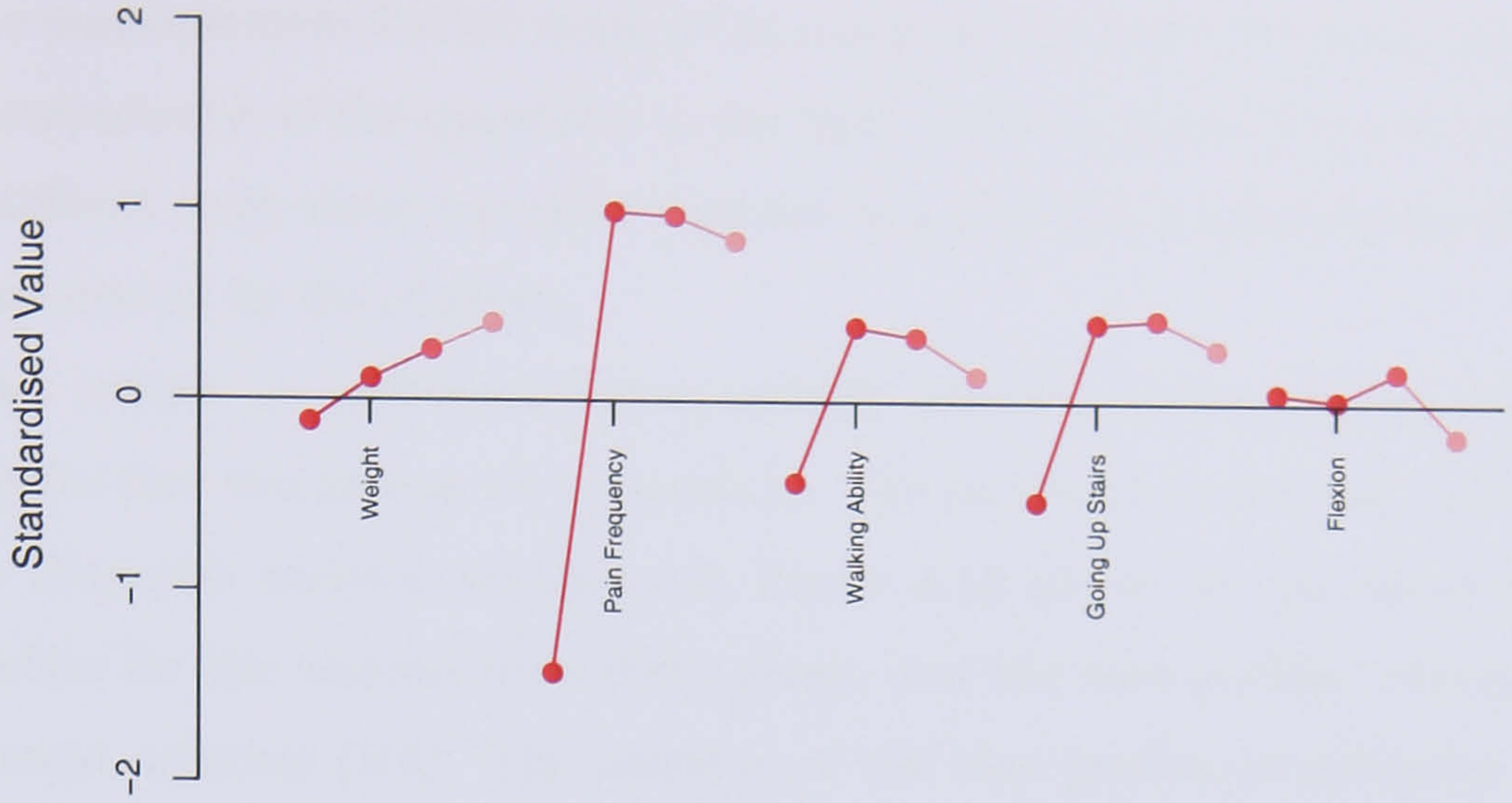


Figure 4.14: Combined paired profiles for five variables of the knees data using (4.7) and showing relative sample size by colour intensity.

locations in Figure 4.11. The true merit of offsetting the locations of the profiles is not immediately obvious when graphing only one group - the true benefits will be apparent when considering several subgroups of the data.

4.3.5 Results

Applying the combined profile plot methodology to the knees data yields the plot given in Figure 4.15. This plot displays the changes in the variables over time and all variables are now easily comparable. At first glance, one can observe that there is a common pattern in many of the variables - that of a steep increase from pre- to post-operation and then a slow decline over the subsequent time points. Those variables for which this is the case are the ordinal variables which correspond to the pain and mobility variables. The interpretation of this pattern is that there is a pronounced improvement in the average patient's condition as a result of the operation they received in the intervening time. After the second time point at one-year after the operation there is no longer a continued improvement and the mean levels of pain and mobility slowly deteriorate over time. However, they do not return to the level they were at prior to the operation. The improvement in the pain scores seem to be relatively greater than those of the other mobility scores

and the post-operative decline seems to be relatively less severe for these quantities. The interpretation of the quantities to the right of *Going Down Stairs* is somewhat more difficult since these represent anatomical angles and measurements that are best understood by the clinician.

If we include an additional factor variable into the profile plot, we can split the profiles into two groups for comparison. The results of performing such a split on the *Diagnosis* variable are shown in Figure 4.16 where the red lines represent the profiles for the osteoarthritis (OA) group, and the blue profiles correspond to rheumatoid arthritis (RA). The paleness of the blue profiles is indicative of the relatively small size of the rheumatoid group when compared with the osteoarthritis group. The plot shows that the two groups typically follow the same pattern of sharp improvement and slow decline. However, it appears that rheumatoid group display poorer average walking ability and their mean condition at 10-years is, for some measurements, actually poorer than the pre-operative state. However, since the sample size for this final sub-group is very small ($n = 9$) there will be a large variance attached to this mean value reflecting a high degree of unreliability in the position of that point. Nonetheless, we can see that the profiles are typically quite close together suggesting a similarity of patient's conditions and their evolution between the two groups. The only variable for which there is a possible separation of the groups is *Weight*, as previously discovered in Section 3.2, where patients with rheumatoid arthritis are typically lighter. This is a characteristic that is widely recognised within the orthopædic community. The only other distinctions between the groups are smaller in scale and occur for a small group of variables including *Walking Ability* and *Walking Aids*.

Repeating the process for the two operations yields the plot displayed in Figure 4.17. From this plot it is quite easy to see that there is very little difference between the two operations with the two sets of profiles all lying very close to one another. This corroborates previously made statements on the nature of the *Operation* variable in Chapter 3. The only noticeable differences are in the values of variables such as *Weight* at the final time point, it being significantly greater for the uncemented operation (blue profile) than for the cemented. However, this is another case

Figure 4.15: Profile plot for the key variables of the knees data.

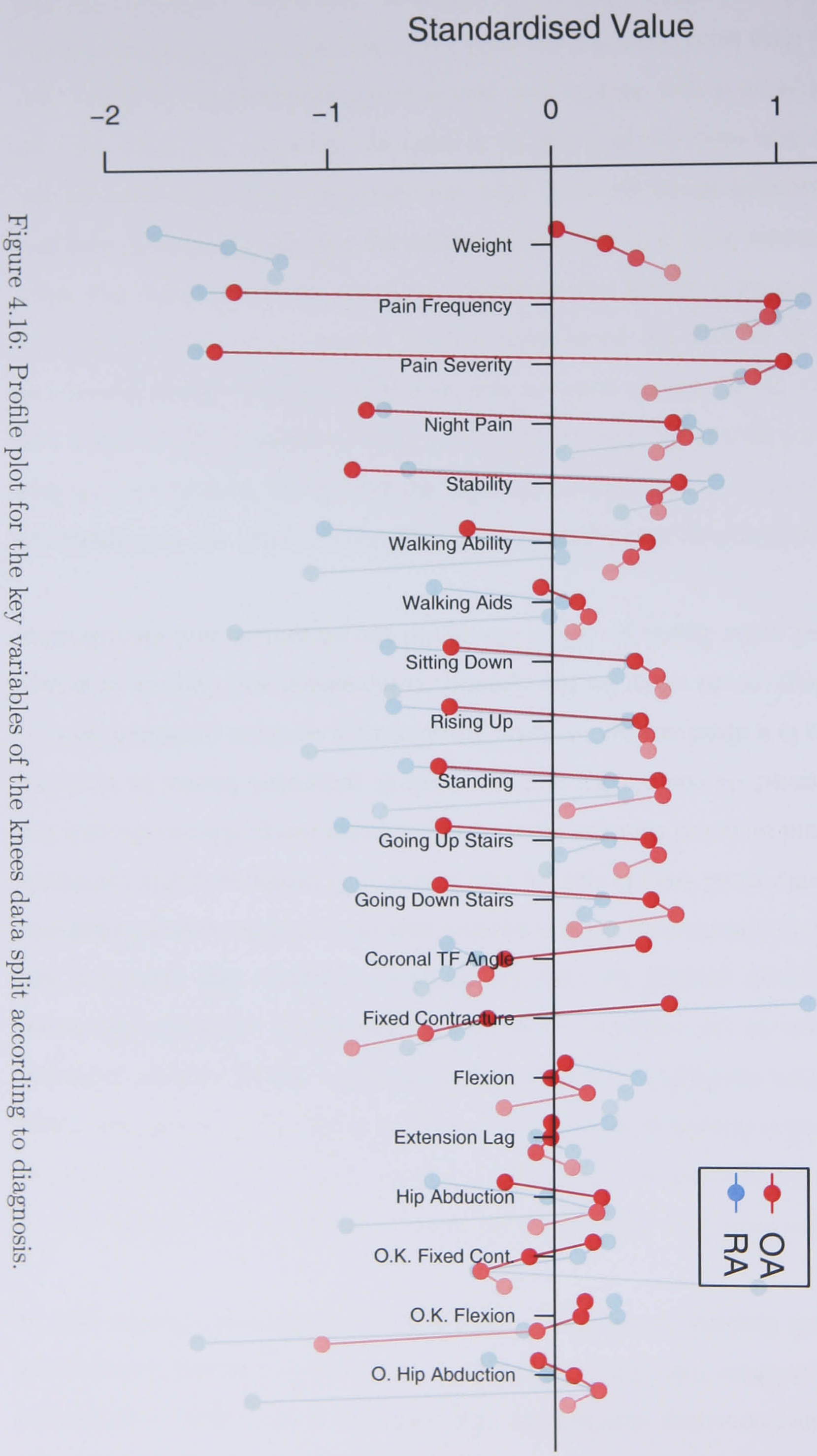


Figure 4.16: Profile plot for the key variables of the knees data split according to diagnosis.

of a relatively small sample size ($n = 33$) giving a value with comparatively high variance and corresponding unreliability. The addition of error bars or confidence bounds would indicate this unreliability however the graph is already complicated and adding such features would likely only confuse inexperienced viewers such as clinicians.

One can also introduce both diagnosis and operation into the profile plot to allow for further comparisons. The results are shown in Figure 4.18. The recurring steep rise and slow descent pattern is still present in many variables and the strong difference between the diagnoses according to the patient's weight remains. However, the eye is drawn to the extreme values for the rheumatoid arthritis/uncemented combination (purple profile) for variables *Fixed Contracture*, *Hip Abduction* and *Other Knee Fixed Contracture*. This is another case of the problem described above, where a relatively small sample size (in this case $n = 7$) produces highly unreliable results. It should be noted that the addition of the extra two profiles to the plot has markedly increased its complexity and has made it somewhat harder to read and interpret. This is a disadvantage of static graphics such as those presented here. A computer package could allow for the dynamic combination and division of groups of variables and other functions which would aid the separation and identification of subgroups.

These techniques were also applied to the hips data, though the full results are omitted for brevity. A standard profile plot for five of the variables in the hips data is given in Figure 4.19. The variables have been re-scaled so larger values now correspond to a better patient state, in line with the knees data. From this we can see that there is, as with the knees data, a sharp increase from before to after the operation suggesting a strong improvement. However we can also see that there is a second increase from 3 to 12 months suggesting the average patient condition continues to improve up to one year after the operation. Since the times for the post-operative hips data are at 3 and 12 months and those for the knees data are at 1, 5 and 10 years it may be the case, and it is not unreasonable to expect that had additional data been observed we could have observed a slow decline in patient condition after 1 year. When comparing different pathologies, there was minimal

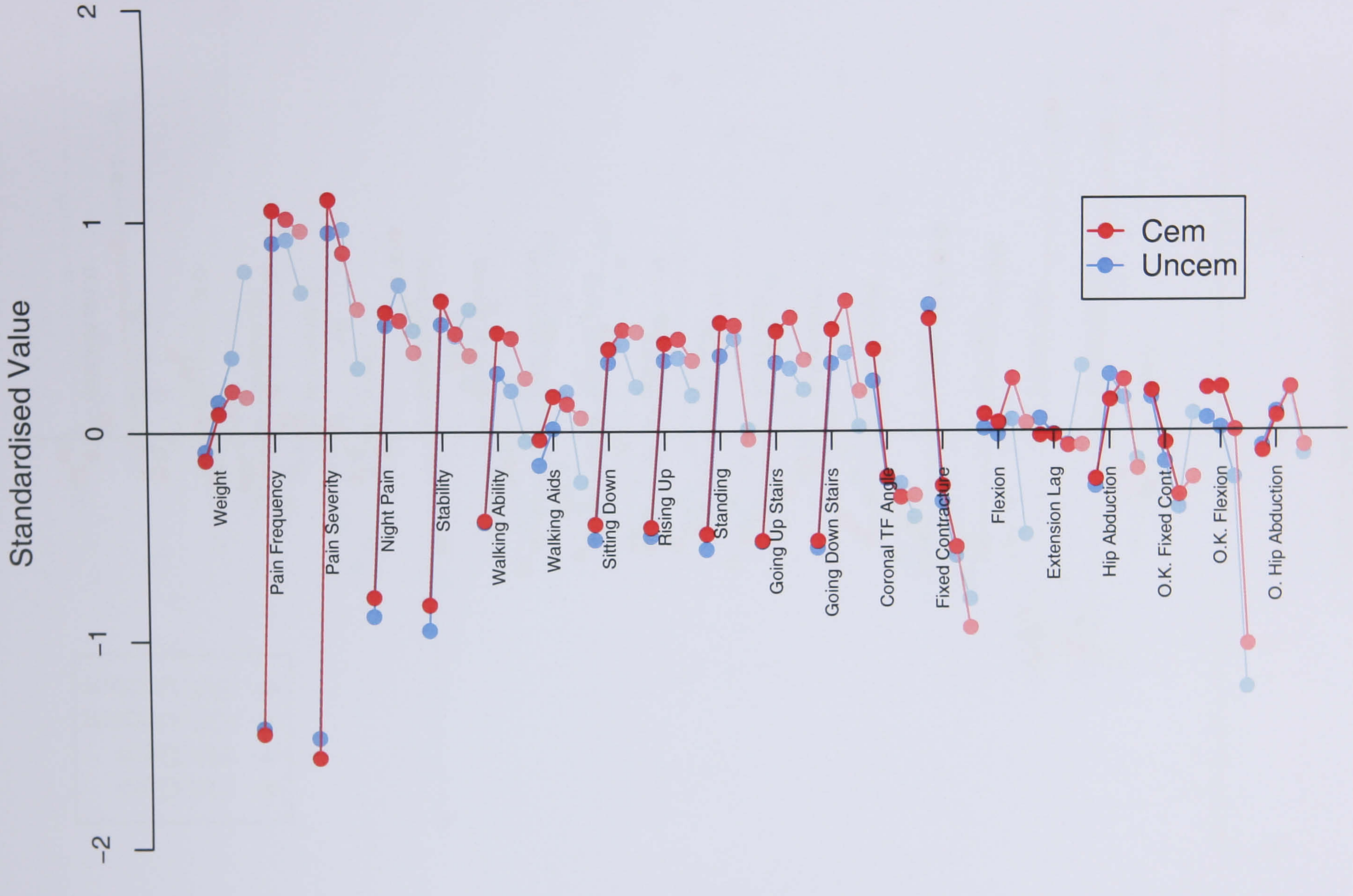


Figure 4.17: Profile plot for the key variables of the knees data split according to operation type.

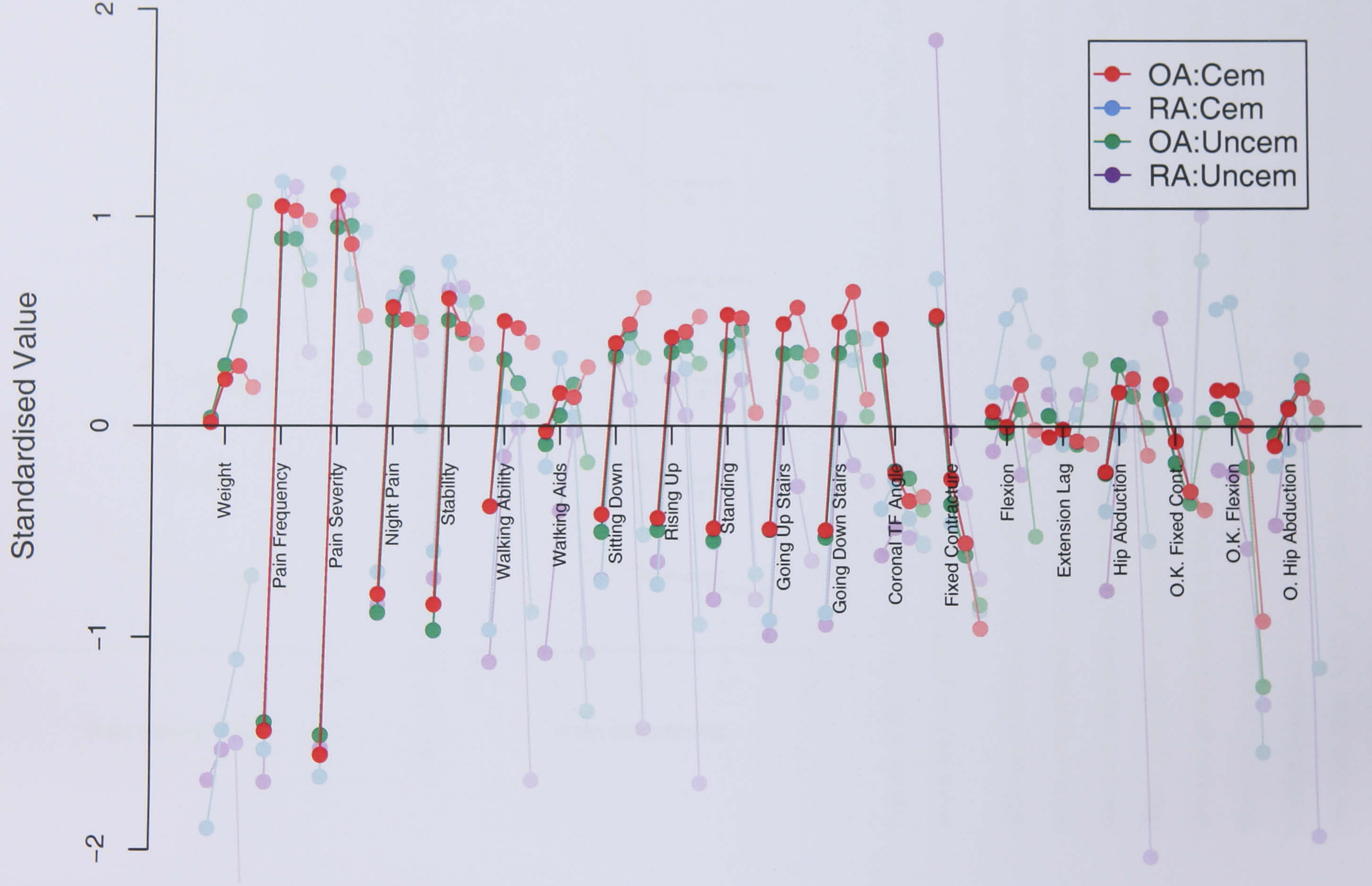


Figure 4.18: Profile plot for the key variables of the knees data split according to both diagnosis and operation type.

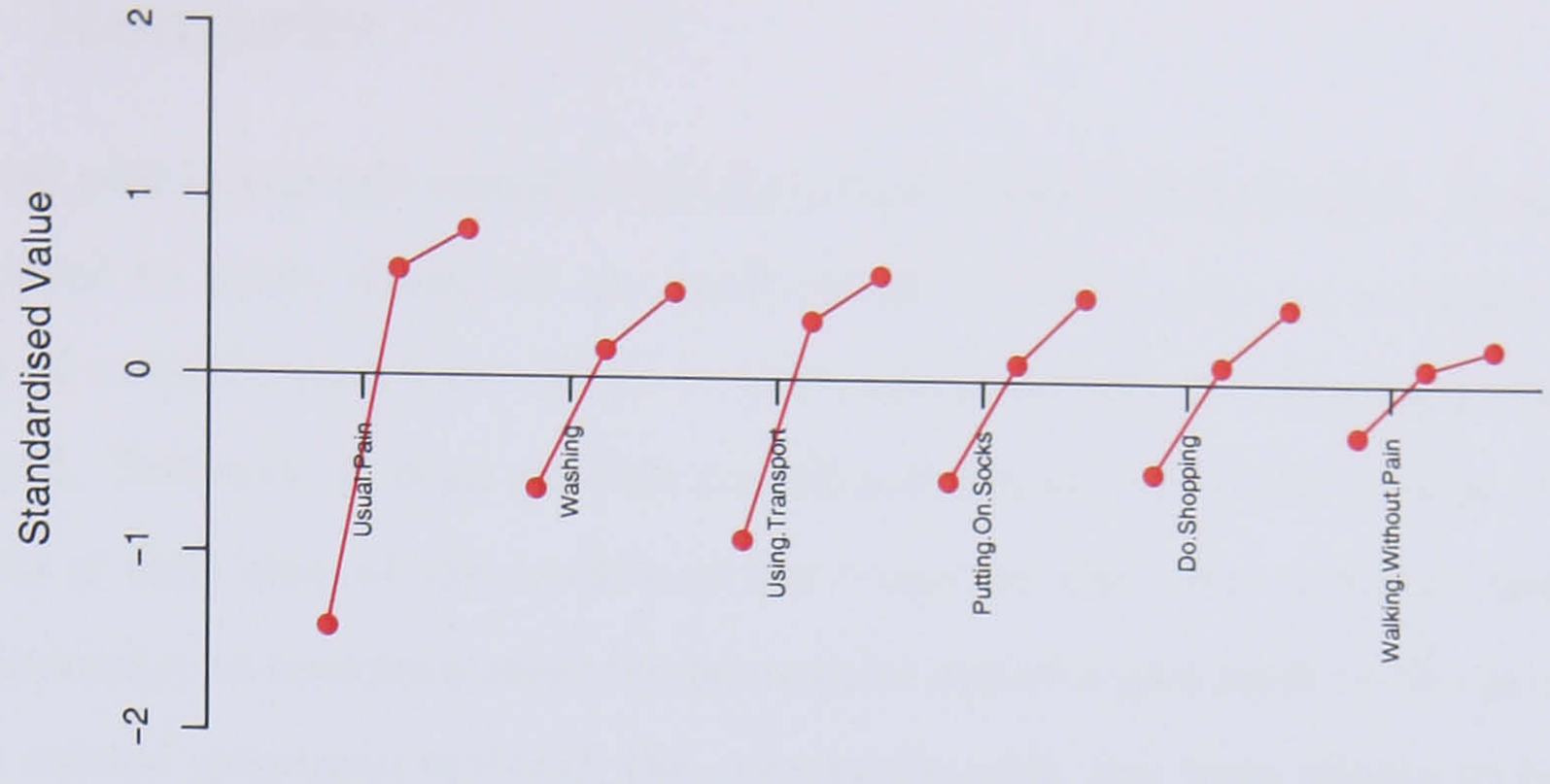


Figure 4.19: Profile plot for the five of the key variables of the hips data.

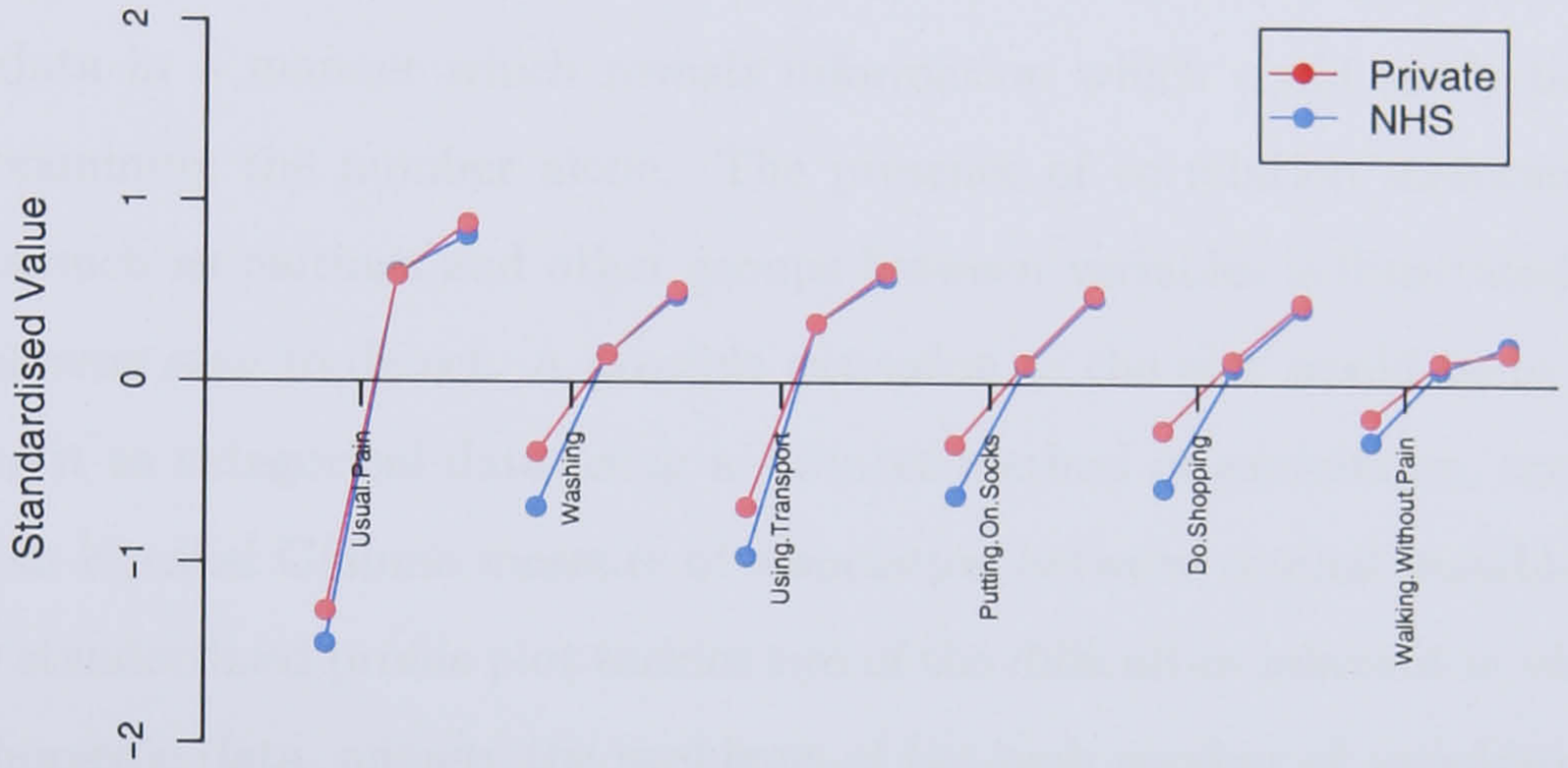


Figure 4.20: Profile plot for the five of the key variables of the hips data split according to private status.

difference between the profiles with the exception of the patient's age where people with osteoarthritis were, on average, younger than the other patients. The hips data also had information on the NHS/private status of the patient and a profile plot with these groupings is displayed in Figure 4.20. The plot shows that post-operatively the two groups fare in a very similar manner with little distinction between the two. However, pre-operatively it would appear that NHS patients typically have a slightly poorer condition compared to those patients who went private. This may be due to the fact that NHS patients would likely have to spend some time on a waiting list before their initial consultation resulting in their condition being more developed.

4.4 Remarks

The t-test plot is a simple visualisation for a basic statistical technique. Its usefulness is restricted to cases where we are performing several t-tests on a relatively large number of variables and it is subject to the usual assumptions associated with t-tests in general. However, it does provide an efficient summary of the results. A second weakness is that due to the nature of the t-test we can only compare two groups simultaneously; to compare more would require use of a plot such as the profile plot.

The second graphical method, the correlation plot, has been shown to be an efficient and compact representation of a standard correlation matrix. The correlation plot straightforwardly displays the information contained in the correlation matrix of the data in a manner which reveals information which could easily be missed when examining the number alone. The presence of correlation structure within the data such as pairings and other groups between variables is illustrated making such patterns easy to detect. A possible extension to the plot would be to consider applying it to categorical data using a suitable method of association, such as the Goodman-Kruskal Gamma measure of association between ordinal variables [55].

The standardised profile plot tackles two of the difficulties inherent in visualising the orthopædic data, namely the problems of the high number of variables and the time series aspect to the data. The profile plot presents data for all variables at all time points on the same plot, eliminating the need to produce reams of graphs and the difficult task of then performing several comparisons between these to interpret the data. Possibly the most useful aspect of the plot is that it gives insight into the evolution of the patient's condition over time and enables use to observe the recurring pattern of a sharp increase followed by a slow decline. One limitation of the procedure is that when comparing multiple groups on the same plot, such as in Figure 4.18, the plot becomes swamped with information making it difficult to interpret. Another problem occurs when the sample sizes are small, which results in unreliable values for the conditional means producing extreme values when plotted, such as at later time points in Figure 4.18. Using the colour intensity method to illustrate the relative sample sizes of each point is a possible method to combat this.

The standardised profile plots assume that the individual time points are inde-

pendent of one another. This is not the case, and so the paired profile plot attempts to remedy this. Being similar in structure and appearance to the standardised profile plot, the paired methodology is more suitable for illustrating the change and evolution of the mean patient condition over time. However, information on the differences in location that is shown in the standardised profile plots is lost in the paired framework. To address this, examining both plots should allow us to gain a sound impression of the nature and changes in patient condition over time and across groups. The consequent development into the combined profile plot successfully combines the best properties of both plots and presents a single, informative plot to display the evolution of the patient status data over time.

Chapter 5

Graphical Modelling

A graphical model is a statistical model which represents dependencies among random variables by a graph in which each variable is a node and each dependency is an edge. Graphical modelling as a statistical methodology can be traced back to the early 20th century with foundations in statistical physics [128] and path analysis [54]. However, the major developments in the field are of a more recent origin with notable contributions by Lauritzen, Wermuth and Cox [25, 124, 22, 83].

The goal of this chapter is twofold: first to provide a brief overview of the fundamentals of the graphical modelling methodology; and secondly to apply these methods in an exploratory fashion to our orthopædic data. The chapter is therefore divided into two principal sections covering these two topics respectively. The first section provides a review of the notions of conditional independence, the independence graph, the graphical models themselves and model selection methods. This section is intended only as an overview - a more comprehensive treatise on the subject can be found in the books by Edwards [38], Whittaker [125], or finally Lauritzen [81] who gives a thorough exposition of the theoretical foundations of the methodology. Some applications of these methods available in the literature are also presented.

The second section deals with the application of the graphical modelling methods to the two orthopædic data sets studied previously. The section reviews the assumptions needed to use the graphical models and the analysis of the pre-operative and post-operative time points for both data sets. The chapter concludes with a review

of the limitations of the methodology encountered in the course of the analysis in this chapter with a view to detailing the problems to be addressed in subsequent chapters.

5.1 Graphical Models

5.1.1 Data and Independence

Suppose we have a set of variables, $V, W, \ldots, Z$ say, for which we have a set of n observations $(v_i, w_i, \ldots, z_i)$. To model these data we would assume that $V, W, \ldots, Z$ are random variables with a joint probability density function:

$$f_\theta(v, w, \ldots, z),$$

where θ is some unknown parameter or vector of parameters. Thus we can imagine that our data set is simply a random sample of size n from f_θ. We now base our inferences about the population on the values of this unknown parameter θ.

It is important at this stage to discriminate between two types of variables: *continuous* variables which can take any real value in $\mathbb{R}$, and *discrete* or *factor* variables which can take values only from a finite set. If a variable X is continuous then its density function is written as $f_X(x)$, whereas if X is discrete, this density may be written as $\mathrm{P}[X = j]$ where j is one of the possible *levels* of X with $j \in \{1, 2, \ldots, \#X\}$.

For graphical models, two key notions are those of *marginal* and *conditional independence*. For a thorough coverage of these topics the reader is referred to Chapter 2 of Whittaker [125]. Two variables X and Y are *marginally independent*, written $X \perp\!\!\!\perp Y$, if their joint density can be written as the product of their marginal densities:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y). \tag{5.1}$$

Equivalently, two variables are marginally independent if the conditional density of, say Y given $X = x$ is not a function of x:

$$f_{Y|X}(y, x) = f_Y(y). \tag{5.2}$$

The advantage of this second expression is that it illustrates that the conditional density of Y given X is unaffected by the value of X.

Conditional independence is an extension of this notion of marginal independence. Suppose we now have three variables X, Y, and Z. If, for every value $Z = z$, we have that X and Y are independent in the corresponding conditional distribution $f_{XY|Z}$ then we say that X and Y are conditionally independent given Z. In terms of the density functions, this statement equates to:

$$f_{XY|Z}(x, y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z). \tag{5.3}$$

This is now written as $X \perp\!\!\!\perp Y|Z$, where this notion is due to Dawid [26]. We can also see that marginal independence is a special case of conditional independence when Z is trivially empty. It is the interpretation of such conditional independence relationships that is one of the most appealing feature of graphical models. Conditional independence relationships can be seen as statements of *irrelevance*, for example $X \perp\!\!\!\perp Y|Z$ can be interpreted as:

If we know Z, then information about Y is irrelevant or uninformative for knowledge of X.

Or, to put it another way if we have observed a value of Z then observing Y does not provide us with any further information about the possible value of X.

5.1.2 Independence Graphs

The fundamental component of a graphical model is the independence graph which is used to depict the conditional independence relationships between pairs of variables. A *graph*, $\mathcal{G}$, is a structure consisting of a finite set, V, of *nodes* (or vertices) and a finite set, E, of *edges* (or arcs) between pairs of nodes. A graph can be represented as a diagram, such as in Figure 5.1 where nodes are drawn as circles and edges as lines joining those circles. These graphs are called *undirected* since the edges are directionless and are drawn with lines. An undirected edge joins two nodes, X and Y say, in such a way that X is connected to Y and vice versa. A *directed* edge would represent a one-way connection. Familiarity with basic graph concepts such

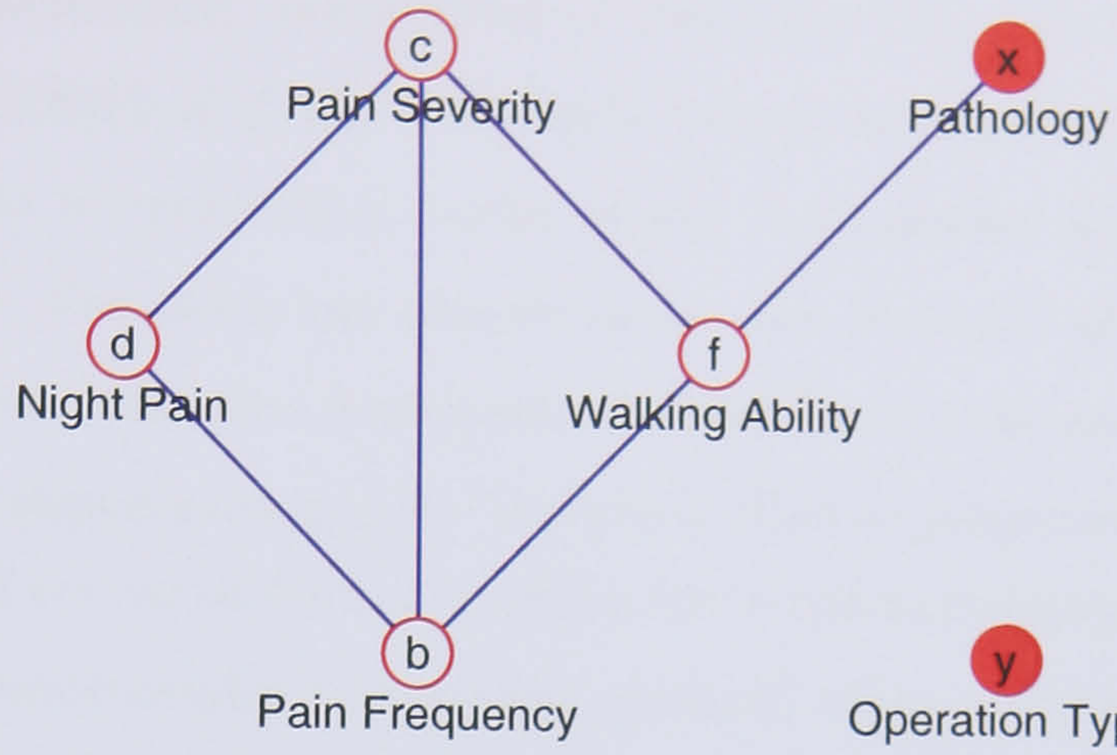


Figure 5.1: A simple graphical model for six variables in the knees data.

as adjacency, subgraphs, cycles, cliques, separation and triangulation is assumed. For more details see section 1.2 of Edwards [38], section 2.1 of Lauritzen [81] or chapter 3 of Whittaker [125].

The relationship between the independence graph and the graphical model is a straightforward one. Variables in the graphical model are represented as nodes in the independence graph. Since we have two types of variables in our data and models we represent this distinction on our model graph, which is now known as a *marked graph* since we have groups of different types of nodes. Continuous variables are drawn as hollow circles, and discrete variables as solid circles – as illustrated in Figure 5.1 where nodes $\boldsymbol{x}$ and $\boldsymbol{y}$ are discrete. Since the focus of graphical modelling is the conditional independence relationships between pairs of variables, these are represented directly on the graph. The edges drawn on the graph join those variables which are *not* conditionally independent given all other variables in the model, i.e. an edge represents a dependency between a pair of variables. Thus for all variable pairs (X, Y) such that $X \perp\!\!\!\perp Y|$ (all other variables), then the edge between X and Y is *omitted* from the graph; all other pairs have edges joining them.

Formally put, the conditional independence graph of the variables $X_1, \ldots, X_m$ is the undirected graph $\mathcal{G} = (V, E)$ where $V = \{X_1, X_2, \ldots, X_m\}$ and (X_i, X_j) is *not* in the set of edges E if and only if $X_i \perp\!\!\!\perp X_j | X_{V \setminus \{X_i, X_j\}}$.

Thus, for example, we can see from Figure 5.1 that if this model holds then *Pain Severity* ($\boldsymbol{c}$) is conditionally independent of *Pathology* ($\boldsymbol{x}$) given the other variables in this model – $\boldsymbol{c} \perp\!\!\!\perp \boldsymbol{x} | \{\boldsymbol{b}, \boldsymbol{d}, \boldsymbol{f}, \boldsymbol{y}\}$. This leads directly to one of the three properties that are crucial to interpreting a model graph, the *pairwise Markov property for undirected graphs.* The other two properties are the *local* and *global Markov properties* which further refine the statements of conditional independence that can be drawn from the independence graph. The three Markov properties are:

Pairwise If two variables are not adjacent in the model graph then they are conditionally independent given all other variables in the model.

Local Each variable is conditionally independent of its non-neighbours given its neighbours.

Global If two sets of variables u and v are separated by a third set of variables w, then $u \perp\!\!\!\perp v | w$.

It has been shown by Pearl and Paz [96] that all three of these Markov properties are, in fact, equivalent.

Here, we can then use the local Markov property to refine our previous statement $\boldsymbol{c} \perp\!\!\!\perp \boldsymbol{x} | \{\boldsymbol{b}, \boldsymbol{d}, \boldsymbol{f}, \boldsymbol{y}\}$ into $\boldsymbol{c} \perp\!\!\!\perp \boldsymbol{x} | \{\boldsymbol{b}, \boldsymbol{d}, \boldsymbol{f}\}$. Using the global Markov property, we can draw the further conclusion that $\boldsymbol{c} \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{f}$. Thus given *Walking Ability*, then *Pathology* is unrelated to *Pain Severity.* We can also use these properties to make a stronger statement of independence, that is to say that if two sets of variables u and v have no edges connecting one set to the other then $\boldsymbol{u}$ and $\boldsymbol{v}$ are marginally independent of one another ($\boldsymbol{u} \perp\!\!\!\perp \boldsymbol{v}$). So from our example, we can see that *Operation Type* ($\boldsymbol{y}$) is marginally independent of all other variables in the model, and is thus not informative about any of those variables.

The local Markov property is especially useful when considering predicting one set of variables from the others and reflects the ideas of irrelevance discussed previously. Suppose we wish to predict the values of $Y \subset V$ from the values of the remaining variables $X = V \setminus Y$. If we are given the independence graph over V, we can partition X into into the set X_b, which contains the variables which are connected to at least one variable in Y (the *boundary* of Y, $\mathrm{bd}(Y)$), and X_r which contains the remaining variables. We know from the local Markov property that

$Y \perp\!\!\!\perp X_r | X_b$. This statement is equivalent to saying that the conditional density of Y given both X_b and X_r can be written in terms of X_b alone, i.e. $f_{Y|\{X_b,X_r\}} = f_{Y|X_b}$. Consequently, given we know the values of X_b, no further information for predicting Y can be obtained from X_r. Thus, X_b are the called the 'optimal predictors of Y' by Whittaker [125].

5.1.3 Association and Causality

Associations between pairs of variables are identifiable in a graphical model. *Direct* associations occur when two variables are immediately adjacent in the model graph. We know that the set of variables connected to a quantity of interest represent the variables which are most strongly associated with that quantity and are the set of its 'optimal' predictors. For example, in Figure 5.1, we can see that *Night Pain* (***d***) is directly connected to *Pain Frequency* (***b***) and *Pain Severity* (***c***).

Indirect associations occur between two variables which are connected via a path through other intermediate variables. For example, since there exists a path between *Pathology* (***x***) and *Pain Severity* (***c***), then *Pathology* will have an indirect association with *Pain Severity* via the intermediate variable *Walking Ability*. Hence, since there is a relationship between *Pain Severity* and *Walking Ability* and between *Walking Ability* and *Pathology*, then changes in *Pathology* will be indirectly associated with changes in *Pain Severity* via changes in *Walking Ability*. These associations are indirect and only important in the case where we may wish to predict *Pain Severity* when its immediate neighbours are unobserved.

The application of graphical models to such data can be useful as it helps to reveal its underlying association structure. Statements of conditional independence about variables and their relationships and interactions can often, as Cox and Wermuth [23] said, "point towards explanations that are potentially causal." This is particularly the case if the model includes a temporal aspect. However, statisticians traditionally prefer to deal with a world of correlations and associations, and are cautious about drawing causal conclusions from their analyses. It is generally held that an association between a treatment and a particular outcome does not imply that the treatment *caused* that outcome. That said however, results from

well-conducted randomised trials are generally regarded as providing good evidence of a causal relationship between treatment and outcome.

The area of causality has traditionally been the domain of the philosopher rather than the statistician, though several probabilistic frameworks for causal inference have been developed, such as Rubin's causal model [109] and Pearl's causal graphs [95]. A detailed coverage of these topics is outside the scope of this thesis, however a brief summary of Eells' theory [40] is as follows. An event C (e.g. treatment) can be said to have a causal influence on an event E (e.g. outcome) if two criteria are satisfied:

1. C occurs before E,

2. For some *carefully chosen set of conditions* $K_1, \ldots, K_n$,

$$\mathrm{P}[E|C, K_1, \ldots, K_n] \neq \mathrm{P}[E|\overline{C}, K_1, \ldots, K_n]$$

or equivalently

$$C \not\perp\!\!\!\perp E|K_1, \ldots, K_n.$$

This is intuitively reasonable, since if C was known to cause E given the conditions K_i were satisfied, then we would expect a dependency between the two, as the occurrence of C would influence the probability of the occurrence of E. However, the pivotal statement in this definition is the phrase "carefully chosen set of conditions". We may find in an analysis that C and E were conditionally dependent. However this dependency may be due to some spurious association with a common cause K such that C and E are conditionally independent given K. It is the identification and inclusion of *all* of these extraneous and potentially influential conditions that would enable a transition between association and causality.

When working with temporally ordered data it can be very easy to erroneously draw causal statements from graphical models, particularly if the cause variable C precedes the effect variable E and they are associated with one another. However, in these cases the model would rarely satisfy Eells' second criteria as this would require including all possible influential factors into the model to prevent any possibly spurious associations being designated as causal relationships. For this reason, care must be taken when interpreting graphical models and the relationships they model.

5.1.4 Types of Models

5.1.4.1 Pure Discrete Models

A prominent advantage of graphical models is that they allow the inclusion of both discrete and continuous data in the same model. These so-called *mixed models* are a combination of two standard methods for dealing with exclusively discrete or exclusively continuous data. In the case where we only have discrete variables (known as a *pure discrete* model), the model corresponds to a loglinear model for discrete data. Loglinear modelling has been a popular method which has had extensive treatment in the literature, though it has now been somewhat superseded by usage of GLMs. To describe such models in the briefest of terms, suppose we had three discrete variables A, B and C with $\#A, \#B$ and $\#C$ levels respectively. We would then construct the contingency table for these variables and calculate the cell probabilities $p_{jkl} = \mathrm{P}[A = j, B = k, C = l]$. The simplest model for these three variables is that of independence which would be expressed via the logarithm of the cell probabilities that is written, using the notation of Edwards, as:

$$\ln p_{jkl} = u + u^A_j + u^B_k + u^C_l$$

where the u's are the unknown *main effects* . .

Since we are dealing with the log of the cell probability, the additivity of the terms here corresponds to multiplicity in the original scale. Therefore we could re-express this model as $p_{jkl} = p_{j++}p_{+k+}p_{++l}$, where $+$ denotes summation over the respective index. Thus the cell probability is written as the product of the marginal probabilities, thereby demonstrating the independence of A, B and C. To represent a dependency, say between variables A and B, we would add a two-factor interaction term u^{AB}_{jk} into the model. For a more substantial coverage of loglinear models, the method is described in many textbooks on discrete data analysis such as that by Agresti [1].

5.1.4.2 Pure Continuous Models

The pure continuous case corresponds to *graphical Gaussian models* [125] or *covariance selection models* [29]. Instead of having a multi-dimensional contingency table

as with discrete data, our data is now in the form of a series of vectors of continuous quantities. To model these data, we first suppose that $Y = (Y_1, \ldots, Y_q)^T$ is a q-dimensional random vector which has distribution $Y \sim \mathcal{N}_q(\mu, \Sigma)$ where $\mu = (\mu^i)$ and $\Sigma = (\sigma^{ij})$ for $i = 1, \ldots, q$, $j = 1, \ldots, q$. Covariance selection modelling is especially interested in the *precision* or *concentration* matrix $\Omega = \Sigma^{-1} = (\omega^{ij})$. The reason for this focus on the inverse variance is that it has beneficial interpretations when we consider the conditional distribution of $(Y_1, Y_2)|(Y_3, \ldots, Y_q)$. The correlation between Y_1 and Y_2 in this conditional distribution is the partial correlation of these two variables given the rest, written $\rho^{12 \cdot 3 \ldots q}$. This partial correlation has the property that:

$$\rho^{12 \cdot 3 \ldots q} = 0 \iff \omega^{12} = 0. \tag{5.4}$$

If $\omega^{12} = 0$, then the joint density of (Y_1, Y_2) can factorised into the product of their marginal densities. Thus by (5.3):

$$Y_1 \perp\!\!\!\perp Y_2 | (Y_3, \ldots, Y_q) \iff \omega^{12} = 0. \tag{5.5}$$

If we have three continuous variables X, Y and Z, then the independence model would equate to setting $\Omega = \text{diag}(\omega^{XX}, \omega^{YY}, \omega^{ZZ})$. To allow a dependency between X and Y we would allow ω^{XY} to be non-zero. Thus these elements of the inverse matrix play the same role as the two-factor interaction terms in the discrete models and these graphical Gaussian models are defined by setting certain elements of Ω to zero.

5.1.4.3 Mixed Models

Graphical models for mixed discrete/continuous data were introduced by Lauritzen and Wermuth [83]. The models for mixed data to be considered here are essentially a combination of both loglinear models and covariance selection models which are known as *hierarchical interaction models* and were introduced by Edwards [37].

Suppose we have p discrete variables and q continuous variables, where the sets of these variables are denoted as Δ and Γ respectively. Now write the corresponding random variables as (I, Y), and a single observation as (i, y). Here i is a p-tuple containing the values of the discrete variables (i.e. the cell of the contingency table

in which this observation lies), and y is a vector in $\mathbb{R}^q$. Further suppose that the probability of observing this combination of discrete values, i.e. our cell probability, is $\mathrm{P}[I = i] = p(i)$, and that given $I = i$ the distribution of Y is normal $\mathcal{N}_q(\mu(i), \Sigma(i))$ where both the conditional mean and variance may depend on i. This is called the *conditional Gaussian* (CG) distribution, which has density:

$$f(i, y) = p(i)|2\pi\Sigma(i)|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y - \mu(i))^T \Sigma(i)^{-1}(y - \mu(i))\right\}. \qquad (5.6)$$

This distribution can be thought of as a being a different multivariate Normal distribution for the continuous data that is found in each of the cells of the contingency table over the discrete data.

The parameters of this distribution are therefore $\{p(i), \mu(i), \Sigma(i)\}_{i \in \mathcal{I}}$, where $\mathcal{I}$ is the set of all possible i. These parameters are known as the *moments* parameters. The dependency of Σ on i results in a model that is *heterogeneous* - this means that we allow the variance of the Normal distributions to change across different cells in the contingency tables. The nature of these changes in Σ will depend on which discrete/continuous interactions are present in the model. For example if all the discrete variables were independent of the continuous variables then we would only fit a single Normal distribution to all of the continuous data and Σ would not depend on i. Conversely, if all possible discrete/continuous interactions were present in the model then there would be a different Σ for every cell in the contingency table.

Removing the dependency and making Σ constant over i gives a *homogeneous* model - this equates to assuming a single variance matrix to all available data irrespective of the values of the discrete variables and any discrete/continuous interactions in the model. It should be noted that a model with one continuous variable and several discrete quantities is similar to analysis of variance (ANOVA) models, the only difference being that cell counts are treated as random in this framework. Leaving the discrete variables as fixed quantities will result exactly in the ANOVA setup. Thus the graphical modelling methodology provides an encompassing framework for a variety of modelling methods.

The CG density function can be re-expressed in a more compact form:

$$\ln(f(i, y)) = \alpha(i) + \beta(i)^T y - \frac{1}{2} y^T \Omega(i) y. \qquad (5.7)$$

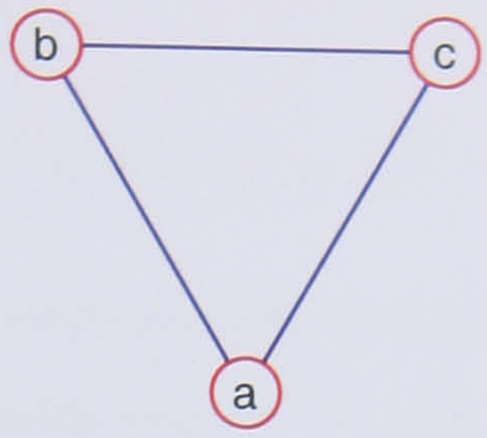


Figure 5.2: A complete graph on three vertices.

where α_i is a scalar, $\beta_i \in \mathbb{R}^q$ and Ω_i is a $p \times p$ positive-definite matrix. These are known as the *canonical* parameters. Hierarchical interaction models are constructed by expanding these canonical parameters into the sums of interaction terms, the models are then defined in a similar way to the loglinear models by setting higher-order interaction terms to be zero. Thus we obtain an interaction expansion for the density of the form:

$$\ln(f(i,y)) = \sum_{a \subseteq \Delta} \lambda_a(i) + \sum_{a \subseteq \Delta} \eta_a(i)^T y - \frac{1}{2} \sum_{a \subseteq \Delta} y^T \psi_a(i) y \tag{5.8}$$

where the sum is taken over all subsets, a, of the discrete variables Δ.

5.1.5 Model Properties

5.1.5.1 Graphical Models

An important subclass of these hierarchical interaction models is the class of models which are *graphical*. These models are defined by setting a set of two-factor interaction terms (and their higher-order relatives) to zero. That is to say, the higher-order interactions are exclusively determined by the presence or absence of the two-factor interaction terms. The significance of these higher order terms is not explicitly tested as one might in a regression context, where after including main-effects terms one then could test for the presence of pairwise interactions.

The benefit of adopting this hierarchical approach allows the models to be interpreted solely in terms of conditional independence. The simplest example of

a non-graphical model is a discrete model with three variables A, B, and C with interactions expressed as :

$$\ln p = u + u^A + u^B + u^C + u^{AB} + u^{AC} + u^{BC}.$$

This model includes all pairwise interactions between the three variables u^{AB}, u^{AC}, u^{BC}, but excludes the higher-order term u^{ABC} and so is non-graphical. Despite its non-graphical nature, the independence graph for this model is given in Figure 5.2. This graph is identical to the independence graph of the complete model with all possible interactions since the pairwise relationships are the same in both cases. If we did not restrict ourselves to models which satisfy the graphical property, then there would be no way of knowing which higher-order interactions were present in the model by looking at the independence graph alone. Thus Figure 5.2 could equally represent either of these models. By restricting ourselves to considering graphical models then Figure 5.2 would always represent the model including the third-order interaction u^{ABC}.

5.1.5.2 Decomposable Models

A second subclass of these models are the *decomposable* models. Whilst being somewhat harder to understand intuitively, decomposable models have beneficial properties when fitting a graphical model to data. Namely, the maximum likelihood parameters of a decomposable model have closed-form expressions meaning that they can be fitted quickly and without the need of iterative fitting algorithms. Furthermore, various computational and theoretical aspects become far more manageable within the decomposable framework than when dealing with a general graphical model. They also have a beneficial interpretation, namely that a decomposable graphical model can be 'decomposed' into a sequence of univariate conditional regressions.

We define a graphical model as being decomposable if and only if:

1. the model graph is triangulated;

2. the model graph does not contain any path between two non-adjacent discrete vertices passing through only continuous vertices.

Whilst there are many definitions for decomposability the above 'forbidden path' definition is the most easily comprehensible. The exact details and theory behind decomposability is a large and expansive area that is well beyond the scope of this thesis. See Lauritzen [81] and chapter 12 of Whittaker [125] for details.

5.1.6 Models, Graphs and Formulae

Graphical models and their associated independence graphs provide a compact representation for the complex association structures of the data and the underlying parametric model. However, it is often useful to be able to express the model in terms of a model formula, rather than relying on the independence graph for its definition. Indeed for a given independence graph for mixed data there are two associated graphical models - the homogeneous model and the heterogeneous model. Therefore, a model formula is useful to eliminate this potential confusion.

Both Edwards and Whittaker present methods for representing a graphical model in a formula, and it is the method of Edwards that shall be discussed here. Given an independence graph for a graphical model, we express the model formula in the following form:

$$\underbrace{d_1, \ldots, d_r}_{\text{discrete}} \;/\; \underbrace{l_1, \ldots, l_s}_{\text{linear}} \;/\; \underbrace{q_1, \ldots, q_t}_{\text{quadratic}} . \tag{5.9}$$

The formula is composed of the following three main components, each being a set of *generators*:

1. The **discrete** generators, d_j, specify the interaction expansion for $\alpha(i)$ as defined in (5.7). These generators are the cliques of $\mathcal{G}_\Delta$, where $\mathcal{G}_\Delta$ is the subgraph of $\mathcal{G}$ over the discrete variables.

2. The **linear** generators, l_j, specify the interaction expansion for $\beta(i)$, the linear coefficient of y in (5.7). These generators are the cliques of $\mathcal{G}_{\Delta \cup \{\gamma\}}$ that contain γ for each $\gamma \in \Gamma$. Note the union of these cliques for a given γ form the discrete boundary of γ and so represent those discrete variables which are directly associated to a continuous variable γ.

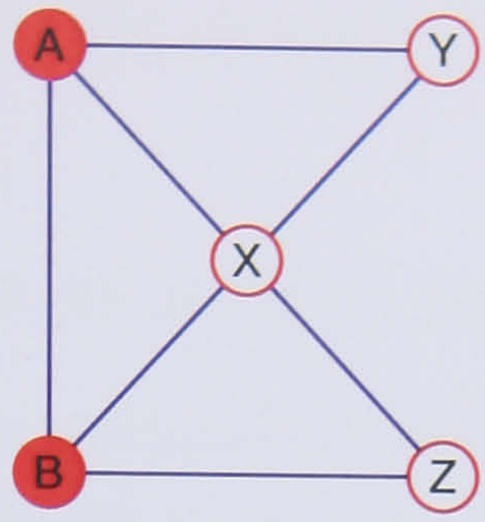


Figure 5.3: A graph on five vertices.

3. The **quadratic** generators, q_j, specify the interactions for the inverse covariance matrix $\Omega(i)$, and depend on whether the model is homogeneous or heterogeneous. If it is homogeneous, then the generators are simply the cliques of $\mathcal{G}_\Gamma$ since the discrete variables do not affect the expansion of $\Omega(i)$. If the model is heterogeneous then the generators are the cliques of $\mathcal{G}$ which intersect Γ, i.e. the cliques of the model graph containing at least one continuous variable.

The reason for defining the model formula in terms of cliques is that the cliques of the graph represent the set of maximal interactions for the model and so are the most compact representation for the model formula. To illustrate the relationship between an independence graph and a model formula, consider Figure 5.3 where we have an independence graph for variables $\Delta = \{A, B\}$ and $\Gamma = \{X, Y, Z\}$. The cliques of this graph are AXY, ABX, BXZ.

The discrete generators for this graph are the cliques of the subgraph over the discrete variables - simply AB. The linear generators must be considered for each continuous variable and so are ABX, AY, BZ. The quadratic generators for the homogeneous model are the cliques of the subgraph of the continuous variables which are XY, XZ. The heterogeneous quadratic generators are the cliques of the model graph containing at least one continuous variable, which are all the cliques of the graph AXY, ABX, BXZ. So we have two possible formulae for this graph -

the homogenous formula:

$$AB/ABX, AY, BZ/XY, XZ$$

and the heterogeneous formula:

$$AB/ABX, AY, BZ/AXY, ABX, BXZ.$$

The process of obtaining an independence graph from a model formula is far simpler. Beginning with a graph with no edges, for every pair of variables (x, y) that appear in the same model generator in the model formula we connect their respective nodes in the model graph. We can then quickly construct an independence graph and determine which interactions are present in the model.

5.1.7 Likelihood, Fitting and Software

Whilst it is perhaps difficult to grasp immediately the reasons for the nature of this model formula given in (5.9), if we consider the model likelihood then the benefits of this specification format may become apparent. To consider the model likelihood we first need some data. Let $(i^{(k)}, y^{(k)})$ for $k = 1, \ldots, N$ be a sample of N independent, identically distributed observations of the variables (I, Y). Now, let (n_i, t_i, SS_i) be the observed counts, variate totals and uncorrected sums of squares and products. For $a \subseteq \Delta$, the marginal cell corresponding to i is written as i_a; similarly for $d \subseteq \Gamma$, the subvector of y is written y^d. We then express the marginal sample statistics corresponding to $a \subseteq \Delta$ and $d \subseteq \Gamma$ as $(n_{i_a}, t^d_{i_a}, SS^d_{i_a})$.

Now let us consider a model with the formula $d_1, \ldots, d_r/l_1, \ldots, l_s/q_1, \ldots, q_t$. It can be shown that a set of minimal sufficient statistics is given by:

1. The set of the marginal tables of the cell counts $\{n_{i_a}\}_{i_a \in \mathcal{I}_a}$ corresponding to the discrete generators, i.e. $a = d_1, \ldots, d_r$.

2. The set of the marginal variate totals $\{t^\gamma_{i_a}\}_{i_a \in \mathcal{I}_a}$ corresponding to the linear generators, i.e. $a = l_j \cap \Delta$,$\gamma = l_j \cap \Gamma$, for $j = 1, \ldots, s$.

3. The set of the marginal tables of the uncorrected sums of squares and products $\{SS^d_{i_a}\}_{i_a \in \mathcal{I}_a}$ corresponding to the quadratic generators, i.e. $a = q_j \cap \Delta$, and $d = q_j \cap \Gamma$ for $j = 1, \ldots, t$.

As these generators dictate a choice of sufficient statistics, they are also pivotal to the likelihood equations. If the graphical model is decomposable then the maximum likelihood estimates of the canonical parameters have closed form expressions. However in general, an iterative procedure known as the MIPS algorithm [53] is used to obtain the parameter estimates. The fitted marginal counts, totals and sums of squares are denoted as m_{i_a}, $ET^{\gamma}_{i_a}$, and $ESS^{d}_{i_a}$. A starting point for the algorithm is required but the fitting process is insensitive to this so typical values for the moments parameters are taken to be $m_i = 1$, $\mu^{\gamma}_i = 0$, and $\Sigma_i = I$ for all $i \in \mathcal{I}$ and $\gamma \in \Gamma$.

We first convert the initial parameters and the sample statistics to canonical form. The algorithm consists of a series of cycles, each of which is composed of three steps. Each step takes the same form by updating the current canonical parameter estimates by adding a value calculated from our sample statistics and subtracting a value computed from the fitted quantities. For example, one step updates the discrete canonical parameters, $\alpha(i)$ for each i using the following update rule:

$$\alpha(i) = \alpha(i) + \ln(n_{i_a}) - \ln(m_{i_a})$$

for each $a \in \{d_1, \ldots, d_r\}$, i.e. a is the set of variables contained in a single discrete generator. The update steps are similar, though more complex, for the other steps. These cycles are continued until convergence occurs. If the model is decomposable then the algorithm will converge after the first iteration provided the steps are performed in a specific order.

We can see that the model generators given in the model formula are not only an efficient specification in terms of the models interactions, but also for sufficient statistics and the calculation of the parameter estimates.

Having obtained a set of parameter estimates, we can then use these parameter estimates to further deepen our understanding of the associations between pairs of variables that thus far have only been shown to be present or absent. Examination of the fitted parameters will give quantitative information on the magnitude and direction of these associations. For example, we know that the 'optimal' predictors for a given variable are its immediate neighbours on the model graph so we can use the fitted distribution to obtain the equation for this relationship. This can

be achieved by marginalising the CG distribution and, essentially, integrating (or summing) out unwanted variables. In the case of a group of continuous variables we will obtain a linear equation for our dependent variable. In the case of discrete variables we will obtain a marginal table of probabilities. However, if the data are mixed then care must be taken when performing such marginalisation. For example, if we have a binary discrete variable, b, and a number of continuous quantities then we will obtain two fitted Normal distributions, one for each level of b. If b is independent of the continuous variables then the fitted distributions will be the same and we can safely marginalise over b. If this is not the case and the two distributions are different, then marginalising out b will not result in a Normal distribution and the values obtained will not be appropriate.

Unfortunately, standard errors are not available for the fitted parameters of graphical model obtained in the manner described. Therefore it is not possible to construct confidence intervals or to perform significance tests on the fitted parameters. However, for pure continuous models some work by Roverato and Whittaker [108] has revealed the form of these standard errors which are expressed in terms of the Isserlis matrix of Σ [64].

The fitting of graphical models is a complex process and is typically accomplished via specialised software packages tailored to manage such statistical models. One such package is the application MIM [38] which allows mixed data modelling of both undirected graphical models and chain graphs (chain graphs will be fully discussed in Chapter 8). MIM supports a wide range of edge deletion methods from standard deviance difference tests and small sample F tests to tests specific to particular data types such as Fisher's exact test and the Kruskal-Wallis test. A further software application dealing with graphical models is the DIGRAM package [74] which tackles the problem of chain graph models for discrete variables again supporting a wide range of tests including Pearson's χ^2 and partial γ-coefficients for assessing conditional independence of variables. CoCo [6] is another program designed to deal with pure discrete models and analyse contingency tables. CoCo contains a variety of exact conditional tests for decomposable models as well as supporting model search via the BIC criterion or p-values. TURNER [79] is another package designed

for analysis of discrete data with loglinear methods that are visualised by graphical hierarchical loglinear model.

5.1.8 Model Selection

The selection of a final graphical model consistent with the data is a complex and computationally difficult process. Two of the simplest methods are:

1. Backward stepwise selection from the full saturated model over the variables,

2. Forward stepwise selection from the independence (main effects) model over the variables,

Both methods begin with an initial model and then attempt to include or exclude additional edges into the model. There are various methods of testing the importance of the inclusion or exclusion of an edge into or from the model. However the method that will be discussed here is that of the asymptotic χ^2 likelihood ratio test. In the case of backward selection, the saturated model is first fitted to the data and its likelihood calculated. Then the candidate edge is removed and this sub-model is fitted, and its likelihood calculated. The resulting likelihood ratio is then compared to a χ^2 distribution with the appropriate degrees of freedom to obtain a significance probability. After testing all edges in the model, the least significant edge is removed.

For forward selection the process is reversed, with the most significant edge being added at each stage. The advantage of backward selection is that it begins with a complete model that will be consistent with the data and the model selection prunes away any unnecessary dependencies from the model. The key disadvantages with this approach are that with large problems it may not be possible to fit the full saturated model due to the potentially huge number of parameters and will typically result in models that are over-complicated. In such cases a forward selection strategy may be employed where we begin with the independence model and include significant edges. However, in this case we are starting with a model that is likely inconsistent with our data and are seeking to improve its performance.

Alternatives to the χ^2 approach include a deviance-based F-test, exact conditional tests and, for specific data, tests such as the Jonkheere-Terpstra test for

ordinal data. Model comparison can also be based on the information criteria AIC and BIC which can be expressed as functions of the maximised likelihood under the model. The various model comparison tests are fully described in chapter 6 of Edwards [38]. Drton and Perlman [35] propose an alternative stepwise selection method to those given in Edwards, which directly tests for zero partial correlations in covariance selection models whilst controlling the error rate for incorrect edge inclusion.

There are, of course, alternatives to the stepwise model selection discussed here - a global model search for example is a process whereby the space of all possible models is searched for one or more suitable models. Roverato and Paterlini propose such a model selection methodology using genetic algorithms [107], whereas Edwards and Havránek developed the EH-Procedure [39]. These global search methods have the advantage that they are less likely to overlook good models than the stepwise selection methods. However, they also suffer from the problems due to the sheer size of the model space that they are searching when the dimensionality is large. If there are m variables in the model then there are $m(m-1)/2$ possible edges in the graphical model for these variables, and hence a total of $2^{m(m-1)/2}$ possible graphical models to consider. This exponential increase in the size of the search space can render these methods either impractical or computationally infeasible. However, with smaller-sized models these methods will likely provide superior results than the simple stepwise methods.

5.2 Application of Graphical Models to the Orthopædic Data

5.2.1 Methodology

The computer package MIM [38] was used to fit graphical models to the various data sets using the `fit` or `cgfit` commands. Models were selected via the `stepwise` command. For both data sets mixed graphical models were constructed by using a forward selection strategy from the main effects model over the patient status

variables plus additional demographics or interesting factors. The reason for using forward selection rather than backward elimination was because the complete saturated model contained too many parameters to be determined using available computer memory. Additionally, to use backward elimination on models when we have so many variables would likely result in a model that was over-complicated. A forward selection strategy would instead lead to a simpler final model for the data. The forward selection was performed in the space of heterogeneous models. That is to say that the variance of the conditional Gaussian distribution was allowed to vary between cells of the contingency table. The assumption of homogeneity whereby all variance matrices were considered to be equal was considered to be unreasonable and tests of this assumption corroborated this.

The edge selection process used the Bayesian Information Criterion [112] rather than the standard χ^2 tests. The motivation for this is that at any stage in the development of the model, a number of highly significant edges will be eligible for inclusion in the model. Due to their large associated test statistics their p values will be close to 0. If the tests of two distinct edges had the χ^2 values of 50 and 50 000 respectively with the same number of degrees of freedom (typically 1 or 2), both would have p-values which were imperceptibly different from 0. In this case, the first examined edge from the pair would be selected, rather than the one with the largest value of the test statistic. This often leads to models that have many significant edges attached to the first few important *variables* due to their ordering in the data matrix, rather than models that have included the most important *edges*. Use of an information criterion such as AIC [2] or BIC eliminates this problem as the criterion values for every tested edge are compared directly and the most important edge is then included. The reason for favouring BIC over AIC can be given by considering the formulae for both criteria:

$$\begin{aligned} AIC &= -2\hat{L} - 2p, \\ BIC &= -2\hat{L} - \sqrt{n}p, \end{aligned}$$

where $\hat{L}$ is the maximised log-likelihood under the model, p is the number of free parameters of the model and n is the number of observations. We can observe that

since BIC incorporates a term in n, as we have an increasing number of observations a selection method using BIC will favour simpler models compared to selection using AIC. Since the orthopædic data have many observations, we will obtain models that are not overcomplicated and are relatively straightforward to interpret.

5.2.2 Verification of Initial Assumptions

5.2.2.1 Ordinal Data

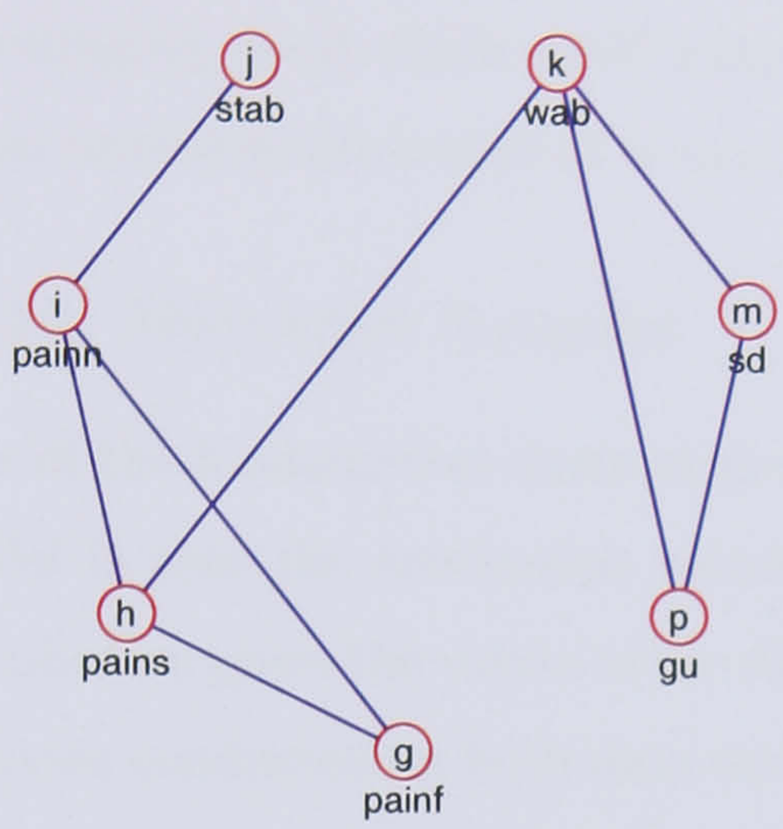
The patient status variables in the orthopædic data sets are either purely ordinal (in the case of the hips data), or a mixture of ordinal and continuous quantities (for the knees data). Both data sets have ten or more of these ordinal variables, each with a total of five possible values. If, for example, we treat these ordinal variables as such when fitting a graphical model to the pre-operative hips data we would seek to build a pure discrete model over the underlying contingency table. However the hips data is composed of twelve five-point ordinal variables resulting in a 12-dimensional contingency table containing in excess of 200 million cells. Including any additional factor variables into this model such as sex, diagnosis, or treatment type would further increase the size and dimensionality of this model.

Needless to say such a model is unwieldy to the point of being impossible to fit for three reasons. Firstly, the vastness of the contingency table means that given the available data we would likely end up with a sparse table with many zero counts due to the fact that a large number of the possible combinations of the values of the variables will be unobserved. Secondly, since the table is so large there would be a correspondingly huge number of parameters to estimate for the fitted model and there would almost surely be insufficient data to do so. Finally, the computation and processing power required to construct and store the contingency of the data is far beyond that of the 2.3GHz Pentium IV machine with 512Mb RAM that was used to fit these models, which is to say nothing of what would be required to perform the stepwise model selection and fitting of each successive model. Even if these three problems were eliminated, the interpretation of such a model would be extremely difficult.

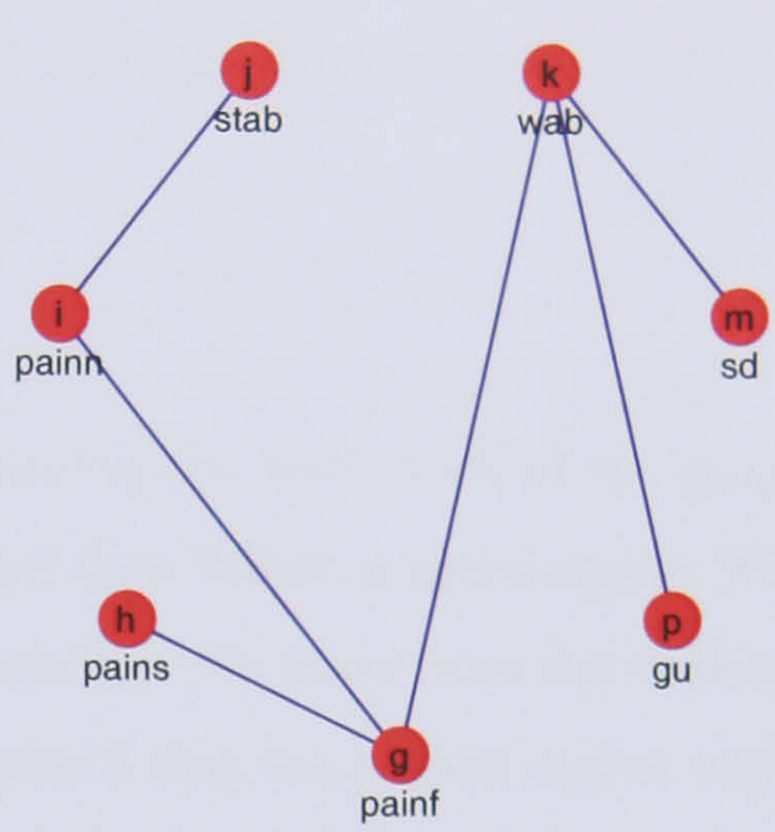
Therefore, before the graphical modelling techniques are applied to the orthopædic data sets we assume, as we did in Chapter 3, that all ordinal variables within both data sets can be approximated as continuous variables for the purpose of this analysis. This changes our patient status models to pure continuous graphical models which correspond to a multivariate Gaussian distribution, or a conditional Gaussian distribution given any further discrete factors included in the model. Using such an approximation will allow for the selection and fitting of a graphical model to the data.

However, the model selection procedure may have differing results when comparing the final model for the ordinal variables with the final approximate continuous model. Whilst the models underlying the ordinal and continuous models are fundamentally different and incomparable, the conditional independence structure of the model can be compared. To investigate this possible discrepancy, models were fitted over a subset of variables from the knees data. Whilst it was not possible to obtain a discrete model over all the ordinal variables, it was possible when restricting ourselves to examining a subset of seven variables. Using the package MIM, the variables could be considered as discrete, ordinal or continuous. Therefore pure discrete and pure ordinal graphical models and an approximate continuous graphical model were fitted to the same subset of seven variables. The three models were obtained by forward selection from their respective main effects models. The resulting model graphs are shown in Figure 5.4.

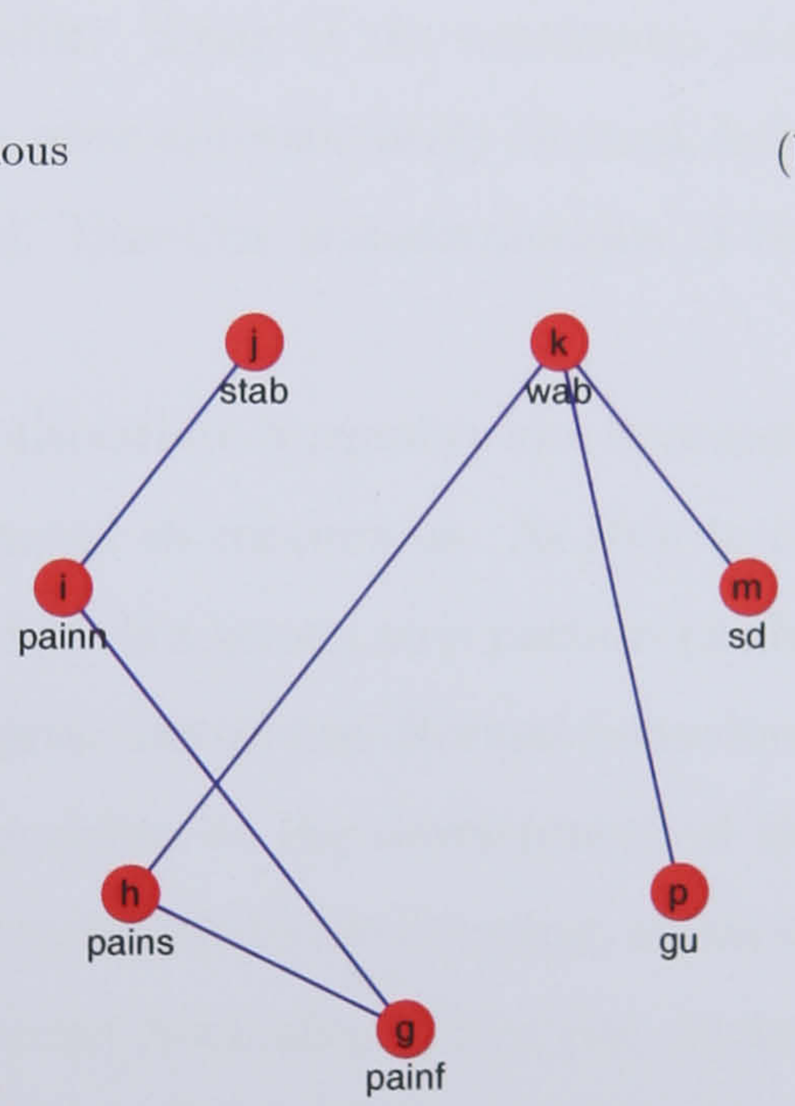
We can see immediately from Figure 5.4 that there is a strong overlap in terms of the detected dependencies between the seven variables with all three graphs being similar. The continuous model introduces arcs which are not present in the ordinal or discrete models - specifically *Pain Severity* (`pains`) and *Night Pain* (`painn`) are joined, as are *Going Up Stairs* (`gu`) and *Sitting Down* (`sd`). Closer investigation reveals that the ordinal model graph is, in fact, a subgraph of the continuous model graph suggesting a good degree of similarity in terms of the independence relationships between the variables. The model graph resulting from treating the data as categorical variables is almost identical to the ordinal model with the exception that in this discrete case it is *Pain Frequency* (`painf`) that is associated to *Walking*



(a) Continuous



(b) Discrete



(c) Ordinal

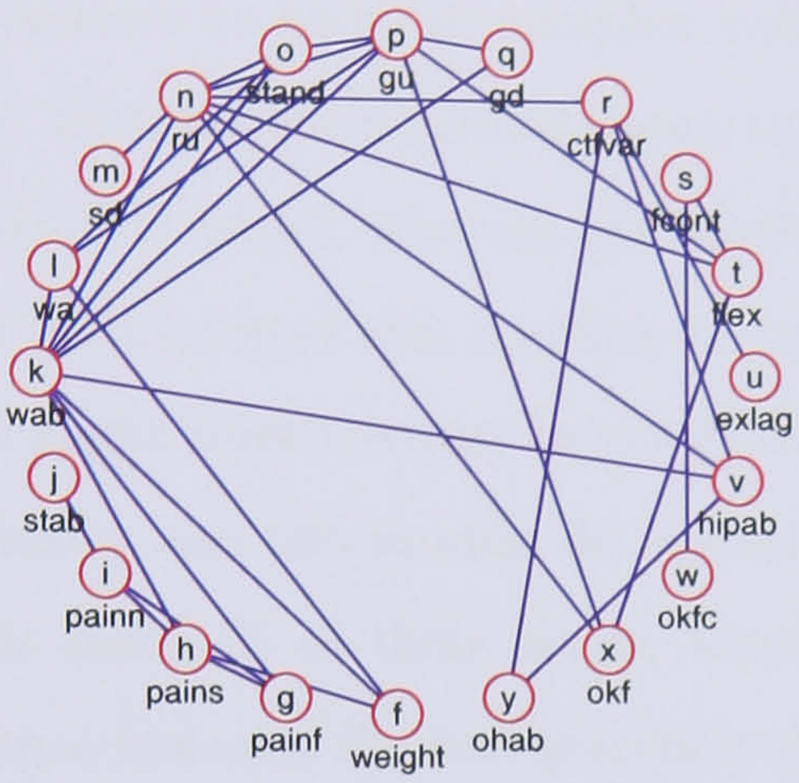
Figure 5.4: Comparison of model graphs obtained for seven ordinal variables from the knees data when the variables are treated as ordinal, discrete or approximated as continuous.

Ability (wab), whereas in the ordinal model this association is with *Pain Severity* (pains) instead. In conclusion, despite some deviations the various models appear to be quite similar in terms of their conditional independence structure despite the fact that the underlying models are fundamentally different. This appears to be an encouraging result which could suggest that a continuous approximation would not be an unreasonable course of action.

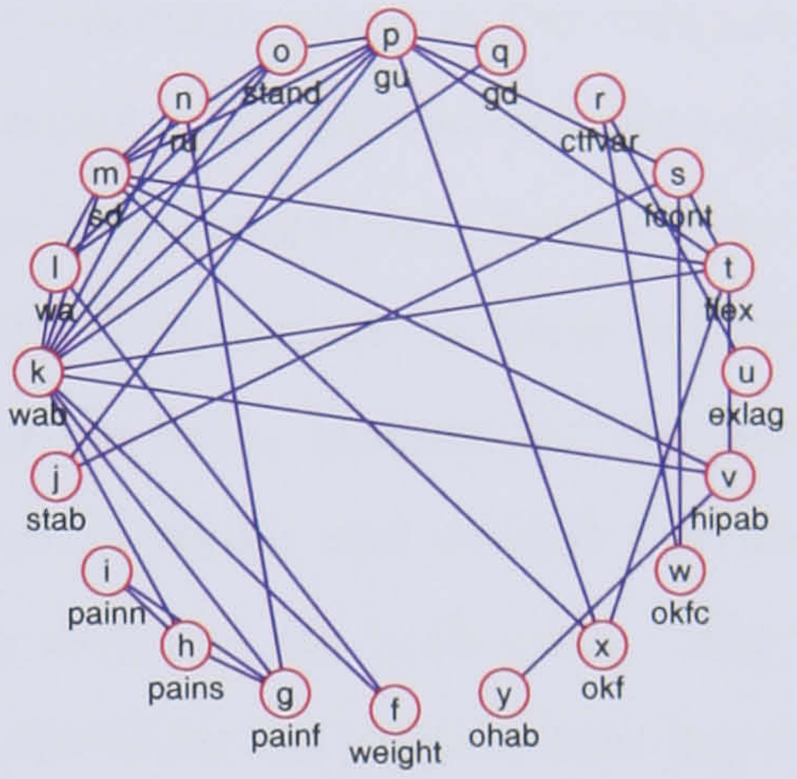
5.2.2.2 Multivariate Normality

One of the fundamental assumptions underpinning the framework of the graphical model is that the continuous variables in the data follow a multivariate Normal distribution given the values of the discrete variables. We know from the exploratory analyses conducted on both data sets in Chapter 3 that the patient status variables seldom exhibited Normality. Some of the continuous variables in the knees data, such as *Age* and *Weight*, were approximately Normal, but others such as *Extension Lag* were heavily skewed. Box-Cox transformations of these variables did little to improve this situation.

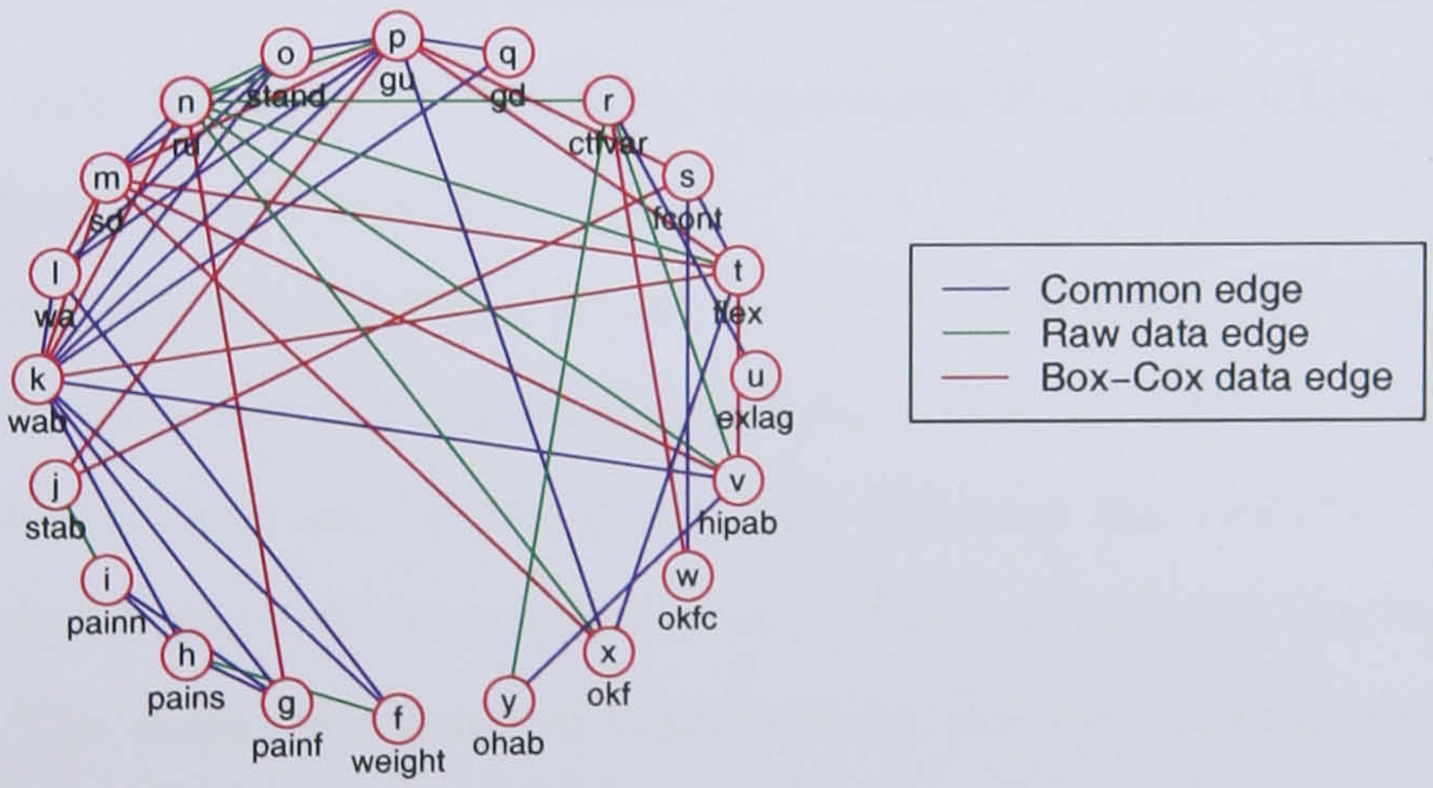
The assumption of multivariate Normality also becomes less valid now the ordinal data are to be approximated as continuous. As seen in the quantile plots of these variables in Chapter 3, there is a strong step pattern produced by the granularity of the ordinal data. This pronounced non-Normal behaviour of the ordinal variables could pose a potential problem to the development of the model. Unfortunately, there is little information available in the literature about the sensitivity of graphical modelling to departures from Normality. There has, however, been some work done on examining the sensitivity of the model selection process to the contamination of data by outliers [77]. It was demonstrated that outlier observations in the data exert a definite influence on the final selected model. Furthermore, it was shown that the degree to which the data set has been contaminated by these outliers governs the level of this influence with the most extreme contamination resulting in the most profound departures from the 'true' model. However, there were typically few extreme observations in the orthopædic data and those cases which were extreme were removed from the analysis.



(a) Raw Data



(b) Box-Cox Transformed Data



(c) Overlayed graph

Figure 5.5: Comparison of model graphs obtained for using the raw knees data and the Box-Cox transformed data.

To assess whether there would be any benefits to performing Box-Cox transformations of the data prior to the modelling process, a pair of models will be constructed one using raw data and the other using the transformed data. The powers used in the transformation are those given in Table 3.1. The resulting model graphs are displayed in Figure 5.5. Firstly, we can see that the model for the transformed data appears to be more complex with apparently more edges in the independence graph. Further investigation shows that the model graph for the raw data contains a total of 35 edges, whereas the model for the transformed data contains 40 edges which corroborates this increase in complexity. This increase is likely attributable to the slight improvement in the Normality of the data. Despite the differences in complexity, the two models do appear to be quite similar and we find that the two models share 25 of their edges, which is the majority for both cases. The main difference between the two graphs is that the model for the original data has several edges emanating from the node $\boldsymbol{n}$, corresponding to *Sitting Down*. The transformed model instead connects these edges with node $\boldsymbol{m}$ (*Rising Up*). Both of these variables are closely correlated to one another ($r = 0.994$) and so one of the pair could be viewed as being a proxy for the other, suggesting that these edges are actually common structure to both models.

In conclusion, we can observe a strong overlap in the independence graphs obtained from the raw and the transformed data. This reflects the fact that the Box-Cox transformations are having a minimal effect on the associations between variables thereby giving rise to a model that is fairly similar to the model of the original data. The main advantage of transforming the data is that we (hopefully) improve the Normality of our data and so are in a better position to apply the graphical modelling techniques. However, the price of obtaining this slight improvement is the loss of the interpretability of our model. Specifically, this is of importance in the clinical setting where information about the relationships between pain measures could be easily understood, but relationships between the pain variables to the powers of 0.68, 0.83 and 0.88 could be meaningless.

Since the model of the transformed data is rather close to the model for our original data we have obtained little practical benefit from performing the transfor-

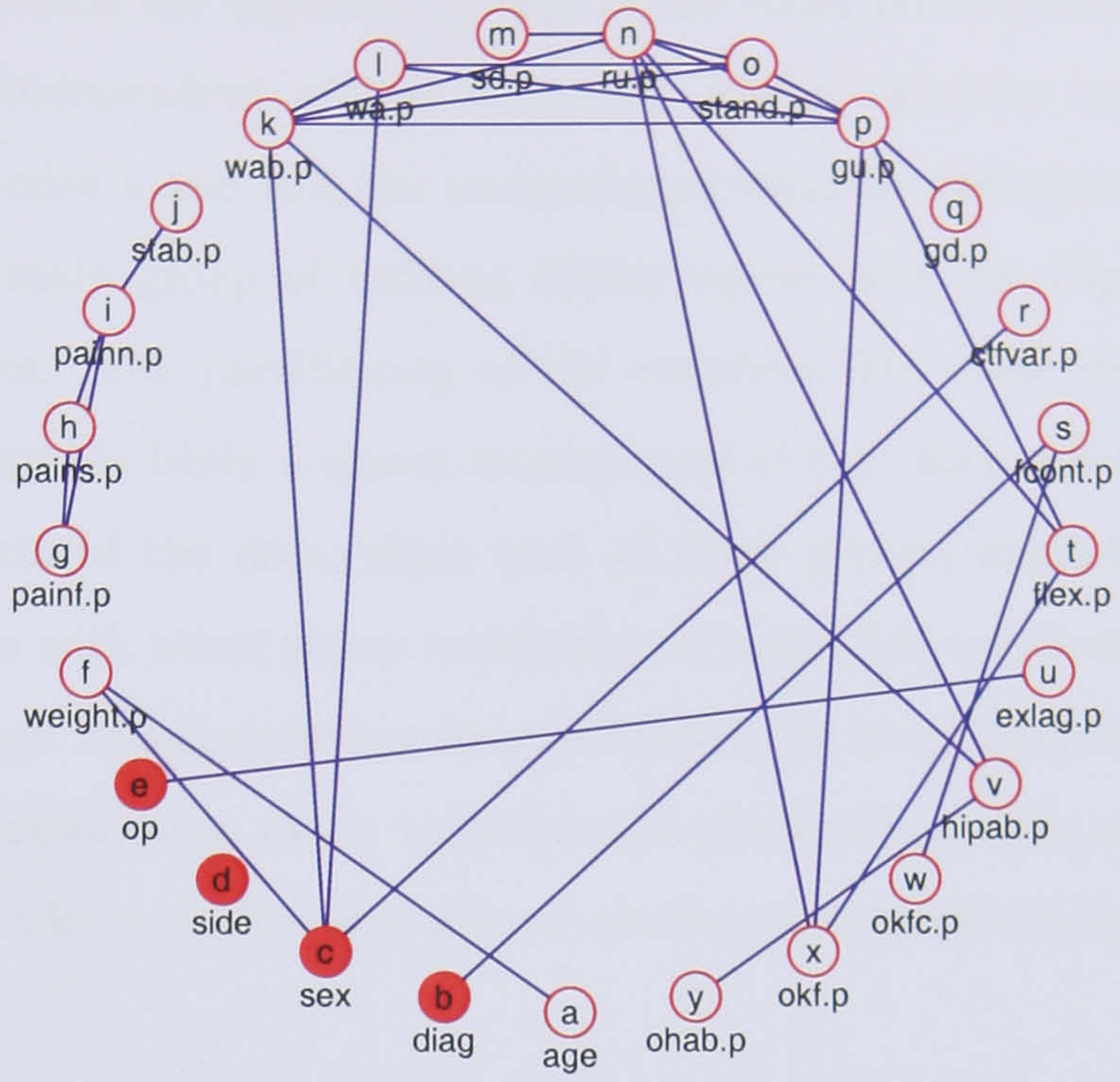
mation. Therefore the original untransformed data will be used, despite the violation of the multinormality assumption.

5.2.3 Results - Knees Data

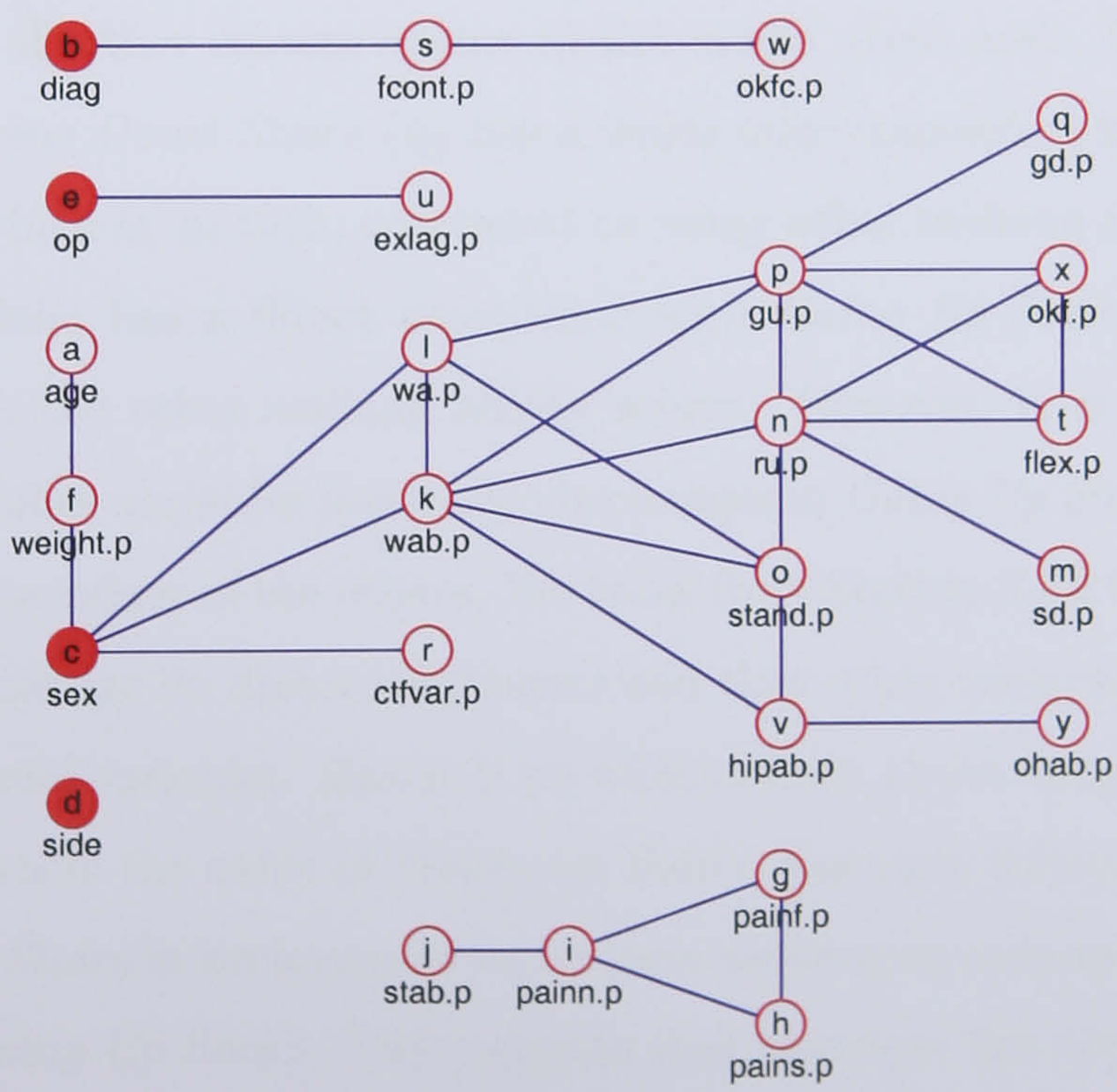
5.2.3.1 Pre-operative Data

Under the above assumptions, the graphical modelling techniques discussed in Section 5.1 were applied to the 20 patient status variables with the addition of five further variables that are either relevant demographics or potentially interesting factor variables. These variables are: *Sex*, *Age*, *Diagnosis*, *Operation* and *Side*, where the final variable encodes whether it is the left or right knee that is to be replaced. A mixed graphical model was constructed by using a forward selection strategy from the main effects model for the 25 variables. The reason for this was that to perform a backward selection strategy from the saturated model where we have so many variables would likely result in a model that was over-complicated. Therefore to obtain a simpler final model for the data a forward selection strategy was employed instead. A further reason for not performing backward elimination from a full saturated model was that the saturated model could not be fitted in MIM. By iteratively including significant arcs into the model we can progressively develop and refine a conditional independence framework for the data. The resulting model graph is displayed in Figure 5.6 where the graph is presented in a standard ring layout and also in a rearranged layout which emphasises the structure present in the model.

Whilst the conditional independence structure can be read from the first graph in Figure 5.6, it is far clearer in the rearranged layout in the second graph. From this graph we can observe that there is a strong group structure present among the variables along the lines mentioned in Chapter 3 and also seen in the correlation plots in Section 4.2. For example we can see that the three pain variables (nodes $\boldsymbol{g}$, $\boldsymbol{h}$ and $\boldsymbol{i}$) are all interconnected indicating that all three variables are mutually dependent. This is to be expected as they measure a similar quantity. The pain variables also appear to be closely related to *Stability* (node $\boldsymbol{j}$). This marginal independence of



(a) Standard Layout



(b) Rearranged Layout

Figure 5.6: Final model for the pre-operative knees data.

these quantities suggests that they represent an aspect of the patient's status that is neither represented nor captured by any of the other patient status variables.

Other such independent groups of patient status variables include the *Fixed Contractures* (nodes $\boldsymbol{s}$ and $\boldsymbol{w}$), the uncorrelated variable *Extension Lag* (node $\boldsymbol{u}$), and finally the main group of walking ability scores plus the *Hip Abduction* and *Flexion* measures. The partitioning of the variables into these distinct groups is to be expected and is likely a direct consequence of the block structure within the correlation matrix of the data, since each of these groups exhibited moderate to high correlations with other group members and negligible correlations to variables outside the group. The correlation structure aside, its is not surprising to observed this group behaviour since many variables are notionally similar, such as the pain scores and the walking ability measures, or are the same measurement taken on both knees (or hips).

A particularly useful feature of graphical models is that one can read conditional independence relationships directly from the model graph. For example, a node A that has only a single edge connecting it to another variable B is conditionally independent of all other measurements in the model given node B. For example, the node for *Going Down Stairs* ($\boldsymbol{q}$) has a single edge connecting it to node *Going Up Stairs* ($\boldsymbol{p}$) which is, in turn, connected to many other walking ability measures. *Going Down Stairs* has a direct association with *Going Up Stairs*, but no direct association with the other walking ability scores. However, it is associated with other walking ability variables *indirectly* via changes in *Going Up Stairs*, which could then influence the values of the others. We know from Section 5.1.2 that a quantity's optimal predictors are its direct neighbours and that other variables are redundant given these optimal variables. Hence, if we wish to learn about the patient's walking ability and we know the value of *Going Up Stairs*, then any information contained in *Going Down Stairs* is irrelevant to us because we have already as much as we can by observing *Going Up Stairs*. This suggests that variables like *Going Down Stairs* which 'hang off' the graph in such a way are uninformative when one wishes to learn about the possible values of other variables in the graph, such as in prediction.

There are many nodes in the model graph which are connected by a single

edge. These commonly appear to occur where we have variables that were strongly correlated, such as *Going Up Stairs* and *Going Down Stairs* as discussed above. However, this pattern also holds for the *Fixed Contractures* (nodes $\boldsymbol{s}$ and $\boldsymbol{w}$), the *Hip Abductions* ($\boldsymbol{v}$ and $\boldsymbol{y}$), and the pair {*Rising Up, Sitting Down*} ($\boldsymbol{n}, \boldsymbol{m}$). Uncorrelated variables such as *Extension Lag* and *CTF Angle* also have only a single association, but in these cases the associations are to discrete variables rather than continuous ones. Both of these variables remain marginally independent of the other patient status indicators in the model.

Since every arc in the model graph represents a dependency between a pair of variables then examining the number of arcs associated with each variable would give an indication of how important a particular variable is in terms of determining values of others. Such values are tabulated in Table 5.1. A variable with a large number of edges forms a 'hub' or 'focus' in the model graph - this is the case with *Going Up Stairs*, *Walking Ability* and *Rising Up* (nodes $\boldsymbol{p}$, $\boldsymbol{k}$ and $\boldsymbol{n}$) which have many connections to other variables. This indicates, for example, that 7 other patient status variables are dependent on the value of *Going Up Stairs* making it a useful variable for determining the values of these other variables. Conversely, we can say that *Going Up Stairs* itself can be best determined using a combination of the same 7 measurements. It should be noted that since the model graph is separated into disjoint groups of variables these three variables are the foci of the group of walking ability variables only, indicating their importance to that particular subgroup of variables. Nonetheless, these variables are assessments of the patients mobility and were all strongly correlated to many other walking ability variables therefore it is not unreasonable to see these variables as foci of that component of the graph.

At the other end of the scale, variables with few arcs feeding into them exhibit greater conditional independence with variables such as *Other Hip Abduction* and *Going Down Stairs* which are conditionally independent from the majority of the model graph given those variables to which they are connected. Variables with no arcs joining them to others, such as *Side*, are marginally independent. This is to say that they are neither informed by nor informative for the values of any other variable.

Variable	Num. Edges	Variable	Num. Edges
Going Up Stairs Walking Ability Rising Up	7	Age Diagnosis Operation Stability Sitting Down Going Down Stairs CTF. Angle Extension Lag OK. Fixed Cont. Other Hip Abduction	1
Sex Standing	4		
Pain Severity Night Pain Flexion Hip Abduction OK. Flexion	3		
Weight Pain Frequency Fixed Contracture	3		
		Side	0

Table 5.1: Table of the number of edges connected to each in the graphical model for the pre-operative knees data.

In terms of the medical implications of this model, firstly we can see that there is definite structure to the various component measurements of the Nottingham Knee Score described in Section 2.2.1. It is clear that the components form a number of distinct groups of closely associated variables. These comprise a large group of walking ability measures plus the flexion and hip abduction scores, a group of the pain scores plus *Stability*, the group formed by the pair of *Fixed Contracture* scores and the two independent variables *Extension Lag* and *CTF Angle*. The fact that the majority of the variables form a single group suggests that there is some overlap in the information represented by these quantities and that by observing all of the variables in this group we will likely introduce some redundancy into the composite score by replicating similar information. There is also evidence of several measurements that are conditionally independent given a small number of other variables suggesting that they are perhaps representing quantities that are more independent of the main body of patient status variables. This may correspond to them conveying relatively novel information that is not captured by the other

variables. If so, these variables would be important inclusions into the composite score as they capture novel information. However conversely these variables will be of little use for prediction as they are minimally related to one another.

Of the demographic variables, *Sex* (***c***) has some association with the walking ability scores via *Walking Ability* itself and *Walking Aids*. There will then be a consequent indirect association with the other variables of the group since there is a path from *Sex* to those other measurements. *Sex* also is the only variable in the model to display an apparent association with the value of *CTF Angle*. These dependencies between discrete and continuous variables correspond to significant sex differences in the values of these measurements. The other two demographic variables were *Age* and *Weight* (***a*** and ***f***). *Age* was only connected to *Weight*, and *Weight* to *Sex*. Hence they had little direct impact on the model, though their association to the *Sex* may indicate a possible, if indirect, bearing on the patient's status.

In terms of the patient's *Diagnosis* and *Operation* (nodes ***b*** and ***e***), neither variable had any major association to the patient's pre-operative status. An association was found between *Diagnosis* and *Fixed Contracture*, but it remained marginally independent of the walking ability and pain measurements suggesting a minimal association between the pathology and the patient's status. The veracity of this assertion appears to be in doubt however, as we observed there to be a noticeable difference between the weights of patients between the two pathology groups. This association in the form of an edge joining nodes ***f*** and ***b*** is not present in the model. Further investigation reveals that such an edge would be highly significant if included into the final model, however doing so would render it neither graphical nor decomposable. The loss of these two properties would pose significant problems for the interpretation and calculation of the model. However, if added the edge would represent a significant and obvious association. Nonetheless, in the model selection it appears to have been overlooked in order to preserve the mathematical properties of the resulting model. This raises the question of how many other potentially significant associations were present in the data that could not be included in the final model due to the constraints of the graphical modelling framework.

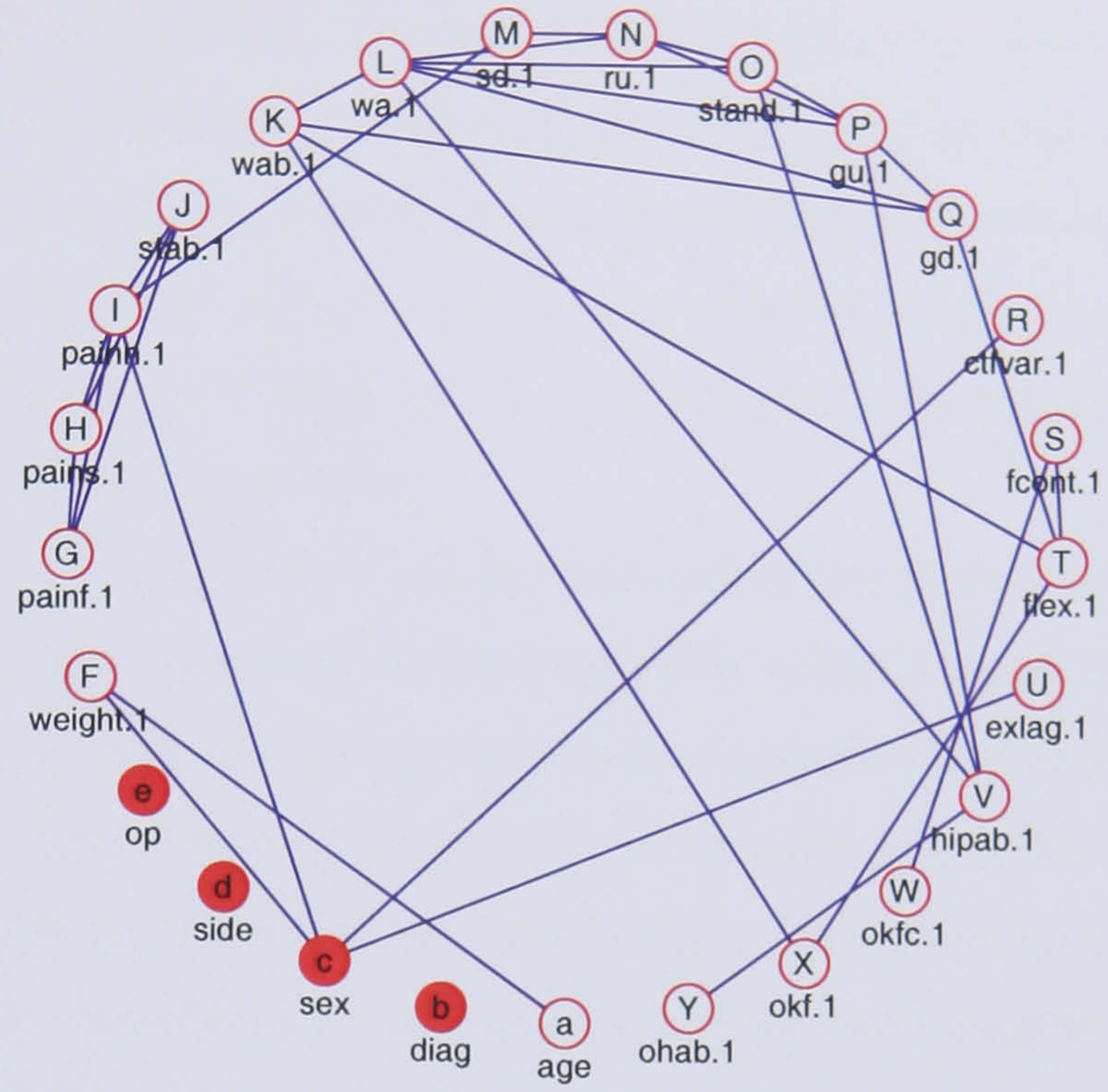
The *Operation* variable is also marginally independent of the majority of the patient status variables with the exception of *Extension Lag* which, under the model, displays a small change in mean and a large change in variance between the two treatment groups. It is important to keep in mind that in chronological terms that the value of *Operation* would be determined ***after*** we have observed the patient's status. Furthermore, since the data was from a trial where *Operation* was randomised we would expect to have no associations. However this association to *Extension Lag* must be present due to either sample variation or as a consequence of the profound non-Normality of the variable resulting in a spurious association. The variable *Side* (node ***d***) is marginally independent of all other variables, implying that the side of the body that is affected has no influence on the patient's condition.

5.2.3.2 One-Year Post-operative Data

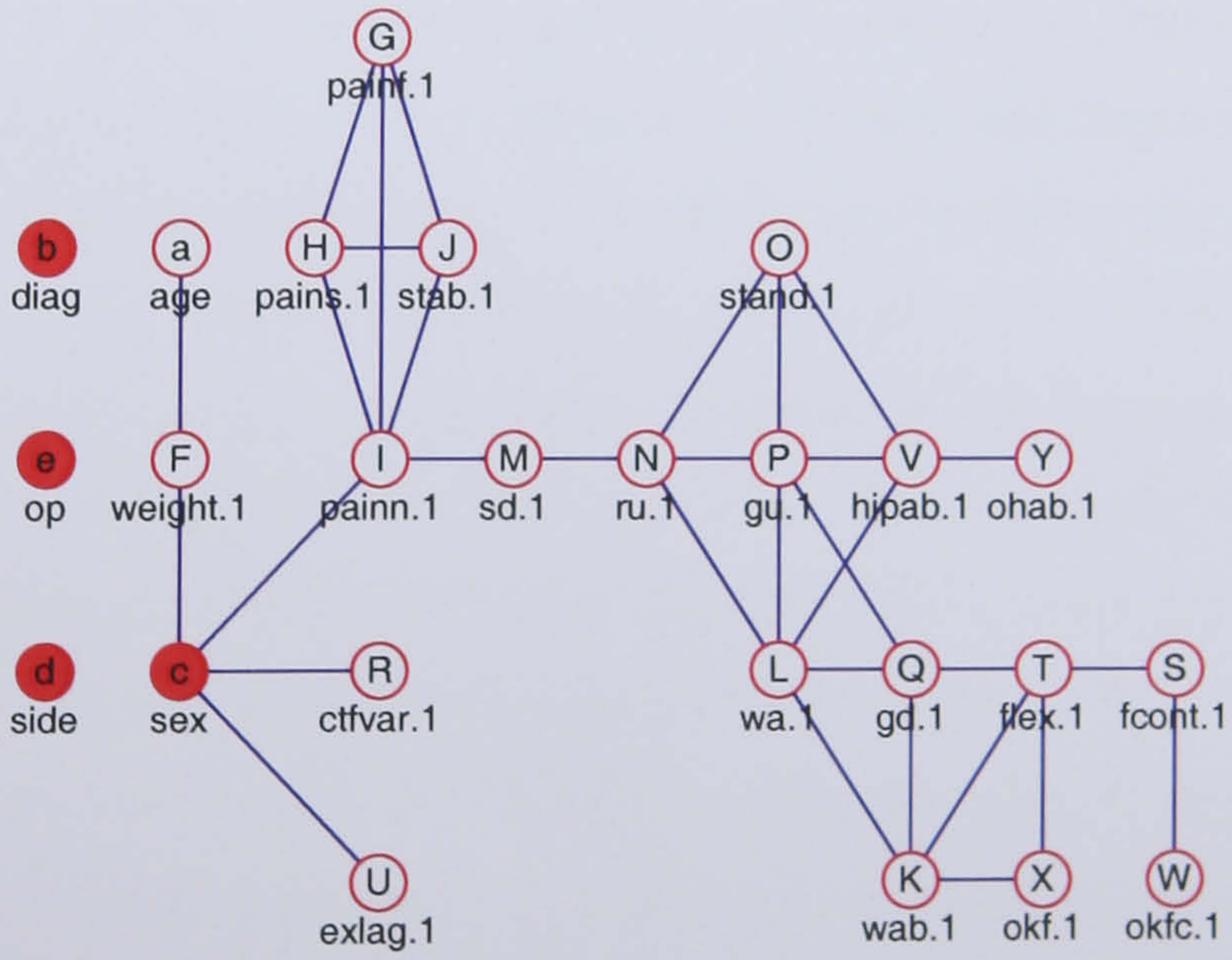
The final graphical model for the 1-year post-operative data was obtained using the same process as with the pre-operative data. The model's independence graph is shown in Figure 5.7 again using both the default ring layout and a rearranged layout to ease the interpretation of the model (note that the rearranged layout for the post-operative model is different from that for the pre-operative model).

We can see by superficial comparison of the model graphs in Figure 5.6 and Figure 5.7 that the pre- and post-operative models differ slightly in terms of the arcs present. Comparison is most easily made between the graphs where the nodes are arranged in a ring since all the variables occur in the same locations in both graphs. From these graphs, we can observe that there are some changes in the edges present, the most obvious of which being those emanating from *Sex* (node ***c***). Looking at the rearranged layout in Figure 5.7, we can notice that post-operatively all the patient status variables are now connected; the disjoint groups of variables seen in the pre-operative model are no longer present. This is likely due to the fact that the post-operative data are more correlated which would suggest more dependencies within the data.

These apparent differences aside, further investigation reveals that 19 of the edges in the pre-operative model remain present in the post-operative graph. This



(a) Standard Layout



(b) Rearranged Layout

Figure 5.7: Final model for the 1-year post-operative knees data.

equates to 63% of the pre-operative edges being preserved, suggesting that there is a persistence in a core number of relationships which remain unaffected by the intervention of the operation. Looking at the table of edge counts in Table 5.2 we can see that *Going Up Stairs*, *Walking Ability* and *Rising Up* now have fewer edges implying their importance in terms of the other variables is reduced. Additionally, the post-operative model differs from the pre-operative model by the increased number of associations with *Walking Aids* - a variable that was pre-operatively far less dependent on other measurements.

The obvious groups of conditionally independent variables observed in the pre-operative model also persist in the post-operative model, though they are perhaps less evident as the variables are typically more connected than before. Nonetheless, the groups of variables such as the pain scores and the pairs of hip abductions and fixed contractures can be seen on the model graph. Their associations have not been disrupted by the intervention of the operation. This could suggest that there exist natural fixed relationships that exist among these variables giving rise to a degree of constancy in the structure of the model.

The three demographic variables of *Age*, *Weight* and *Sex* have similar relationships in the post-operative model as they did pre-operatively. The only differences are in those variables which show a dependency on *Sex*, and hence exhibit notable sex differences such as *CTF. Angle.* The other two variables are *Extension Lag* and *Night Pain.* This suggests that there is still a difference between the status of patients of different sexes post-operatively. However, it also suggests that sex now only directly associates with the patient's pain rather than their walking ability. Nonetheless, since all variables are connected there will be an indirect sex relationship carried through the model via its association to *Night Pain.*

The marginal independence of *Diagnosis* and *Operation* is new in the post-operative model. However the interpretation of these apparent independences now differs from that made with the pre-operative data. Since both variables were determined prior to gathering the post-operative patient status data, the directions of implication are reversed. Therefore, we can infer that the choice of treatment has no significant association with the patient's status at 1-year after the operation. That

Variable	Num. Edges	Variable	Num. Edges
Walking Aids	6	Weight Sitting Down Fixed Contracture OK. Flexion	2
Going Up Stairs Night Pain	5		
Sex Walking Ability Rising Up Standing Going Down Stairs Flexion Hip Abduction	4	Age CTF. Angle Extension Lag OK. Fixed Cont. Other Hip Abduction	1
Pain Frequency Pain Severity Stability	3	Diagnosis Operation Side	0

Table 5.2: Table of the number of edges connected to each in the graphical model for the post-operative knees data.

is to say that the two operations types *Cemented* and *Uncemented* perform in a similar manner with neither treatment being apparent superior to the other on the basis of the given patient status information. This directly agrees with the profile plot in Figure 4.17. Additionally, we can say that all patients respond in a similar manner irrespective of their underlying pathology with the rheumatoid arthritis and the osteoarthritis groups showing no significant differences. We can also infer that both groups respond to the two treatments in similar ways with neither of the possible treatments being most applicable to patients with one condition or the other. The marginal independence of *Side* remains in the post-operative data.

5.2.3.3 Joining the Pre-op and Post-op Models

In order to construct a model over both the pre-operative and the 1-year post-operative data requires including all variables from both time points into the model. Consequently the initial model would be ultimately far more complex than those studied in the previous sections, and the number of possible pairwise interactions

would be huge. Nevertheless, an attempt was made to construct such a model encompassing both time points. The initial model was, as before, that of independence and forward selection was begun. However, due to the sheer size of the model, the amount of data available and the number of parameters associated with this model each edge inclusion test was taking approximately one minute to evaluate. This resulted in a model selection process that was easily running for several hours with little progress.

This unacceptably slow progress of fitting a prospective model to the data is not reasonable for application to a clinical data setting. It is not feasible to require many hours or even days to fit a single two time-point model. The inclusion of the other two time points would obviously render this process far less suitable. Therefore, this graphical modelling framework as it stands is not directly suited to the analysis of data of this complexity. A possible solution to this problem could be to reduce the size of the data set and operate with a reduced number of variables, thereby reducing the model size and complexity. A further advantage could be gained by exploiting the ordering of the data given by its temporal structure.

5.2.4 Results - Hips Data

5.2.4.1 Pre-operative Data

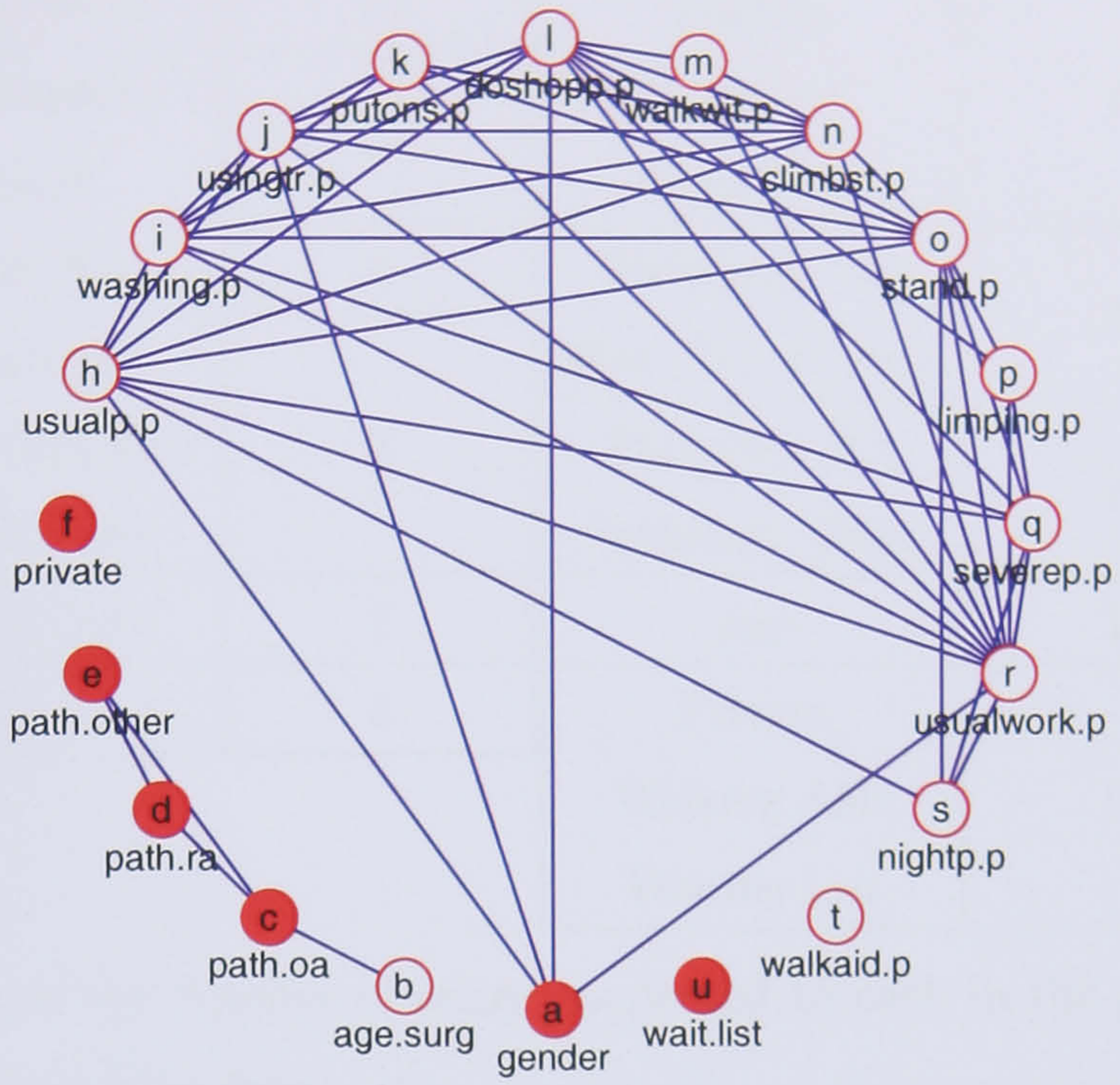
Applying the same model selection process to the pre-operative hips data as performed previously yields the model graph in Figure 5.8. First examination of the model graphs shows there to be a great many more edges present in the model than we found for the knees data. Indeed, there are a great many pairwise dependencies found within the 12 patient status variables as indicated by the edge counts in Table 5.3. This is likely due to the fact that these variables were all fairly well correlated implying a strong degree of association amongst such measurements. Such a high level of dependency within these patient status variables indicates that there will likely be a great deal of overlap, duplication and redundancy in the information conveyed by these measurements. When these measurements are combined into the composite Oxford Hip Score, these duplications and redundancies will compound

themselves.

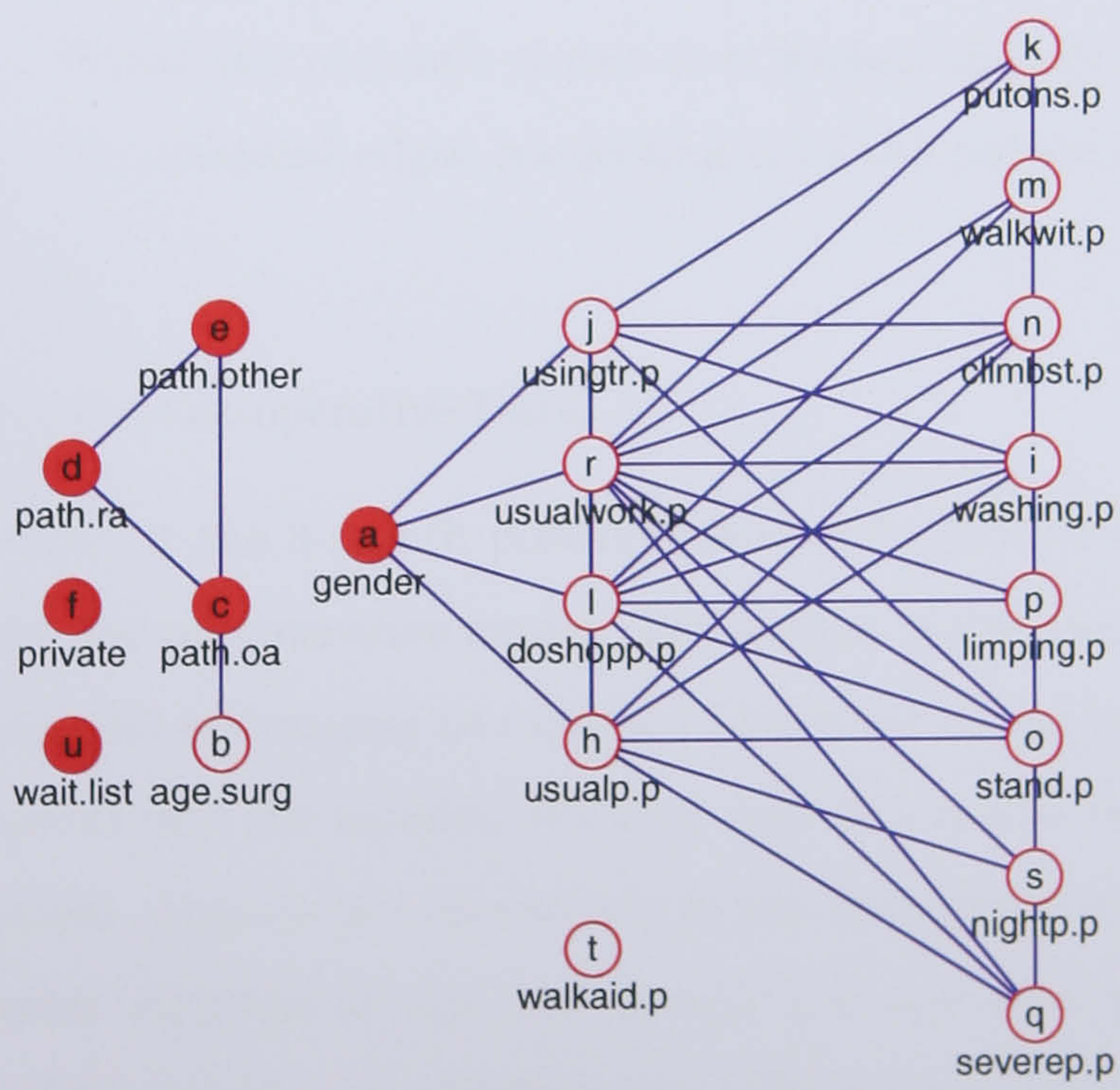
The one status variable which is not part of the composite score is *Walking Aids* (***t***). Under the graphical model, this variable is marginally independent of and has negligible correlations with the patient status indicators. This quantity is measuring information that is quite different from that of the hip score measures and as such is a novel descriptor of the patient's status.

Turning to the demographic variables, the variable *Gender* (***a***) has a relationship with four of the patient status variables indicating their values are not the same for both sexes. The variables *Using Transport* (***j***), *Usual Work* (***r***), *Do Shopping* (***l***) and *Usual Pain* (***h***) are all connected to the rest of the patient status variables which will mean that *Gender* will have an indirect association throughout the patient's status. Additionally, the variable *Age* (***b***) has no relationship with the patient status measurements as it is marginally independent in the model. The three variables encoding the patient's underlying pathology also have no association with the patient status variables suggesting that there is no difference between patients in the different pathology groups.

The factor variable *Private* (***f***), which represents whether the patient is receiving private or NHS treatment, appears to have no association to the patient's pre-operative state. This is slightly unexpected as we saw in Chapter 4 that the pre-operative states for private patients was typically slightly better than for NHS patients. It transpires that we are, again, victim of the restrictions of working in a decomposable graphical modelling framework. If we examine the significance of including edges from *Private* to the patient status variables we find significant evidence to include arcs to any of the measurements bar *Walking Aids*. As with the knees data, the introduction of such an edge would destroy the model's graphical and decomposable properties. Leaving such significant relationships absent from the model would typically cause one to conclude that such relationships were not present in the data when they evidently are present but have been overlooked to preserve the mathematical properties of the model. This behaviour is a significant problem and a potential barrier to interpretation of the model. Another factor variable, *Waiting List* (***u***), which represents the length of time the patient has been on a



(a) Standard Layout



(b) Rearranged Layout

Figure 5.8: Final model for the pre-operative hips data.

Variable	Num. Edges	Variable	Num. Edges
Usual Work	12	Gender Limping Night Pain	4
Do Shopping Stand	10		
Usual Pain	9	Pathology OA Walk W/out Pain	3
Washing Using Transport Climb Stairs	8	Pathology RA Pathology Other	2
Severe Pain	7	Age	1
Put On Socks	5	Private Walking Aids Waiting List	0

Table 5.3: Table of the number of edges connected to each in the graphical model for the pre-operative hips data.

waiting list, also appears again to have no relationship with the patient's status. Examination of the data suggests that in this case the absence of relationship appears to be reasonable. Whilst the patient's status does worsen slightly as *Waiting List* increases, none of the potential edges connecting it to the patient status variables are significant.

5.2.4.2 Three-month Post-operative Data

The graphical model for the 3-month post-operative hips data exhibits some common structure with the pre-operative model. The patient status variables all remain strongly associated post-operatively as expected from their strong correlations. Furthermore, post-operatively the variable *Walking Aids* ($\boldsymbol{T}$) is now dependent on the other status variables, whereas pre-operatively it was apparently independent.

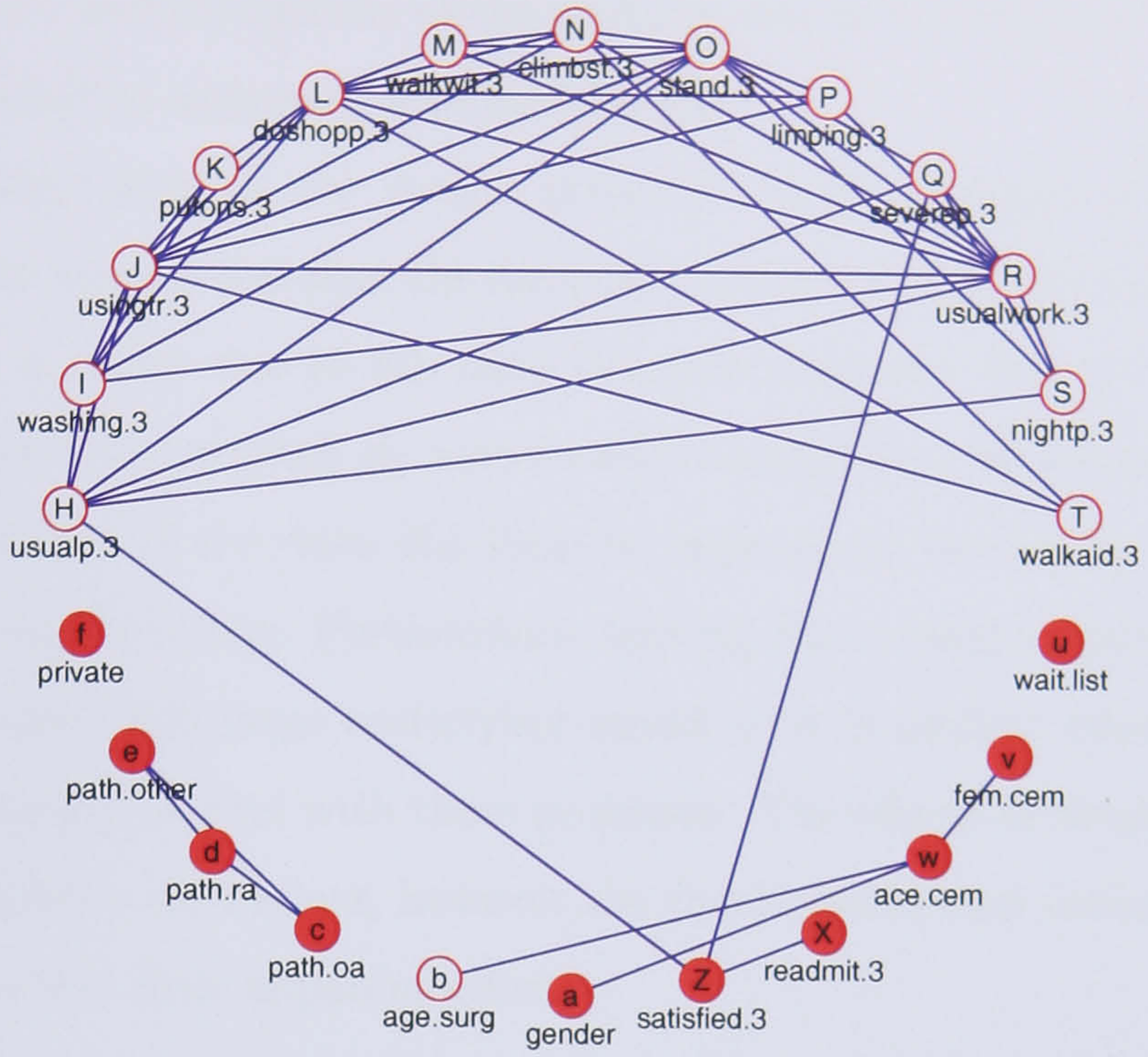
The demographic variables of *Age* and *Gender* are now both marginally independent of the patient's status, suggesting no sex differences or associations between age and the post-operative condition of the patient. Neither do we have any associations with the variables *Private* and *Waiting List* and other variables. Furthermore, both the pathology variables and the two treatment variables ($\boldsymbol{v}$ and $\boldsymbol{w}$) are also

independent of the patient's condition suggesting no significant differences between these groups either.

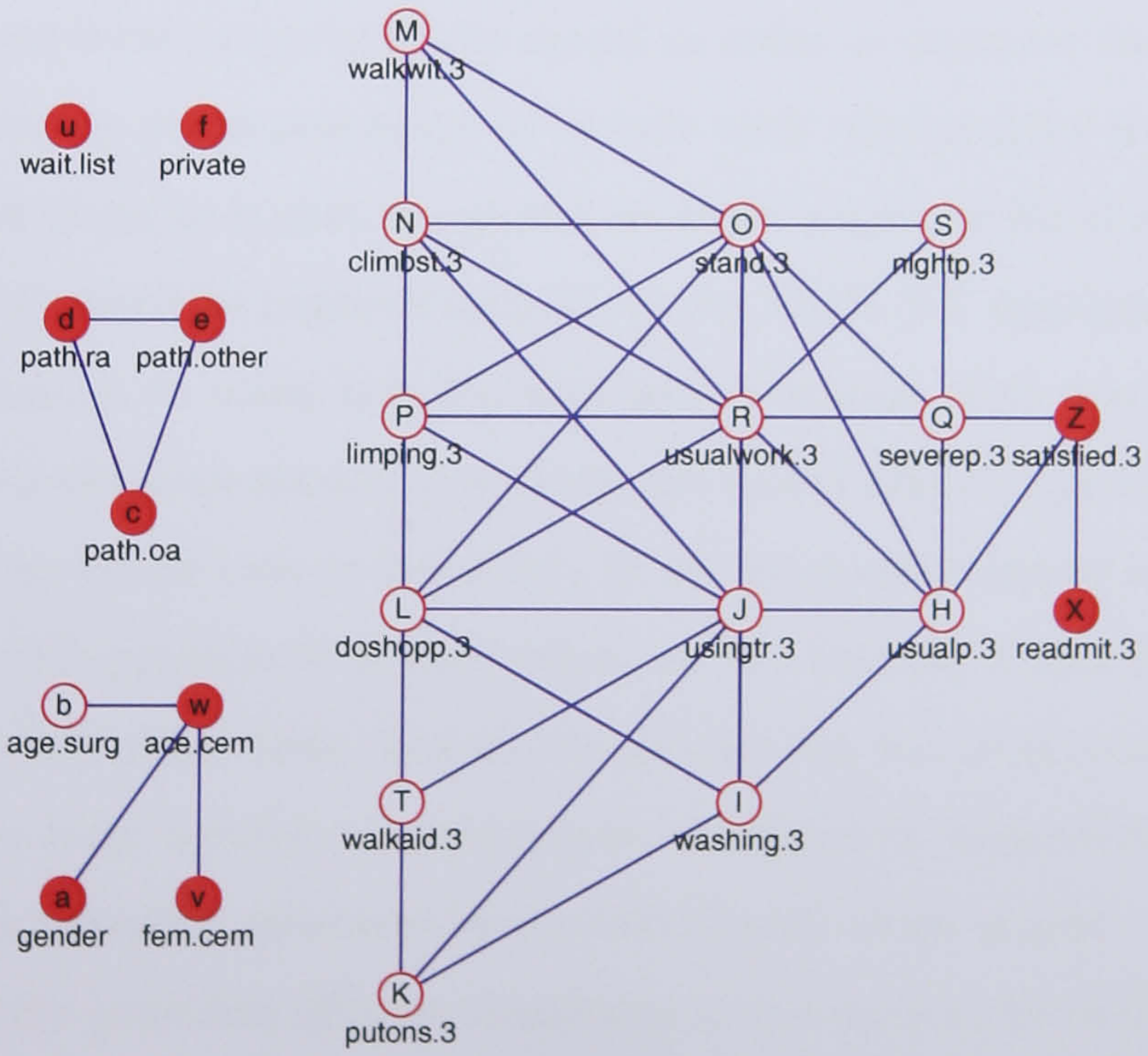
The introduction of the variables *Satisfaction* and *Readmitted to Hospital* (nodes $\boldsymbol{Z}$ and $\boldsymbol{X}$) reveal some interesting information about the patient's satisfaction. The model graph shows that the patient's satisfaction is directly associated to only three other variables in the model. These are *Severe Pain*, *Usual Pain* and *Readmitted to Hospital*. This implies that pain levels and whether they needed to return to hospital can explain the patient's level of satisfaction. Whilst this may be a reasonable relationship for *Satisfaction*, it seems unusual that *Readmitted to Hospital* would be conditionally independent of the patient's status variables given *Satisfaction*. Indeed one would expect direct associations between the status variables and *Readmitted to Hospital* with patient's who were not readmitted faring better than those who were. Further investigation revealed potentially significant associations between *Readmitted to Hospital* and variables such as *Using Transport* and ($\boldsymbol{J}$ and $\boldsymbol{N}$) which were not added due to model constraints as discussed previously. However, there does still remain a path between *Readmitted to Hospital* and the status variables which indicates an indirect association between those variables.

5.3 Limitations

Throughout the course of this chapter, it has become clear that there are a number of limitations to the graphical modelling approach. The most obvious is that it becomes progressively harder to fit models when the number of variables increases. This was particularly true when both the pre-operative and post-operative data were included in the same model. The sheer size of the model became unwieldy resulting in a long time to perform the edge tests resulting in a slow model selection process. This issue of dimensionality, model complexity and the required computation involved is highly important as it was evident that MIM was struggling to fit the model over only two time points. Including the data from the later time points would result in a further increase in size and complexity of the model that would likely be impossible to fit without the aid of specialised high-performance computers. This would mean that



(a) Standard Layout



(b) Rearranged Layout

Figure 5.9: Final model for the 3-month post-operative hips data.

the procedure would be wholly impractical in a clinical setting. Some mechanism by which the size and complexity of the problem can be reduced is then required if a graphical modelling approach is to be followed.

A further limitation to the results presented in this chapter is that we typically violate the assumption that the data have a multivariate Normal distribution. This violation is either due to the data being intrinsically skewed or being originally ordinal and then treated as continuous for simplicity. However, we saw that the transformations of the data did little to improve its Normality and sacrificed the model's interpretability. Furthermore, leaving the ordinal variables as discrete results in an impossibly large underlying model so it is unclear whether there are any ways to adequately deal with these problems. The effects of these violations on the resulting models are unclear, however the checks performed earlier showed that these violations had little apparent effect.

Another problem encountered during the selection of the graphical models was that certain significant edges were not included in the model as this would result in the model no longer being decomposable, or graphical or both. Whilst it is necessary to retain the graphical property of the model in order to interpret the model graph correctly, sacrificing decomposability to include such relationships is possible. The disadvantage of doing so is that the model could no longer be fitted exactly and an iterative method would be required such as via the MIPS [53] algorithm. Whilst not necessarily a barrier to using non-decomposable methods, with many variables in the model this could dramatically slow down the model selection process as for each potential edge we would have to iteratively fit the prospective model in order to test it. Of course, this problem is averted when we use backward selection. However, this is not possible with these data as the models are too large and required the estimation of a huge number of parameters, rendering it impossible to fit. This problem can be partially addressed by modelling with *chain graphs*.

To address the problems of over-complexity in the model, the next two chapters discuss a mechanism to reduce the number of variables. We could then build smaller models over this reduced set of variables which would eliminate some of the problems we had observed with the dimensionality of the data. In Chapter 8, we return to the

graphical modelling of these data using an extension of graphical models known as *chain graphs* which allow for an efficient representation of the temporal structure of the data. They also suffer less from the problems of significant edges being ignored to preserve decomposability.

Chapter 6

Variable Reduction and Principal Variables

6.1 Introduction

In multivariate analysis with many variables it is often desirable to be able to highlight which of those variables are the most important and which variables could be deemed to be redundant or uninformative. By determining such a subset of *principal* variables, it may be the case that the remaining variables can simply be discarded and attention can subsequently be focused on the reduced subset. In practical terms taking this step not only serves to reduce the dimensionality of the problem, but by working with a smaller group of variables we reduce the problem's complexity and thereby save time, reduce computational overhead and facilitate interpretation of the ultimate results of the statistical investigation. Such approaches can help to combat some of the problems encountered in Section 5.3. In addition, identifying the most important subset of variables can have significant implications for future data collection endeavours with fewer measurements needing to be recorded.

This chapter proposes and develops a method for determining such a reduced subset of the original variables. The chapter begins in Section 6.2 with a review of several methods found in the literature. Motivated by this, in Section 6.3 a suitable measure of variability of a variable based on the principal component analysis of the correlation matrix of the data is proposed. In Section 6.4 a simple stepwise procedure

for variable selection is developed and refined using the defined statistic. Section 6.5 introduces several extensions to the selection procedure - the first is the incorporation of temporal information and the second addresses using utility information in the selection process. The final extensions are proposed improvements to the search procedure and the presentation of some graphical aids similar to the commonly-used scree plots. Section 6.6 presents a discussion of ways to ascertain the effective dimensionality of the data in terms of the number of variables which correspond to a stopping rule for the selection process. The chapter ends with a presentation of the results of both a Monte Carlo simulation study and the analysis of real data to assess the performance of the selection procedure and to compare it with existing methods.

6.2 Existing Data and Variable Reduction Techniques

6.2.1 Preliminaries

In what follows, we suppose that we collect n observations on a $p-$dimensional measurement vector into the $n \times p$ data matrix $\boldsymbol{X}$. Suppose that the sample covariance and correlation matrices are respectively $\boldsymbol{\Sigma}$ and $\boldsymbol{R}$. Our aim is to select some subset of m principal variables (PVs) where $m < p$ and which best (in some sense) represent the original variables. Suppose that we partition the variables $\boldsymbol{X}$ into the subsets $\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}$ we can assume without loss of generality that $\boldsymbol{X}_{(1)}$ represent our subset of PVs and $\boldsymbol{X}_{(2)}$ are the remaining variables. It will be helpful to consider partitioning $\boldsymbol{\Sigma}$ correspondingly as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then, the partial covariance matrix for $\boldsymbol{X}_{(2)}$ given $\boldsymbol{X}_{(1)}$ is

$$\boldsymbol{\Sigma}_{22 \cdot 1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{21}, \tag{6.1}$$

and the partial correlation matrix, $\boldsymbol{R}_{22 \cdot 1}$, is obtained by scaling $\boldsymbol{\Sigma}_{22 \cdot 1}$ so that diagonal elements are unity. If we begin with a correlation matrix $\boldsymbol{R}$, then we further define

the *unscaled* partial correlation matrix:

$$\widetilde{\boldsymbol{S}}_{22\cdot 1} = \boldsymbol{R}_{22} - \boldsymbol{R}_{12}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{21}. \tag{6.2}$$

Another type of correlation that arises in variable reduction is the *canonical correlation*. A canonical correlation analysis seeks the vectors a and b such that the random variables $a^T\boldsymbol{X}_{(1)}$ and $b^T\boldsymbol{X}_{(2)}$ maximise the correlation:

$$\rho = \mathrm{Cor}[a^T\boldsymbol{X}_{(1)}, b^T\boldsymbol{X}_{(2)}]. \tag{6.3}$$

The random vectors $U = a^T\boldsymbol{X}_{(1)}$ and $b^T\boldsymbol{X}_{(2)}$ which attain this maximum are the first pair of canonical variables and the first canonical correlation is given by ρ for these variables. The process then continues seeking pairs of vectors which maximise the same correlation subject to the constraint that they are to be uncorrelated with all preceding pairs of canonical variables; this gives the remaining pairs of canonical variables and the rest of the canonical correlations. A canonical correlation analysis consists of finding eigenvalues and eigenvectors of the following matrix:

$$\widehat{\boldsymbol{R}} = \boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}.$$

The canonical correlations are given by the square roots of the eigenvalues of $\hat{\boldsymbol{R}}$ and the canonical variates use the eigenvectors in determining their composition.

Finally, we define the multiple correlation coefficient as the correlation between a response variable y and its associated covariates $x_1, \ldots, x_n$. The multiple correlation coefficient, R, is given by:

$$R = \sqrt{c^T\boldsymbol{R}_X^{-1}c}, \tag{6.4}$$

where $\boldsymbol{R}_X$ is the matrix of correlations of the covariates and c is the vector with elements $c_i = \mathrm{Cor}[y, x_i]$. The square of the multiple correlation coefficient is known as the *coefficient of (multiple) determination* and indicates the amount of variation in y that is explained by the predictors $x_1, \ldots, x_n$.

6.2.2 Variable Reduction Methods

6.2.2.1 Jolliffe's PCA-based Methods

One of the most commonly used tools within multivariate data analysis for reducing dimensionality is Principal Component Analysis (PCA) which has been well

documented in the literature [71, 68]. Principal Component Analysis examines the variance or correlation matrix generated from data on p random quantities and seeks to explain this p-dimensional random variation by decomposing it into p separate and orthogonal 1-dimensional components. These components correspond to the eigenvectors of the variance matrix and their corresponding eigenvalues represent the variance of the associated component. It is therefore not surprising that there exist several variable selection methods based on this methodology.

Jolliffe has discussed various techniques for variable selection based on the principal components methodology [66, 68]. These techniques fall into three groups - the first using multiple correlation coefficients, the second using the principal components (PCs) themselves and the third using cluster analysis. Of the three groups, the techniques which used the principal components were the most successful. These techniques are referred to (using Jolliffe's notation) as methods **B1**, **B2** and **B4**. Method **B3** was determined to be unsatisfactory in its performance and so will not be discussed here.

Each of Jolliffe's **B** methods associates a single variable with each of the principal components of $\boldsymbol{\Sigma}$ (or $\boldsymbol{R}$). The PCs are ordered according to the size of their associated variances with the component with the largest variance being first. The variable chosen is then that variable which has the largest absolute loading in the principal component under consideration. Methods **B1** and **B2** then begin this association process with the last principal component, under the reasoning that the last few components typically represent near-constant relationships and are usually dominated by a single variable, and thus these are ideal candidates for exclusion. Hence the variable associated with this last component is discarded and attention then proceeds to the penultimate component dismissing its associated variable, and continues until sufficient variables had been eliminated. Method **B1** iteratively performs the PCA on each remaining subset of variables, whereas method **B2** operates only on the initial PCA. By contrast, method **B4** is a forward selection approach and associates and retains variables with high loadings in the first m principal components.

These methods were demonstrated to be both fast and efficient via a simulation

study. However, it is important to note that Cadima & Jolliffe [14] state that the underlying selections of such methods can be 'seriously unreliable' as both the loading of a variable in a PC and the associated variance of that PC are required in order to determine that variable's importance. Neglecting the information conveyed by one or the other would likely produce inappropriate results.

6.2.2.2 McCabe's Principal Variables

McCabe [87] considered the various principal components optimality criteria and solved them directly. Consequently, he showed that the variable selection problem motivated by a PCA approach has a non-unique solution. In fact, there were four solutions to the 'best' variable subset problem where the set of *principal variables* is that which satisfies one of these criteria:

$$\mathbf{M1} \quad \max |\mathbf{\Sigma}_{11}| \equiv \min |\mathbf{\Sigma}_{22\cdot 1}| \equiv \min \prod_i \lambda_i \tag{6.5}$$

$$\mathbf{M2} \quad \min \operatorname{tr}(\mathbf{\Sigma}_{22\cdot 1}) \equiv \min \sum_i \lambda_i \tag{6.6}$$

$$\mathbf{M3} \quad \min ||\mathbf{\Sigma}_{22\cdot 1}||^2 \equiv \min \sum_i \lambda_i^2 \tag{6.7}$$

$$\mathbf{M4} \quad \max \textstyle\sum_{i=1}^{p} \rho_i^2 \tag{6.8}$$

Here $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22\cdot 1}$ are defined as (6.1); $|\boldsymbol{A}|$ and $\operatorname{tr}(\boldsymbol{A})$ are the determinant and trace of the matrix $\boldsymbol{A}$, respectively; $||\boldsymbol{A}||^2$ is the squared norm $(\sum\sum a_{ij}^2)$; λ_i are the eigenvalues of $\mathbf{\Sigma}_{22\cdot 1}$; and the ρ_i are the canonical correlations between the variables not selected and those selected as defined in (6.3). As McCabe points out, since $\mathbf{\Sigma}_{22\cdot 1}$ represents the information left in the remaining variables, once the chosen ones have been removed it is quite plausible that three of the optimality criteria should be functions of this matrix.

Of all these four possible solutions to the variable reduction problem, McCabe noted that only solution **M2** can be arrived at in a stepwise fashion. The stepwise method for finding a near-optimal set of variables that satisfy **M2** can be determined by exploiting the following equation attributed to Okamoto [93]:

$$\min \operatorname{tr}(\mathbf{\Sigma}_{22\cdot 1}) = \min \sum_{i=1}^{m} \lambda_i = \max \sum_{i=1}^{m} \sigma_{ii} \eta^2(v_i; v^{(1)}, \ldots, v^{(j)})$$

where λ_i are the eigenvalues of $\mathbf{\Sigma}_{22\cdot1}$, σ_{ii} is the standard deviation of variable i. $\eta^2(x;\boldsymbol{y})$ is the squared multiple correlation between x and $\boldsymbol{y}$ (see (6.4)). v_i is the i-th *remaining* variable and $v^{(k)}$ is the k-th *selected* variable. The stepwise procedure then involves, at step j, selecting the variable that maximises the sum on the right. This solution **M2** is equivalent to the RM criterion discussed in [13] (see [15] for details). The RM criterion is defined to be the cosine of the angle between $n \times p$ data matrix $\boldsymbol{X}$, and the $n \times m$ matrix whose columns result from regressing each of the p (centred) observed variables on M (i.e. orthogonally projecting them on M), where M is the subspace of $\mathbb{R}^n$ spanned by the m variables in our chosen subset.

The remaining three solutions would require an exhaustive evaluation of all possible variable subsets in order to determine which subset was the optimum. Whilst possible, but computationally expensive, for smaller subsets of variables this would rapidly become computationally infeasible as the number of variables increases.

6.2.2.3 Multiple correlation method

Beale *et al* [7] discuss a method for discarding variables based upon multiple correlation. They suggested that one should retain the subset of m variables which maximise the minimum multiple correlation between the m selected variables and any of the remaining variables. Let us call this method **A1**. However, this method was determined by Beale *et al* [7] and Jolliffe[68] to be too slow to be practically useful at the time of publication. This was due to the fact it required exhaustive enumeration of all subsets of size m, in addition to the calculation of the corresponding multiple correlations for each subset. As a potential alternative to this method, Jolliffe proposed a stepwise version whereby at each stage the variable with the highest multiple correlation with the remaining variables was excluded until only m variables remain (method **A2**). It does not appear that either of these methods has been investigated since Jolliffe in the 1970s.

6.2.2.4 Krzanowski's Procrustes method

Adopting an entirely different approach, Krzanowski [75] proposed a method based on Procrustes Analysis (call this method **KP**). To compare the various subsets, the

sum of squared differences between data points after being transformed to principal component space based on the PCA of all the variables are compared with the sum of squared differences when transformed to the PCA-space based on a *reduced* subset of variables. Krzanowski's method intends explicitly to preserve the multivariate structure of the original data as much as possible in the final variable subset, rather than selecting a set which seeks to maximise some variance measure over the variables.

6.2.2.5 A method based on graphical Gaussian models

A different variable selection method has been proposed by de Falguerolles *et al* [28] (**DF**). The proposed method is based on graphical Gaussian models and seeks to choose a subset of variables which are the focus of the model graph of the graphical model. That is to say the selected variables should have many connections to other variables and leave the unselected variables conditionally independent given those selected. To do this they seek the variable subset that minimises the deviance of this hypothesised model from the saturated model via the expression:

$$D^2 = -N \log \left(\frac{|\widetilde{\boldsymbol{S}}_{22\cdot 1}|}{|\mathrm{diag}(\widetilde{\boldsymbol{S}}_{22\cdot 1})|} \right),$$

where N is the sample size and $\widetilde{\boldsymbol{S}}_{22\cdot 1}$ is as (6.2). Whilst likely being predictively useful for the variables, it could include potentially redundant variables. For example, suppose we have one such key variable which is associated to all other variables and that given this measurement the remainder are all conditionally independent. We could introduce a second key variable equal to the first measurement plus some random noise. This variable would exhibit similar relationships as the first and may be selected along with the first as a focal point in the graphical model, thereby introducing unnecessary duplication of the same information. Furthermore, if we want to select $p - 1$ variables, then for all p possible variable subsets we have $\widetilde{\boldsymbol{S}}_{22\cdot 1} = \mathrm{diag}(\widetilde{\boldsymbol{S}}_{22\cdot 1})$ resulting in all subsets giving a D^2 value of 0 preventing us from making a sensible choice of variable using this method. To compensate in this case, de Falguerolles *et al* recommend eliminating that variable with the maximum diagonal element in $\boldsymbol{R}^{-1}$.

6.2.2.6 Simple component analysis

Another technique based on PCA is the method of simple component analysis (SCA) proposed by Rousson and Gasser [106]. Whilst technically a technique to reduce dimensionality rather than variables, it has some interesting features which are worthy of mention. The goal of SCA is replace the standard PCs with a set of components that, whilst being suboptimal, are more easily interpreted. The results of SCA are most appropriate to the situation where the initial correlation matrix is either approximately or exactly block diagonal. First, SCA seeks to obtain a single component to represent each block of variables within the data - these are termed the *block components.* Secondly, SCA then obtains a set of *difference components* which represent information about the structure within a single block.

The methods of SCA reduce the dimensionality of the data set in manner that is far more intuitive than PCA. By first obtaining the block components one gains insight into the overall block structure of the data. The difference components then give further details of the relationships between the variables in a single block. Interpreting these simple components, one could suppose each of these blocks represents a single latent quantity. The selection of a single variable from each block could thus be a possible variable reduction strategy.

6.2.2.7 CUR decompositions

In many applications, huge data sets are being constructed in areas such as credit scoring, complex manufacturing, and image analysis for astronomical data. For these cases, the recent developments concerning CUR decompositions may form a possible means to dimension reduction [34]. The CUR decomposition is a method by which one seeks to replace the $(n \times p)$ data matrix $\boldsymbol{X}$ by an approximation $\boldsymbol{A}$. The need to use such an approximation is peculiar to the situation with prohibitively large data sets whose size is such that it becomes impossible to load or manipulate these data within the confines of a computer's available RAM.

The approximated data matrix is defined as $\boldsymbol{A} = \boldsymbol{CUR}$ where $\boldsymbol{A}$ is the approximate decomposition of $\boldsymbol{X}$, and $\boldsymbol{C}, \boldsymbol{U}, \boldsymbol{R}$ are smaller, more easily computed matrices. $\boldsymbol{C}$ is an $(n \times c)$ matrix formed from c randomly chosen columns of $\boldsymbol{X}$, and similarly

$\boldsymbol{R}$ is an $(r \times p)$ matrix of r random rows of $\boldsymbol{X}$. The $(c \times r)$ matrix $\boldsymbol{U}$ is then determined from $\boldsymbol{C}$ and $\boldsymbol{R}$. This method is most appropriate for enormous data sets; application to more data of more manageable dimensions would likely result in approximations that were likely more crude than obtainable by other methods. Additionally, the selection of the variables which constitute $\boldsymbol{C}$ is random and not informed. This is due to the scale of the data preventing such an investigation. However, despite being unsuitable for data sets of modest size these decompositions may have significant potential for enormous and otherwise unmanageable data sets.

6.2.2.8 Regression-based methods

Stepwise selection procedures are well established within the framework of regression whereby terms are added to or removed from the original regression model. The main distinction between such methods and those discussed above is that regression-based methods have a defined response and a goal to provide the best possible prediction of the response using the remaining variables. This structure leads to several natural criterion functions which are used to motivate the selection process. These include Mallow's C_p, the prediction error sum of squares (PRESS), the multiple correlation coefficient and F-ratios. Hocking [61] provides a comprehensive review of variable selection in regression.

6.2.3 Assessing Dimensionality

A further problem entwined with the topic of variable selection is the determination of the number of the original variables to be retained. In the domain of PCA there exist many techniques for assessing the number of principal components to retain.

Jolliffe [68] proposed some techniques for determining the number of PCs to keep based upon the eigenvalues associated with the principal components. For example, one could seek to retain sufficient components so that the cumulative proportion of the original variability in the data was above a particular level, say 75%. A second approach, when working with a correlation matrix, is known as Kaiser's rule [69] and is to drop all PCs whose associated eigenvalues fall below a threshold of 1. The rationale for this is that 1 is the average eigenvalue and a PC with variance

less than one conveys less information than a single original variable. However, Peres *et al* [97] point out that due to sampling variation, one-half of the sample eigenvalues from randomly generated data will exceed this threshold. Consequently, Jolliffe [66] determined that a reduction of the threshold from 1 to 0.7 would be more appropriate and would serve to allow for this variation.

A common graphical method for assessing dimensionality using a PCA of the data is the scree plot [16] where the variances of each PC are plotted in descending order of magnitude. This is a method also used for determining the effective dimensionality where one seeks the PC beyond which the variances decrease in a linear fashion - the PCs beyond this point are regarded as representing noise within the data.

Wold [127] and Eastment and Krzanowski [36] propose a cross-validatory approach for determining the number of components to retain. Eastment and Krzanowski's approach is based on successively predicting each element of the data matrix after its row and column have been deleted, whereas Wold's method considers using larger subgroups within the data for the cross-validation. Both methods then consider the prediction error sum of squares (PRESS) for different numbers of components and construct statistics based on functions of the various values of PRESS.

Velicer [122] adopted an approach based on partial correlation in order to determine the number of components. He proposed considering the partial correlations between p variables given the first m principal components in order to determine the number to retain. He proposed considering the statistic:

$$V = \sum_{i=1;i\neq j}^{p} \sum_{j=1}^{p} \frac{(r_{ij}^{\star})^2}{p(p-1)}$$

where $r_{ij}^{\star}$ is the partial correlation between the ith and jth variables given the first m PCs. Velicer noted that V decreases and then subsequently increases as m, the number of retained components, increases. The optimum value for m was then suggested to be at the minimum value for V. However, Jolliffe [68] noted that whilst this was a reasonable approach for factor analysis, it was inappropriate for PCA as it unfairly dismisses PCs that are dominated by a single independent variable as

they have low partial correlations but provide information unavailable from other sources.

A further method operates under the assumption that the total variance in the data is randomly divided amongst the principal components. Under this assumptions, the expected distribution for the eigenvalues can then be assumed to follow a broken-stick distribution [51]. Under this model one can then calculate the expected proportion of the total variation associated with the individual eigenvalues. This expected proportion, b_k, is given by:

$$b_k = \frac{1}{p} \sum_{i=k}^{p} \frac{1}{i}$$

where p is the number of variables. If proportion of the total variation associated with the eigenvalue for the kth component exceeds the corresponding value of b_k then the component is retained.

Peres *et al* [97] provide a comprehensive overview of many stopping rules for determining the number of non-trivial principal components. The results of the simulation study testing the effectiveness of many such stopping rules showed that the original Kaiser's rule performed poorly whereas Velicer's method was among the most accurate though it did tend to underestimate the number of dimensions. Stepwise methods for regression models typically have associated stopping rules which govern the complexity of the associated model, though such methods are typically inappropriate in the absence of a defined response variable.

6.3 A Measure of Variability

As McCabe [87] stated that three of the optimal subsets of variables can only be determined exhaustively and, furthermore, since the orthopædic data sets under examination have large numbers of variables, it would be sensible to pursue a stepwise rather than exhaustive approach to variable selection even though the solutions thus provided will likely be non-optimal. The exhaustive evaluation of all $2^m - 1$ possible subsets for all $m = 1, \ldots, p$ can be computationally infeasible, especially with the large orthopædic data sets discussed in Section 2.2. Therefore it may be the case

that a 'good' solution to the subset selection problem achieved with reasonable effort is to be favoured to an optimal solution obtained at great cost.

One fundamental component of stepwise procedures is a criterion function which yields a numerical statement of the suitability or unsuitability of the various elements being considered at each stage. In the context of variable selection, this would be a numerical expression of the desirability to retain or discard a particular variable. The need for such a statement for the value of a single variable is generally peculiar only to stepwise procedures, as fully exhaustive methods would consider the value of every possible subgroup of variables in order to determine the optimum. Whilst guaranteeing optimality these exhaustive methods suffer a corresponding increase in computation and time in order to examine all possible variable subset combinations and as the number of variables increases we would suffer a combinatorial explosion in computing required to determine this subset.

Our focus shall be directed towards the methods of both Jolliffe and McCabe who obtained suitable variable subsets by examining the variance or correlation matrix of the data using a principal components approach. Whilst these methods are expressed in terms of the variance matrix, $\boldsymbol{\Sigma}$, it may be preferable to consider the correlation matrix. The reason for this is that variables on larger scales will have correspondingly higher variances which will, in turn, dominate the principal components of the variance matrix. By scaling the variance to correlation form we give equal emphasis to all variables.

Consider the $(p \times p)$ variance matrix $\boldsymbol{\Sigma}$, for which we shall assume full rank for convenience. It can be expressed in terms of its spectral decomposition as:

$$\boldsymbol{\Sigma} = \sum_i \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^T$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ are the ordered eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p$ the associated eigenvectors. $\boldsymbol{A}$ is then the $(p \times p)$ orthonormal matrix whose columns are the $\boldsymbol{a}_i$ and $\boldsymbol{\Lambda}$ is the $(p \times p)$ diagonal matrix with entries λ_i.

Of the four solutions to the principal variable selection problem proposed by McCabe, it is to solution **M3** in (6.7) that we turn. Method **M3** states that we would seek to retain those variables which will minimise $||\boldsymbol{\Sigma}_{22 \cdot 1}||^2$ - the squared norm

of the matrix of partial covariances of the variables remaining given those selected. To arrive at an expression for a suitable variable criterion function we consider the composition of the squared norm of a general variance matrix $\boldsymbol{\Sigma}$:

$$\begin{aligned} ||\boldsymbol{\Sigma}||^2 &= ||\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^T||^2 \\ &= ||\boldsymbol{A}\boldsymbol{\Lambda}||^2 \\ &= \sum_i \lambda_i^2 \sum_j a_{ji}^2 = \sum_i \lambda_i^2 \qquad (6.9) \\ \text{or} \quad &= \sum_j \sum_i (\lambda_i a_{ji})^2 = \sum_j h_j \qquad (6.10) \\ \text{or} \quad &= \text{tr}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}) = \sum_j \sum_i \sigma_{ij}^2 \qquad (6.11) \end{aligned}$$

Decomposition (6.9) expresses $||\boldsymbol{\Sigma}||^2$ as a sum of the squared eigenvalues each representing the variability associated with each principal component or column in the $\boldsymbol{A}$ matrix. This leads to an obvious route of variable reduction as the basis for the principal components methods of Jolliffe. This approach uses λ_i and $\boldsymbol{a}_i$ to select variables by concentrating first on eigenvectors with large eigenvalues, and then looking to the variables with high factor loadings on them. We can see from (6.9) that the first eigenvalue provides the greatest contribution to $||\boldsymbol{\Sigma}||^2$ with the remaining eigenvalues providing progressively less input. This of course corresponds to the fact that the first principal component has the largest variance of any linear combination of the original variables.

The second decomposition of $||\boldsymbol{\Sigma}||^2$ in (6.10) suggests that we may instead examine values $h_1, \ldots, h_p$ which represent the variability of each row or variable in $\boldsymbol{A}$.

Definition 6.3.1 Given a variance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ over the variables $v_1, \ldots, v_p$, define the h statistic for the jth variable v_j as:

$$h_j = \sum_i (\lambda_i a_{ji})^2 = \sum_i \sigma_{ij}^2 \qquad (6.12)$$

Proof: Let $\boldsymbol{e_i}$ be the vector of p zeros, with 1 in position i. Now let $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^T$ be the spectral decomposition of the variance matrix. We can now write $\sigma_{ij} = \boldsymbol{e_i}^T\boldsymbol{\Sigma}\boldsymbol{e_j}$.

So

$$\begin{aligned}\sum_{i=1}^{p}\sigma_{ij}^2 &= \sum_{i=1}^{p}\boldsymbol{e}_j^T\boldsymbol{\Sigma}\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{\Sigma}\boldsymbol{e}_j \\ &= \boldsymbol{e}_j^T\boldsymbol{\Sigma}(\sum_{i=1}^{p}\boldsymbol{e}_i\boldsymbol{e}_i^T)\boldsymbol{\Sigma}\boldsymbol{e}_j \\ &= \boldsymbol{e}_j^T\boldsymbol{\Sigma}\boldsymbol{\Sigma}\boldsymbol{e}_j \\ &= \boldsymbol{e}_j^T\boldsymbol{A}\boldsymbol{\Lambda}^2\boldsymbol{A}^T\boldsymbol{e}_j \\ &= \sum_{k=1}^{p}(\lambda_k a_{jk})^2\end{aligned}$$

□

The h statistics are essentially the mean squared covariance between variable j and other variables. This decomposition leads to a possible route for variable selection, which again concentrates on high eigenvalues and high loadings, but instead averages these across components to yield a value appropriate for individual variables rather than linear combinations. Under this framework, we observe that the variable $v_{(1)}$ with the largest value of h provides the greatest contribution of all the variables to the squared norm of the variance matrix. Hence, in this context, this would suggest that this variable is providing the most variability of all variables. Thus we conclude that this is the most desirable variable (in terms of $||\boldsymbol{\Sigma}||^2$) to retain in a selection process.

Consequently, this value h_i is a useful numerical statement on the variability of the i-th variable. It also, by definition, combines information from both eigenvalues and loadings which would not lead to the problems of unreliability as mentioned in [14]. However, it should be remembered that by using the correlation matrix, the value for h_i provides a measurement of the level of inter-correlations between variable i and other variables. Hence, a variable which is correlated with many other variables will score higher than a single independent variable. This means that on the basis of h_i values alone, independent variables would seem to be undesirable, whereas in fact they may be highly desirable as they will contain information that cannot be expressed in terms of other correlated variables.

If we restrict ourselves to correlation matrices, we can make the following statement about the properties of the h:

Proposition 6.3.1 For a $(p \times p)$ correlation matrix $\boldsymbol{R}$, the value of h_j for variable j lies in $[1, p]$.

Proof: The proof is simple and relies on the identity $h_j = \sum_i r_{ij}^2$. When all variables are independent $\boldsymbol{R} = \boldsymbol{I}$ and all $h_j = 1$. Conversely, if all variables were perfectly correlated then $\boldsymbol{R} = \mathbf{1}$ and all $h_j = p$. □

6.4 Stepwise Selection Procedures

6.4.1 The Simple Selection Procedure

A simple stepwise variable selection procedure can be arrived at immediately. Following the strategy employed in Jolliffe's **B2** or **B4**, we calculate the h scores associated with each individual variable and select that with the highest value (our first principal variable or PV). We could then continue selecting the variables with the highest h in the remaining set until we have enough variables (the issue of how many PVs is 'enough' will be addressed in Section 6.6.)

However, this approach is somewhat simplistic and has two key limitations. The first is that it is important to compensate for the fact that one or more variables have been selected and removed from the analysis. One such way of achieving this is to make use of the partial variance or correlation by transforming the initial variance or correlation matrix to the corresponding partial form after having identified the variable to select. By using the partial variances of the variables under consideration given those already removed, we eliminate the effects of the selected PVs from the subsequent analysis and ensure that we have compensated for their absence when determining the next PV for selection. To an extent, this seeks to mirror the notion of orthogonality of the principal components in the principal variables. However, there are complications when working with correlation instead of variances - this will be addressed in Section 6.4.2.

Using the decomposition of $\boldsymbol{\Sigma}$ into the h_i statistics, coupled with the use of partial variance, it is then straightforward to arrive at a slightly more sophisticated stepwise procedure for variable selection. Beginning with the original variance/correlation

matrix, we determine the h_i values for each variable v_i for $i = 1, \ldots, m$. Then we identify that variable with the largest h_i value and remove it from the set of remaining variables. This is our first selected variable and is the most important and informative in terms of h_i. Then we update the variance matrix to the partial variance matrix of the remaining variables given the variable we have just removed. The process then repeats: calculating h values, identifying candidate variables, removing them and updating the correlation matrix to reflect the fact that variables have been removed from the analysis.

This **H** procedure thus corresponds, in part, to a naive pursuit of the set of variables which satisfy McCabe's solution **M3** - $\min ||\boldsymbol{R}_{22\cdot1}||^2$. This is since the selection of the variable with maximum h_i will provide the greatest contribution, in terms of the h statistics and single variables, to $||\boldsymbol{R}_{11}||^2$. The selection of this variable should then provide the greatest reduction in $||\boldsymbol{R}_{22\cdot1}||^2$ and thus is the logical choice *at this stage.* Obviously, the first principal component will always provide the biggest component of variation, however this is a linear combination of all variables and so is not useful in the variable selection scenario. This procedure to extract the 'best' $m < p$ variables from the set V of all p variables is expressed more formally in Figure 6.1. This stepwise variable selection procedure using h_i statistics will be subsequently referred to as **H**.

6.4.2 On the problems of re-scaling to correlation form

It is likely that we will begin the variable selection process with an initial correlation matrix $\boldsymbol{R}$ in order to eliminate the effects of variables on larger scales dominating the results. We then use $\boldsymbol{R}$ to identify the first PV and then wish to transform to partial form. The corresponding partial correlation $\boldsymbol{R}_{22\cdot1}$ is calculated via

$$\boldsymbol{R}_{22\cdot1} = \boldsymbol{D}^{-\frac{1}{2}} \widetilde{\boldsymbol{S}}_{22\cdot1} \boldsymbol{D}^{-\frac{1}{2}}$$

where $\widetilde{\boldsymbol{S}}_{22\cdot1}$ is given in (6.2), and $\boldsymbol{D} = \text{diag}(\widetilde{\boldsymbol{S}}_{22\cdot1})$ is used to scale the matrix so the diagonals are unity. However, in this case it will be advantageous not to re-scale this resulting matrix back to correlation form. The reason for this is that a variable that is tightly associated with a selected PV will have particularly small

1. Set $V_1^{(1)} = \emptyset$, $V_2^{(1)} = V$, and $\widetilde{\boldsymbol{S}}_{22\cdot1}^{(1)} = \boldsymbol{R}$ or $\boldsymbol{\Sigma}$.

2. For $j = 1, \ldots, p$

   (a) Calculate h_i from $\widetilde{\boldsymbol{S}}_{22\cdot1}^{(j)}$. Select variable $v^{(j)}$ with the largest h_i.

   (b) Set $V_2^{(j+1)} = V_2^{(j)} \setminus \{v^{(j)}\}$ and $V_1^{(j+1)} = V_1^{(j)} \cup \{v^{(j)}\}$.

   (c) Update $\widetilde{\boldsymbol{S}}_{22\cdot1}^{(j)}$ to $\widetilde{\boldsymbol{S}}_{22\cdot1}^{(j+1)}$ using:

$$\widetilde{\boldsymbol{S}}_{22\cdot1}^{(j+1)} = \widetilde{\boldsymbol{S}}_{22}^{(j+1)} - \frac{\widetilde{\boldsymbol{S}}_{21}^{(j+1)} \left(\widetilde{\boldsymbol{S}}_{21}^{(j+1)}\right)^T}{\tilde{s}_{(j)}},$$

   where

$$\begin{aligned} \widetilde{\boldsymbol{S}}_{22\cdot1}^{(j+1)} &= \widetilde{\boldsymbol{S}}_{22}^{(j+1)} - \widetilde{\boldsymbol{S}}_{21}^{(j+1)} (\widetilde{\boldsymbol{S}}_{21}^{(j+1)})^T, \\ \widetilde{\boldsymbol{S}}_{22}^{(j+1)} &= \mathrm{Cov}[V_2^{(j+1)}], \\ \widetilde{\boldsymbol{S}}_{21}^{(j+1)} &= \mathrm{Cov}[V_2^{(j+1)}, v^{(j)}], \end{aligned}$$

   and further $\widetilde{\boldsymbol{S}}_{22}^{(j+1)}$, $\widetilde{\boldsymbol{S}}_{21}^{(j+1)}$ are simply submatrices of $\widetilde{\boldsymbol{S}}_{22\cdot1}^{(j)}$, and $\tilde{s}_{(j)}$ is the partial variance of $v_{(j)}$ on the diagonal of $\widetilde{\boldsymbol{S}}_{22\cdot1}^{(j)}$.

Figure 6.1: The iterative variable selection algorithm using h values and partial covariance (**H**).

corresponding values in the unscaled correlation matrix $\widetilde{\boldsymbol{S}}_{22\cdot1}$. This information is useful as it informs us that the merit of including this variable is now small due to its low h value, and so we would make an alternative choice for the next selected variable.

If we were to return to correlation form, then we would rescale the variance and covariances associated with this variable such that the corresponding diagonal element of $\widetilde{\boldsymbol{S}}_{22\cdot1}$ was 1. This would artificially inflate the h score of this variable and would appear, erroneously, to be more desirable than it actually was. This is especially damaging as independent variables have h values close to 1 and convey unique information that is not represented elsewhere in the data, and will now appear less desirable when compared with other variables. Furthermore, repeatedly selecting variables from groups of tightly correlated variables would introduce redundancy as each such variable would be conveying similar information.

To illustrate the merits of not returning to correlation form, consider the following numerical example. Let X_1, X_2, X_3, X_4 be random quantities with correlation matrix $\boldsymbol{R}$ such that:

$$\boldsymbol{R} = \begin{pmatrix} 1.00 & 0.90 & 0.80 & 0.00 \\ 0.90 & 1.00 & 0.75 & 0.00 \\ 0.80 & 0.75 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix}.$$

Thus we have that variables X_1, X_2, and X_3 form a group of tightly correlated variables and variable X_4 is uncorrelated with the others and seemingly independent.

Calculation of the h_i gives values of 2.45, 2.30, 2.13, and 1.00 for X_1, X_2, X_3, and X_4 respectively. Thus the selection procedure identifies X_1 as being the first PV and it is removed. The correlation matrix $\boldsymbol{R}$ is transformed to the partial form of X_2, X_3 and X_4 given X_1. This matrix $\widetilde{\boldsymbol{S}}_{22\cdot1}$ is then scaled to give the partial correlations, $\boldsymbol{R}_{22\cdot1}$:

$$\widetilde{\boldsymbol{S}}_{22\cdot1} = \begin{pmatrix} 0.190 & 0.030 & 0.000 \\ 0.030 & 0.360 & 0.000 \\ 0.000 & 0.000 & 1.000 \end{pmatrix}, \boldsymbol{R}_{22\cdot1} = \begin{pmatrix} 1.000 & 0.115 & 0.000 \\ 0.115 & 1.000 & 0.000 \\ 0.000 & 0.000 & 1.000 \end{pmatrix}.$$

Re-calculating the h_i over the correlation matrix, $\boldsymbol{R}_{22\cdot1}$, for this second stage of the variable selection gives us values of 1.08, 1.08 and 1.00 for X_2, X_3, and X_4. So there is a tie between X_2 and X_3 resulting in one of these two being chosen randomly. At the next stage we get $\boldsymbol{R}_{22\cdot1} = \boldsymbol{I}$, giving h_i values of 1 for both remaining variables which results in another tie. Thus the independent variable X_4 will either be the third or final variable, depending on how this final tie is broken. If we consider using the partial covariance matrix, $\widetilde{\boldsymbol{S}}_{22\cdot1}$, instead then the h scores are 0.037, 0.1296 and 1.0 for X_2, X_3, and X_4. Thus our second choice of PV is now the independent variable X_4.

Hence we can see that independent uncorrelated variables are dismissed in favour of including multiple closely correlated variables. This may be inappropriate when there is evidence of structure in the correlation matrix. Therefore calculating partial correlations rather than partial variances would confound and complicate the selection process. The selection procedure **H** uses partial variance and so will calculate $\widetilde{\boldsymbol{S}}_{22\cdot1}$ at each stage rather than $\boldsymbol{R}_{22\cdot1}$. Consequently, it does not suffer from these problems of rescaling.

Since we now use partial variance, the properties of the h scores given in Proposition 6.3.1 will no longer apply. This is because we are no longer restricted to having a 1 as the diagonal element of the matrix. Consequently the h values for individual variables will now fall in the interval $[0, p]$.

6.4.3 The Correlation-Based Selection Procedures

In the previous section we have dismissed working with partial correlation in favour of partial variance as it prevents problems when we have multiple tightly correlated variables. However, it is interesting to consider further what happens with a selection procedure where we rescale our partial variance to correlation form at each stage. The format of this new correlation-based selection process (**HC**) is similar to that of the original method **H** and is given in Figure 6.2.

We know already that there are some notable shortcomings to using correlations as a basis for variable selection. Therefore this new method **HC** may suffer from the same negative behaviours towards independent variables and groups of corre-

1. Set $V_1^{(1)} = \emptyset$, $V_2^{(1)} = V$, and $\boldsymbol{R}_{22\cdot1}^{(1)} = \text{Cor}[V_2^{(1)}|V_1^{(1)}] = \text{Cor}[V] = \boldsymbol{R}$.

2. For $j = 1, \ldots, p$

   (a) Calculate h_i from $\boldsymbol{R}_{22\cdot1}^{(j)}$. Select variable $v^{(j)}$ with the largest h_i.

   (b) Set $V_2^{(j+1)} = V_2^{(j)} \setminus \{v^{(j)}\}$ and $V_1^{(j+1)} = V_1^{(j)} \cup \{v^{(j)}\}$.

   (c) Update $\boldsymbol{R}_{22\cdot1}^{(j)}$ to $\boldsymbol{R}_{22\cdot1}^{(j+1)}$ using:

$$\boldsymbol{R}_{22\cdot1}^{(j+1)} = \boldsymbol{D}^{-1/2}\boldsymbol{S}_{22\cdot1}^{(j+1)}\boldsymbol{D}^{-1/2}$$

where

$$\begin{aligned}
\boldsymbol{S}_{22\cdot1}^{(j+1)} &= \boldsymbol{R}_{22}^{(j+1)} - \boldsymbol{R}_{21}^{(j+1)}(\boldsymbol{R}_{21}^{(j+1)})^T \\
\boldsymbol{R}_{22}^{(j+1)} &= \text{Cor}[V_2^{(j+1)}] \;\; (= \text{Submatrix of } \boldsymbol{R}_{22\cdot1}^{(j)} \text{ determined by } V_2^{(j+1)}) \\
\boldsymbol{R}_{21}^{(j+1)} &= \text{Cor}[V_2^{(j+1)}, v^{(j)}] \;\; (= \text{Partial row of } \boldsymbol{R}_{22\cdot1}^{(j)} \text{ determined by } v^{(j)}) \\
\boldsymbol{D} &= \text{diag}(\boldsymbol{S}_{22\cdot1}^{(j+1)})
\end{aligned}$$

Figure 6.2: The correlation-based iterative variable selection algorithm (**HC**).

lated variables discussed previously. However, it may be possible to modify this correlation-based **HC** to compensate for these problems. One way to do this could be to construct weights w_i for each variable and select variables based upon the product $w_i h_i$. The nature of the w_i must be such that the value of w_i will be close to 0 for variables that are tightly correlated to a selected PV in order to reduce their desirability for selection; and conversely for a completely uncorrelated variable w_i should be 1. All the w_i should also be 1 at the first stage, so we make the first selection based on the values of h_i only. This would counteract the problems due to working with correlations and may improve the performance of **HC**.

A logical choice is to base the weighting w_i on the size of the correlation between variable i and the last variable selected. Thus we could define the weight for variable v_i at the $j+1$ stage of selection as one of the possible $w_i^{(j+1)}$ given below:

$$w_i^{(j+1)} = 1 - |\text{Cor}[v_i, v^{(j)}|V_1^{(j)}]|, \tag{6.13}$$

$$\text{or } w_i^{(j+1)} = 1 - (\text{Cor}[v_i, v^{(j)}|V_1^{(j)}])^2, \tag{6.14}$$

where $v^{(j)}$ is the j-th PV and $V_1^{(j)}$ is the set of the first $j-1$ PVs. From this expression, we see that the weight will be inversely proportional to the size of the correlation between variables v_i and the PV $v^{(j)}$. However this only reflects the association between this variable and the last PV.

A second possibility for a suitable weighting is a recursive version of the above. Using the weight based on the squared correlation in (6.14) for reasons that will become clear in the next section, we could express our recursive weight in the form:

$$w_i^{(j+1)} = w_i^{(j)}(1 - (\text{Cor}[v_i, v^{(j)}|V_1^{(j)}])^2). \tag{6.15}$$

The potential advantage of this recursive formulation is that it will have a stronger 'memory' of the correlations between variable v_i and the previously selected variables. This method will be more aggressive in its discounting of inter-correlated groups of variables, especially if more than one variable from the same group is selected.

Both of these possible forms of weighting address the problems described above as independent uncorrelated variables will have small correlations with the selected

1. Set $V_1^{(1)} = \emptyset$, $V_2^{(1)} = V$, and $\boldsymbol{R}_{22\cdot1}^{(1)} = \text{Cor}[V_2^{(1)}|V_1^{(1)}] = \text{Cor}[V] = \boldsymbol{R}$, $w_i^{(1)} = \mathbf{1}$.

2. For $j = 1, \ldots, p$

   (a) Calculate h_i from $\boldsymbol{R}_{22\cdot1}^{(j)}$. Select variable $v^{(j)}$ with the largest $w_i^{(j)} h_i$.

   (b) Update weightings w_i by one of:

$$\begin{aligned} w_i^{(j+1)} &= 1 - (\text{Cor}[v_i, v^{(j)}|V_1^{(j)}])^2 \\ \text{or } w_i^{(j+1)} &= w_i^{(j)}(1 - (\text{Cor}[v_i, v^{(j)}|V_1^{(j)}])^2) \end{aligned}$$

   (c) Set $V_2^{(j+1)} = V_2^{(j)} \setminus \{v^{(j)}\}$ and $V_1^{(j+1)} = V_1^j \cup \{v^{(j)}\}$.

   (d) Update $\boldsymbol{R}_{22\cdot1}^{(j)}$ to $\boldsymbol{R}_{22\cdot1}^{(j+1)}$ as in Figure 6.2.

Figure 6.3: The modified correlation-based algorithm which incorporates weighting of the variable h values (**HW1**, **HW2**).

variables giving a w_i close to 1 and will therefore be minimally altered. Groups of tightly correlated variables will be reduced in importance once one candidate variable from the group has been selected since the high correlation with the selected variable will result in a value of w_i close to 0 and so the remaining variables will be strongly down-weighted. The second formula in (6.15) has a more cumulative property which will more aggressively dismiss variables that are correlated to two or more of the selected variables.

The weighted selection procedure using weighting in (6.14) will be referred to as **HW1** and the procedure using the cumulative weighting in (6.15) will be labelled **HW2**. The algorithm for the weighted procedures is given in Figure 6.3.

If we look closer at the expression for the recursive weights in (6.15), we notice that this expression is, in fact, closely related to the multiple correlation coefficient. We have from Equation 28.56 in Stuart *et al* [115] that:

$$1 - R_{1(2\ldots p)}^2 = (1 - \rho_{12}^2)(1 - \rho_{13.2}^2)\ldots(1 - \rho_{1p.2\ldots(p-1)}^2), \tag{6.16}$$

where $R_{1(2\ldots p)}$ is the multiple correlation coefficient between variable 1 and variables

2 to p, ρ_{12} is the correlation between variables 1 and 2, and $\rho_{1p.2\ldots(p-1)}$ is the partial correlation between variables 1 and p given the variables 2 to $p-1$. The term on the right of this equation is in the same form as the expression form as the recursive weights in (6.15), leading to the identity for the recursive weight of variable i at step $j+1$:

$$w_i^{(j+1)} \equiv 1 - R^2_{i\left(V_1^{(j)}\right)}, \tag{6.17}$$

where $V_1^{(j)}$ is the set of variables selected so far.

Thus the weights can simply be expressed using the multiple correlation coefficient. Furthermore, by Equation 28.57 in Stuart *et al* [115], we can express the multiple correlation between variable 1 and a set of variables, s, using the proportion of remaining variance:

$$R^2_{1(s)} = 1 - \frac{\sigma^2_{1\cdot s}}{\sigma^2_1}, \tag{6.18}$$

where σ^2_1 is the variance of variable 1 and $\sigma^2_{1\cdot s}$ is the partial variance of variable 1 given the set of variables s. Using this expression, we can re-express the recursive weight for variable i at step $j+1$ as:

$$w_i^{j+1} = \frac{\sigma^2_{i.V_1^{(j)}}}{\sigma^2_i}, \tag{6.19}$$

where $V_1^{(j)}$ is the set of PVs selected up to the current stage. Thus this form of weight provides the most succinct and intuitive definition of the weighting, simply being an expression for the proportion of variation in variable i that we have thus far failed to account for by the selection of the variables in $V_1^{(j)}$. The weight thus will remain close to 1 if i is independent of the selected variables and will drop towards zero if one or more of the selected variables are correlated to i.

If we apply the weighing methods to the illustrative example we can observe their effects on the selection process. The variable selected at the first stage is unaffected by the weighting strategy and is chosen on the basis of h values alone, and so remains as X_1. Having identified X_1 as our first selected variable, we can now construct weights for the remaining variables. Both weighting methods give weights of 0.1, 0.2 and 1.0 for variables X_2, X_3 and X_4. Taking the product of the weights and the h values gives scores of 0.108, 0.216 and 1.000 for the three variables respectively.

Thus the next selected variable is X_4, the independent variable that was previously dismissed. If we continue to the next stage using the first weighting strategy from (6.14), we obtain weights for X_2 and X_3 of 1.00 and since both variables have h scores of 1.013 this results in a tie and a random choice determined their order of selection. Using the recursive definition from (6.15) yields weight values of 0.1 and 0.2 which causes us to prefer variable X_4 over X_3 due its lower correlation to the already-selected variable X_1.

6.4.4 Comparison of the Simple and Cumulatively-Weighted Selection Procedures

The modification of the correlation-based stepwise procedure by weighting h scores, which was discussed in the previous section, is a necessary amendment to prevent the unfair dismissal of variables which are uncorrelated or independent. In procedure **HW2**, this mechanism operates by cumulatively weighting each remaining variable by the product of the $(1 - r^2)$ where at each step r is the partial correlation between the variable in question and the variable selected at this stage given all the variables previously selected. However, it is conceivable that this weighting is an attempt to reverse the standardisation imposed by rescaling to correlation form at every step and return us to a matrix of partial variances thereby approximating the selections of the original simple selection method **H**.

The connection between these two selection procedures - the simple selection method (**H**) and this cumulatively-weighted correlation-based method (**HW2**) - is still somewhat unclear, therefore the specifics of these two methods shall be considered in some detail.

Lemma 6.4.1 The weighting of h values for **HW1** and **HW2** is equivalent to finding the h scores over a modified correlation matrix $\widetilde{\boldsymbol{R}} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{R}$, where $\boldsymbol{R}$ is the original correlation matrix and $W = \text{diag}(()w)$ is a diagonal matrix of weights.

Proof: With an initial correlation matrix $\boldsymbol{R} = (r_{ij})$, the h statistic for variable j is

defined as $h_j = \sum_i r_{ij}^2$. Weighting the h statistics gives:

$$\begin{aligned} w_j h_j &= w_j \sum_i r_{ij}^2 \\ &= \sum_i (\sqrt{w_j} r_{ij})^2. \end{aligned}$$

Thus weighted h values for $\boldsymbol{R}$ are equivalent to taking unweighted h values over the matrix with elements $\sqrt{w_j} r_{ij}$, i.e. $\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{R} = \widetilde{\boldsymbol{R}}$. □

The simple selection procedure **H** uses $\boldsymbol{R}$ to inform its selection of principal variables. However we can see from the above lemma that the correlation-based weighted methods employ a modified matrix $\widetilde{\boldsymbol{R}}$ instead. The selection procedures are the same for both methods – both calculate h statistics over their input matrices and select the variable with the maximum h value – so the difference is in the input matrix alone. It is the relationship between these matrices $\boldsymbol{R}$ and $\widetilde{\boldsymbol{R}}$ which will characterise the differences between the performance of the two selection procedures. Since $\widetilde{\boldsymbol{R}}$ is dependent on the weights being used in the selection process, the relationship between **H** and **HW2** will also depend on the form of these weights. For what follows, we assume that the weights are of the form given in (6.15). The connection between these two forms of the correlation matrix are now such that we can make the following statement.

Theorem 6.4.2 At each stage of variable selection, the input matrix $\widetilde{\boldsymbol{R}}$ for the cumulatively-weighted correlation-based variable selection procedure **HW2** is given by

$$\widetilde{\boldsymbol{R}} = \widetilde{\boldsymbol{S}} \boldsymbol{D}^{-\frac{1}{2}},$$

where $\widetilde{\boldsymbol{S}}$ is the unscaled correlation matrix used by the simple selection procedure **H** and $\boldsymbol{D} = \text{diag}(\widetilde{\boldsymbol{S}})$.

Proof: We know from the lemma that $\widetilde{\boldsymbol{R}} = \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{R}$. Since the correlation matrix $\boldsymbol{R}$ is obtained by scaling $\widetilde{\boldsymbol{S}}$ to have diagonals equal to unity, we can say that

$$\widetilde{\boldsymbol{R}} = \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{D}^{-\frac{1}{2}} \widetilde{\boldsymbol{S}} \boldsymbol{D}^{-\frac{1}{2}}.$$

Hence to obtain the given result, all we must show is that $\boldsymbol{W} = \boldsymbol{D} = \text{diag}(\widetilde{\boldsymbol{S}})$. This can be shown by induction.

For the first variable, the initial weights are all 1, so we have that $\boldsymbol{W}^{(1)} = \boldsymbol{I}$. We also know that $\widetilde{\boldsymbol{S}}^{(1)} = \boldsymbol{R}$ at the first step, thus giving $\text{diag}(\widetilde{\boldsymbol{S}}^{(1)}) = \boldsymbol{I} = \boldsymbol{W}^{(1)}$. Hence, the result holds for the first variable.

Now assume that the result holds for variable $k-1$. Under this assumption, the diagonal elements of $\widetilde{\boldsymbol{S}}^{(k-1)}$ are given by the weights $\boldsymbol{W}^{(k-1)}$, so

$$\widetilde{s}_{i\cdot 12\ldots(k-2)} = w_i^{(k-1)} = 1 \times (1 - r_{i1}^2) \times (1 - r_{i2\cdot 1}^2 \cdots \times (1 - r_{i(k-2)\cdot 12\ldots(k-3)}^2),$$

where $\widetilde{s}_{i\cdot 12\ldots(k-2)}$ is the ith diagonal element of $\widetilde{\boldsymbol{S}}^{(k-1)}$, and $r_{ij\cdot 12\ldots(j-1)}$ is the partial correlation between the ith and jth variables given the first $(j-1)$ variables.

Assuming the result holds for variable $k-1$, we now consider variable k. Since the weights are defined cumulatively, the weight for variable i will be now given by:

$$w_i^{(k)} = w_i^{(k-1)}(1 - r_{i(k-1)\cdot 12\ldots(k-2)}^2).$$

To find the form of $\widetilde{\boldsymbol{S}}^{(k)}$, we can use the definition of partial correlation to obtain the following result for the diagonal elements of $\widetilde{\boldsymbol{S}}$:

$$\begin{aligned} \widetilde{s}_{i\cdot Aj} &= \widetilde{s}_{i\cdot A} - \frac{\widetilde{s}_{ij\cdot A}^2}{\widetilde{s}_{j\cdot A}} \\ &= \widetilde{s}_{i\cdot A}(1 - r_{ij\cdot A}^2), \end{aligned}$$

where i and j are single variables and A is a set of one or more variables. This allows us to define the diagonal elements of $\widetilde{\boldsymbol{S}}^{(k)}$ as:

$$\begin{aligned} \widetilde{s}_{i\cdot 12\ldots(k-1)} &= \widetilde{s}_{i\cdot 12\ldots(k-2)}(1 - r_{i(k-1)\cdot 12\ldots(k-2)}^2) \\ &= w_i^{(k)}. \end{aligned}$$

So we have that if $\boldsymbol{W} = \text{diag}(\widetilde{\boldsymbol{S}})$ holds for variable $k-1$ then it also holds for variable k. Since this property holds for the first variable, it must be true for all $k \in \mathbb{Z}^+$. □

The significance of demonstrating that $\widetilde{\boldsymbol{R}}$ can be expressed in the form $\widetilde{\boldsymbol{S}}\boldsymbol{D}^{-\frac{1}{2}}$ may not be directly apparent. What this result shows is that the weighting process used by the correlation-based method **HW2** is attempting to reverse the scaling effect attributable to using partial correlation instead of partial covariance or unscaled correlation. The weighting is attempting to take $\widetilde{\boldsymbol{R}}$ and return it to its unscaled

form, $\widetilde{\boldsymbol{S}}$, and thereby mimic the behaviour of the simple unscaled selection procedure **H**. However, the weighting does not successfully eliminate all of the effects of the rescaling process, as shown by the presence of $\boldsymbol{D}^{-\frac{1}{2}}$. This failure to completely remove the scale effects will result in **HW2** behaving as a biased and therefore inferior version of **H**. Due to this deficiency in the correlation-based method, in the future we shall exclusively prefer the variable selection procedure **H** which operates directly with the unscaled partial correlation, $\widetilde{\boldsymbol{S}}_{22\cdot1}$.

6.5 Extensions to the Variable Selection Procedure

6.5.1 Incorporating Temporal Information

One key feature of the two data sets described in Section 2.2 is that several of their variables constitute a repeated measures data set. That is to say, a group of variables are repeatedly measured on the same observational units at different time points. These repeated measures data sets are commonplace and are neither peculiar to medicine nor orthopædics. The standard selection procedure as it stands can be applied to each time point in the data individually and thus generate a sequence of variable sets each containing the extracted variables for each time point. However, having different sets of variables for each time point can result in different and contradictory variable subsets especially if the correlations change over time. It would therefore be preferable to determine a single overall subset of *longitudinal* principal variables to represent the full time spectrum. Doing so could provide significant reduction in time and cost for data collection, as well as simplifying subsequent analyses and their interpretations.

Incorporating a temporal aspect into standard PCA has been examined in various different ways. One of the main difficulties in these approaches is to account successfully for both the temporal and multivariate nature of the data, by exploiting both the correlation structure within variables at one time point and their associations between time points. Typically methods based on stationary time series [68]

are inappropriate as it is unlikely that the repeated measures data will constitute a stationary time series, and repeated measures data sets can have only a small number of time points which would pose problems for such analysis. Berkey *et al* [8] discuss a *longitudinal principal components* regression model which performs principal components on the various observations of a single variable over time and then uses the resulting principal components as predictors in a linear model. However, interactions between the variables are not considered, so whilst accounting for the *longitudinal* nature of the data, the *lateral* nature is ignored. Additionally, the model is one of regression with a defined response which is not appropriate to variable extraction from repeated measures data as we have no such defined responses. A further method for incorporating a temporal element is that of *functional data analysis* [102] whereby the data could be considered to be a functional time series. However, this method would typically require a large number of time points and cases than we have available as they are based on fitting splines through the observations.

The method used here to accommodate for the temporal dimension to the data is based on the nonparametric time dependent PCA techniques proposed by Prvan and Bowman [100]. Suppose we have a data matrix $\boldsymbol{X}$ containing n cases $\boldsymbol{x}_i$, $i = 1, \ldots, n$ each observed at a time point t_i. First, we choose a focal time point θ. We then associate with each $\boldsymbol{x}_i$ a weight ω_i defined as:

$$\omega_i = \omega(t_i, \theta, \sigma) = \phi\left(\frac{t_i - \theta}{\sigma}\right) \tag{6.20}$$

where σ is a bandwidth parameter and $\phi(\cdot)$ is the standard Normal density function. The purpose of these weights is to assign each case $\boldsymbol{x}_i$ in the data a weight ω_i that represents the distance between its associated time point t_i and the time point of interest θ. Cases with $t_i = \theta$ will receive the highest weights and the greatest influence over the following calculations. Weights for other time points will become progressively smaller the further in time t_i is from θ giving the associated cases less importance.

Having defined a weighting strategy for the cases in the data, we can now construct the weighted mean $\bar{\boldsymbol{x}}_\omega(\theta)$ and the weighted variance matrix $\boldsymbol{S}_\omega(\theta)$ which are

defined as:

$$\bar{\boldsymbol{x}}_\omega(\theta) = \frac{1}{\sum_{i=1}^n \omega_i} \sum_{i=1}^n \omega_i \boldsymbol{x}_i, \tag{6.21}$$

$$\boldsymbol{S}_\omega(\theta) = \frac{1}{\sum_{i=1}^n \omega_i} \sum_{i=1}^n \omega_i (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_\omega(\theta))(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_\omega(\theta))^T. \tag{6.22}$$

From (6.22) we can obtain the weighted correlation matrix $\boldsymbol{R}_w(\theta)$, by re-scaling $\boldsymbol{S}_w(\theta)$ in the usual way. Thus we have obtained a matrix of correlations over the variables in the data relative to a particular time point θ. The notable feature of this method is that we allow *all* the data to influence the value of the mean and variance at time point θ, rather than just looking at the data directly observed at that time. This means that, by virtue of the smoothing, data observed at a time close to the point under consideration has a perceptible input on the values of $\bar{\boldsymbol{x}}_\omega(\theta)$ and $\boldsymbol{S}_\omega(\theta)$, and data at distant times have a minimal effect due to their low weights. This reinforces the notion that data observed at time points that are close together are likely to be related, whereas that relationship may change in the intervening period.

The magnitude of the effect of temporally adjacent data and the distance in time over which it applies is governed by the bandwidth parameter σ. The choice of σ is typically subjective and is based on the plots of component loadings versus time discussed in [100]. The plots are constructed by first performing the temporal PCA for each time point at which data is recorded, then for each smoothed PC we plot the values of its loadings against time. The idea being that a bandwidth that is too large will mask curvature in the data, whereas a bandwidth that is too small will result in obvious displays of sample variation. Furthermore, if the loadings in each PC remain relatively constant over time then this suggests that there is little evidence of changes in structure over time. More detailed assessments of the time effect can be achieved via the calculation of reference bands [100].

Having established the fundamentals of this temporal smoothing process, we can then use our smoothed correlation matrix $\boldsymbol{R}_\omega(\theta)$ as the input to the variable selection procedure. This still requires us to focus at a particular time, however the effects of data observed at other time points are not ignored and are incorporated into the analysis via the smoothing mechanism. It also allows for the possibility of

exploring results for a time point that has not been directly observed. However, we must still apply the variable selection technique at each time point of interest. This is not desirable as an overall subset of key variables is the intended goal, rather than a series of likely different subsets.

This procedure is reasonable for determining longitudinal PVs if we believe that the multivariate structure is preserved across time points, excepting random fluctuation. This could be checked formally by sphericity-type tests [76], as long as we were prepared to make further distributional assumptions. Informally, we can examine the plots discussed in [100] which are used for choosing σ to assess this.

To determine an overall subset of longitudinal PVs, we construct $\boldsymbol{R}_\omega(t_i)$ for each time point t_i for which we make a measurement. We calculate h_{j,t_i} (6.12) for each variable j at each time point t_i. A simple guide to the selection value of each variable across all time points is then $\bar{h}_j = \sum_i h_{j,t_i}$. The variable with the highest average value of h_i could then be selected. All of the correlation matrices are then updated to partial covariance form given the variable we have selected. The full temporal algorithm is detailed in Figure 6.4, and labelled **HT**.

This process allows for each time point to contribute to the overall selection process and provides a simple way to combine the results to generate an overall 'best' subset. One point of note is that if it were determined that, say, five variables could sufficiently describe each time point and if there was evidence of a change in the structure or relationships between the variables over time, then five variables would not describe the data at all time points to the same degree. Consequently, it may be necessary to increase the number of variables retained in order to adequately describe the data at all times. This could be addressed, for example, by considering the scree-type plots discussed in Section 6.5.4.

Since the calculation of the overall 'best' subset involves taking a simple mean of the weighted scores across all time points, one could also introduce a further area of customisation of the procedure by introducing an additional set of weightings for the time points, κ_{t_i} and then construct a weighted mean $h_j^* = \sum_i \kappa_{t_i} h_{j,t_i}$. This would allow certain time points to have more of a contribution to the selection process. This could be especially applicable when the sample size is not constant across all

1. Set $V_1^{(1)} = \emptyset$, $V_2^{(1)} = V$. For each time point t: $\boldsymbol{S}^{(1)}_{\omega,22\cdot1}(t) = \boldsymbol{R}_\omega(t)$, where $\boldsymbol{R}_\omega(t)$ is the smoothed correlation matrix for time point t.

2. For $j = 1, \ldots, p$

   (a) For each time point t: From $\boldsymbol{S}^{(j)}_{\omega,22\cdot1}(i)$ calculate the $h_i(t)$ for each variable v_i.

   (b) Calculate $\overline{h_i} = \sum_t (h_i(t))/T$, where T is the number of time points under consideration. Select variable $v^{(j)}$ which maximises $\overline{h_i}$.

   (c) Set $V_2^{(j+1)} = V_2^{(j)} \setminus \{v^{(j)}\}$ and $V_1^{(j+1)} = V_1^{(j)} \cup \{v^{(j)}\}$.

   (d) For each time point t: update $\boldsymbol{S}^{(j)}_{\omega,22\cdot1}(t)$ to $\boldsymbol{S}^{(j+1)}_{\omega,22\cdot1}(t)$ as in Figure 6.1.

Figure 6.4: The modified algorithm (**HT**) which incorporates a temporal aspect.

time points and some degree of corresponding compensation is desired. In fact, there are a great many such extensions that could be considered to this selection process, however incorporating further sophistication might make little difference in a given application. Nevertheless, such extensions and developments of the methodology could be the focus for future research.

6.5.2 Utilities

6.5.2.1 Motivation

Whilst the selection of variables based solely on the information conveyed in the data is, in many cases, the best course of action it is likely that there will be cases where one would wish to adjust the selection process. For example, in a medical context it is conceivable that a clinician will have an opinion on the relative usefulness of particular measurements for their diagnosis or monitoring of a patient. One would then wish to utilise this information to guide the selection process towards choosing variables that the clinician deemed more useful than others. Similarly, there may be

variables that one would wish to always include into, or exclude from, the returned subset. In the case of the former one could simply examine the remaining variables separately and then combine the results with that variable after the analysis. However, this would not account for the effects of forcing the selection of that particular variable when the procedure was making choices over the other variables. Finally, the clinician may also wish to penalise variables according to how easy the data is to collect since measuring some variables may require great expense or great discomfort for the patient.

In Section 6.4.3, the flawed correlation-based selection procedure **HC** was modified to prevent the dismissal of independent, uncorrelated variables by including a weight term w_k to reduce the appeal of variables which were highly correlated to other variables that had already been selected (**HW2**). There, the w_k term sought to down-weight the variables' associated h_k scores - if w_k was 1 then h_k was unmodified, however if w_k was close to zero then the associated variable's desirability becomes negligible and it is effectively eliminated from consideration. This weighting procedure was ultimately determined to be inferior to working with the partial variances directly, however the notion of a weighting to modify the desirability of a variable still remains a potentially interesting concept. The questions being asked here are: is there scope to incorporate a set of user-defined subjective weights or utilities for the perceived merit of retaining a particular variable, and furthermore will it help us to tackle situations such as those described above?

The integration of such subjective information into the variable selection process would enable the procedure to be guided in an informed manner and would allow for the combination of information gathered from the data with information available externally. This would be most useful when there are several variables with values of h_i close to the maximum and would allow for an alternative choice to be made depending on the values of the associated utilities. For example, if the utility represented the ease of measurement of a particular variable, then in the case of a tie we would favour the retention of the variable that is the easiest to measure. This would also allow for a degree of customisability of the selection procedure with individual preferences becoming relevant when choosing between variables, rather than basing

the choice exclusively on the data alone. In addition, the use of expert judgements for these utilities would, no doubt, allow for results that were more practically useful and applicable to the context of the problem under investigation.

6.5.2.2 Scaling and Transformations

The incorporation of a utility term into the procedure is straightforward. Suppose that u_i is the utility for retaining variable v_i. A simple way of modifying the selection process is to replace h_i by $h_i^U = u_i h_i$ and to select at each stage the variable with the maximum value of h_i^U. Whereas in the obsolete weighted selection procedure **HW2**, the w_i could only reduce the desirability of the variables, the utilities u_i can both increase and decrease variable suitability. It is important however that the utility measures can guide the selection process, but they must not dominate the results - that is to say the variable selected should be done so on the basis of a combination the information obtained from the data and the subjective utilities.

It would be desirable to configure the utilities so that if a variable was assigned the highest possible score then that variable would be forced into the resulting subset. Conversely, if a variable had the minimum possible score then this would have the opposite effect and prevent the corresponding variable's selection. If the utility score was scaled so that $u_i \in [0, 1]$ then one transformation that would have these properties would be:

$$f(x) = \frac{x}{1-x}.$$

This function would map the u_i from $[0, 1]$ to $[0, \infty]$. This would allow for variables with a utility score of 0 to always have a desirability of 0 thereby effectively preventing its selection; conversely a utility score of 1 would translate to an infinite desirability which would force its immediate selection. Additionally, an untransformed utility score of 0.5 would be equal to 1 after the transformation, leaving the desirability of the variable equal to h_i.

However, there are some possible problems here when we seek to combine these quantities. The first problem could arise if we have a utility score of ∞ and $h_i = 0$, which has an undefined product. In this case, it would be best to use only the utility score to define the variable's desirability. When we have multiple variables with the

equal desirability and we seek to choose between them then the choice should be made, again, on the basis of the utility values. We would seek to favour the expert utility information over that of the data in this case. If the utilities themselves are equal then we should break the tie randomly.

6.5.2.3 Multiple Utilities

Unlike with a single utility measure, the case with multiple utilities poses a different problem. It is not unreasonable to assume that there may exist multiple utilities for each variable. These could be utilities measuring different factors of the variables, for example in the medical context discussed above two such utilities could be the ease of measurement and clinical usefulness of the variable. Equally, we could have multiple utilities which reflect the opinions of different individuals on the same criterion. The primary problem arises in how best to combine such information and then utilise it within the variable selection framework.

Essentially this corresponds with seeking to combine a number of different 'votes' on which variable to choose next with each 'vote' corresponding to a utility measure. This area is known as social choice theory and method of combination we seek is known as a social welfare function. The simplest method of combination is the approach of *utilitarianism* [10] whereby the individual votes are simply summed. Thus from several utility vector $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(l)}$, we construct $\boldsymbol{u}^* = \sum_i \boldsymbol{u}^{(i)}$ to aggregate the individual components and then consider the product $u_i^* h_i$ for selection. Utilitarianism is not the only mechanism for combining multiple utilities in this manner - alternatives such as maximin, maximax, leximin, and Cobb-Douglas could be considered as alternative combination mechanisms. As it stands utilitarianism has several beneficial properties (strong Pareto, anonymity, and continuity) and is an intuitive and easy to apply method. See [10] for a discussion of these methods and their properties.

6.5.3 Improving the Search Procedure

The stepwise progression through the variables in the data corresponds to a form of search, where our goal is the best subset of variables as defined by the h_i. The

stepwise selection procedure equates to a restricted version of a *best first* search with a heuristic equal to h_i. That is to say at each step we evaluate h_i for all remaining variables and choose the variable v which is the best and then proceed to the next step, never entertaining the possibility of selecting any variable other than v at this stage. It may be the case that selecting a non-optimal variable at this step leads to a better subset of variables in the end.

Considering exploring other combinations of variables that were previously discounted is a worthy extension to the variable selection process. Thus the simple stepwise search could be improved upon by considering a tree-based search strategy. The application of simple search algorithms such as A* search [113] would likely improve the quality of the returned subset, at the expense of additional computation. One advantage of using an algorithm and search strategy such as A* search is that it is guaranteed to find the optimum value and do so in an optimally efficient manner provided the heuristic is 'admissible'.

An alternative modification to the search procedure would be to consider using local search procedures, such as those using simulated annealing and genetic algorithms. These procedures were used in [13] to enhance the variable selection process when performing an exhaustive search for a suitable variable subset, rather than using a stepwise approach.

At present, the variable selection procedure does not incorporate an advanced search routine. Such developments could be the focus for future research.

6.5.4 'Scree'-type Plots

The scree plot [16] is a popular method for assessing the number of principal components which represent the 'signal' in the data. The variances of each principal component are plotted in descending order of magnitude and one looks for the principal component beyond which the variances decrease in a linear fashion. The components beyond this point are taken to be consonant with random noise. Such graphical summaries are both useful and informative and an equivalent such plot can also be achieved for the h-based procedure. There are several possibilities for a suitable scree plot which will be mentioned briefly here and developed further in

the context of actual data in Section 6.7.2.2.

The first candidate for a scree-type graphic is to plot the h_i of the selected variables in the sequence in which they were selected. That is to say, if we select variable $v_{(j)}$ at stage j with an associated h-value of $h_{(j)}$, then we plot $h_{(j)}$ against j. We know from the properties of h_i that variables which have a score $h_i > 1$ represent a variable that is correlated to others, but has negligible correlation to the variables already extracted - these variables are the ideal candidates for selection and are removed first. Scores of $h_i = 1$ represent a single independent variable or a correlated variable that now has low partial covariances due to an association with a previously selected variable. Finally, scores of $h_i < 1$ indicate variables which are moderately to strongly associated with one or more variables that have already been selected. Therefore, if the variables are not independent then we would expect to see a decreasing concave curve as the variables are selected in descending order of their h values. The steepness of this descent will depend upon whether there is one large tightly correlated group which would give a steep drop in the h or whether there are several smaller groups which would produce a slower reduction. The curve will then level off at $h = 0$ once the most correlated variables are selected.

Another possibility is to plot the amount of variation explained by our current subset of PVs. After removing j variables we obtain the unscaled partial correlation matrix $\widetilde{\boldsymbol{S}}^{(j)}_{22\cdot 1}$. We could then plot the values $\mathrm{tr}(\widetilde{\boldsymbol{S}}^{(j)}_{22\cdot 1})$ or $||\widetilde{\boldsymbol{S}}^{(j)}_{22\cdot 1}||^2$ vs. j for $j = 0, \ldots, p$ where $\boldsymbol{S}^{(0)}_{22\cdot 1} = \boldsymbol{R}$. Whilst the trace of the matrix is most commonly associated with the traditional proportion of variation, the squared norm is intimately related to the variable selection process and so could be equally useful. The trace will vary less dramatically due to the fact that the trace uses only the diagonal elements of the variance matrix whereas the squared norm uses the square of all the elements. Hence changes in variability will likely be more visible on plots of the squared norm. However, it should also be noted that $\mathrm{tr}(\boldsymbol{R}) = \sum_i \lambda_i$ and $||\boldsymbol{R}||^2 = \sum_i \lambda_i^2$, so there will be a strong degree of similarity between both plots of either value.

If all the variables are independent then $\mathrm{tr}(\boldsymbol{R}) = \sum_i \lambda_i$ and $||\boldsymbol{R}||^2 = \sum_i \lambda_i^2$ would decrease in a perfectly linear fashion since each variable would have an equal contribution to the overall variance. However, if there is correlation then the curve

will be concave with an initial drop as the most correlated variables are removed which prompts a large reduction in the trace or squared norm. As more variables are removed they will become progressively more conditionally independent given the selected variables as the set of selected variables increases. Thus the curve will begin to straighten out as the h_i scores of the remaining variables become close to zero reflecting only a small change in $\text{tr}(\boldsymbol{S}_{22\cdot 1})$ or $||\boldsymbol{S}_{22\cdot 1}||^2$. This is similar to the property of the scree plot for principal components where there is a straightening out of the scree curve once the most important components are selected.

As an alternative to the above plot, we could plot the cumulative proportion of the total variation explained in terms of $\text{tr}(\widetilde{\boldsymbol{S}}_{22\cdot 1})$ or $||\widetilde{\boldsymbol{S}}_{22\cdot 1}||^2$. In this case we would calculate the proportions:

$$p_j = \left(1 - \frac{\text{tr}(\widetilde{\boldsymbol{S}}_{22\cdot 1})}{\text{tr}(\boldsymbol{R})}\right) \tag{6.23}$$

which we plot against j, where π_p is set to 1. This would give a plot that was a reversal of the previous plot with the curve being convex before straightening out, however it is likely more interpretable than the previous plot as it operates in terms of the proportion of total variation than in the values of $||\widetilde{\boldsymbol{S}}_{22\cdot 1}||^2$ themselves. The equivalent plot could be performed using $||\cdot||^2$.

Some examples of possible scree plots are given in Figure 6.5. These scree plots are for a data set with six variables, three of which are merely noisy copies of the other three. Therefore we assume that the effective dimensionality of the data is three. The first plot in Figure 6.5(a) is the standard scree plot of the squared eigenvalues in descending order. The importance of the first three components is evident through the large variances, the components associated with the random noise are indicated by their low values of λ^2. The plot in Figure 6.5(b) is that of h values of the principal variables in the order of selection. The shape clearly mirrors that of the standard scree plot with a sharp drop in variable importance after the selection of the third PV. The plot in Figure 6.5(c) is that of the $||\boldsymbol{S}_{22\cdot 1}||^2$ after the selection of each PV. We can see that after the selection of the third variable there is little subsequent change in the value of $||\boldsymbol{S}_{22\cdot 1}||^2$ suggesting that the remaining variables are consonant with random noise. Finally, the plot in Figure 6.5(d) is of the percentage trace that is explained by the chosen PVs (see (6.23)). We can see there

that amount of variability explained rapidly increases as we select the first three PVS, but after this point we capture progressively less of the variation indicating that these variables are contributing less novel information.

6.6 Assessment of Dimensionality

The assessment of dimensionality in terms of the number of variables is an important component of the variable selection problem as it directly informs us about the number of variables to select. In essence, the dimensionality of the data acts as a stopping rule for the selection process. However, the existing methods discussed in Section 6.2.3 do not easily lend themselves to application in stepwise variable selection. The methods based on principal components are inappropriate when working with individual variables as they require retention of all variables. For example, Kaiser's rule of discarding components with eigenvalues less than one is meaningless when working in terms of variables. The cross-validatory approaches of Wold and Eastment & Krzanowksi are both computationally intensive and time consuming for large data sets and so would not lend themselves to the study of the orthopædic data.

Velicer's method of calculating the average remaining partial correlation given m principal components and stopping at a minimum, may be applicable. However, when using the **H** procedure the average partial covariance given the first m variables will typically be decreasing and will not behave in the same way as Velicer's V statistic.

Methods based on arbitrary thresholds would still be appropriate when dealing with variables rather than dimensions. For example, one could cease the selection process when we have captured $\alpha\%$ of the total variation as defined in (6.23). The selection procedure could then stop selecting variables once we have identified that our selected subset represents at least 75%, say, of the original variation of the data. Furthermore, as an analogue to Kaiser's rule for principal components, we know that the h statistics represent the contribution of an individual variable to the value of $||\boldsymbol{R}||^2$ with an h value of 1 corresponding to the value of a single independent

variable. Therefore, if all the h statistics of the remaining variables fall below such a threshold then the remaining variables contain less novel information than a single independent variable due to their associations with the selected subset. This could be a logical stopping point for the selection process, though the threshold may need modification to tolerate sample variation as did Kaiser's rule.

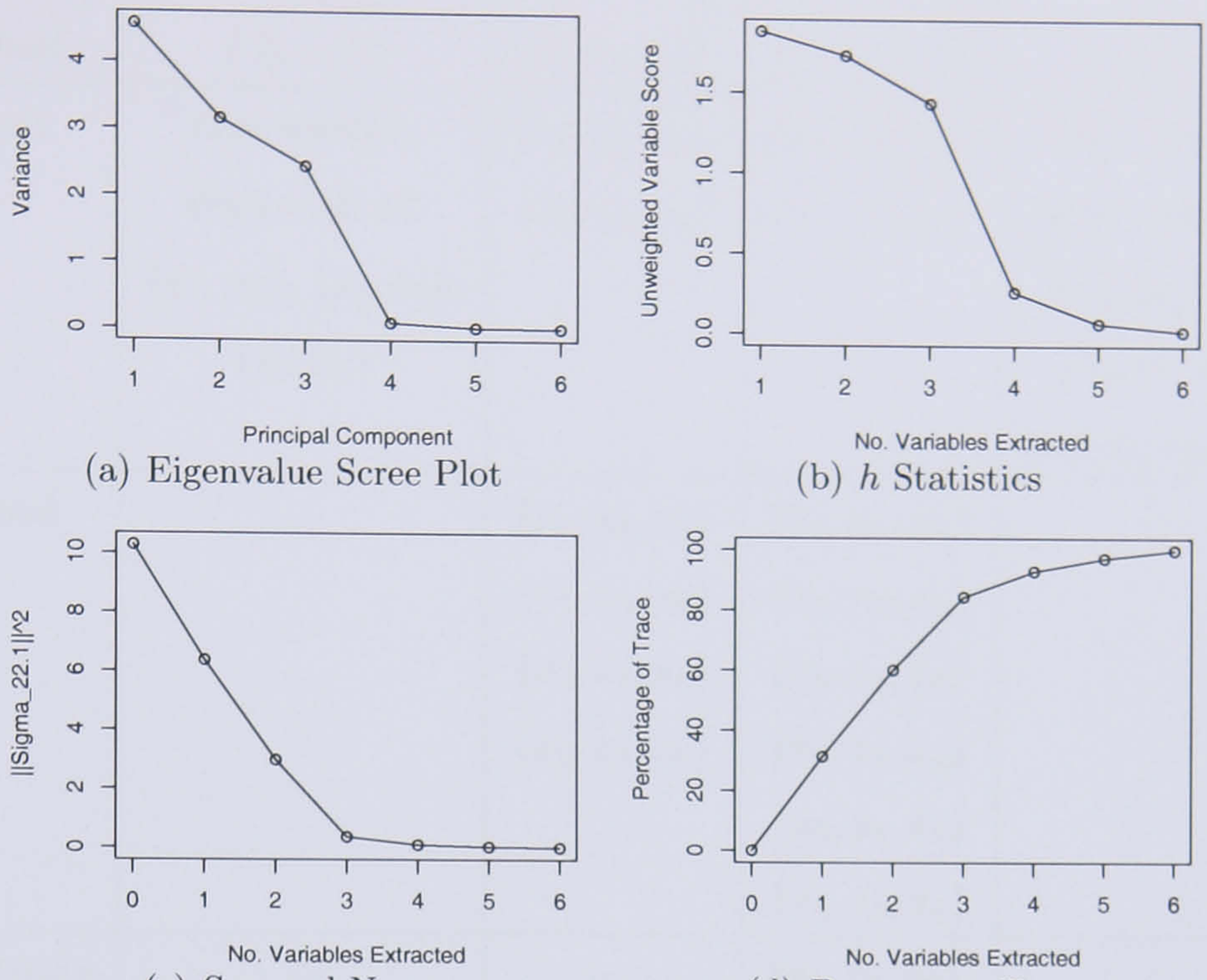
6.7 Results

6.7.1 Artificial Data

6.7.1.1 Simple Models

In order to assess the efficiency of the variable selection procedures developed in the previous sections, a Monte Carlo simulation study was performed. This study was the same as the one used to compare methods of variable selection by Jolliffe [66], and also by Krzanowski [75]. The study was performed in four parts, each part testing performance on simulated data conforming to a different pre-determined model. Each part was then repeated 500 times, each time re-simulating the data to allow sample variation. Each model was created so that certain variables in the model were linear combinations of the others plus random noise - these variables are therefore redundant. With the a priori knowledge of the data structure, it is possible to determine what variables should be extracted from these data. Consequently, a returned subset can be classified as being "Best", "Good", "Moderate" or "Bad."

For each iteration of the simulation on a particular model, 100 cases of data were generated containing between six and ten variables. These data conformed to models defined by Jolliffe [66] as Models *I–IV*; the exact specification of these models is reproduced in Table 6.1. The structure of these models is such that the first contains a set of three pairs; the second is a pair, a triple and a single independent variable; the third is effectively three pairs though with stronger correlations elsewhere; and the fourth is composed of a single variable, a pair, a triple and a quadruple. The first three models have three key variables, whereas the fourth has four. Correlation plots to illustrate this structure are given for a random data set of size 100 in Figure



(a) Eigenvalue Scree Plot (b) h Statistics

(c) Squared Norm (d) Percentage Trace

Figure 6.5: Scree plots for simulated data.

Variable	Model			
	I	*II*	*III*	*IV*
x_1	z_1	z_1	z_1	z_1
x_2	z_2	z_2	z_2	z_2
x_3	z_3	z_3	z_3	$z_2 + z_3$
x_4	$z_1 + 0.5z_4$	$z_1 + 0.5z_4$	$z_1 + 0.8z_2 + 0.6z_4$	z_4
x_5	$z_2 + 0.7z_5$	$z_2 + 0.7z_5$	$z_2 + 0.7z_5$	$z_4 + 0.75z_5$
x_6	$z_3 + z_6$	$z_2 + z_6$	$z_3 + 0.5z_6$	$2z_4 + 0.75z_5 + 1.5z_6$
x_7	—	—	—	z_7
x_8	—	—	—	$z_7 + 0.5z_8$
x_9	—	—	—	$2z_7 + 0.5z_8 + z_9$
x_{10}	—	—	—	$3z_7 + z_8 + z_9 + z_{10}$

Table 6.1: Definition of Jolliffe's simulated data Models I-IV.

The z_i are independent standard Normal variables.

Type of subset	Model			
	I $(p=3)$	II $(p=3)$	III $(p=3)$	IV $(p=4)$
Best	One variable from each of: $\{x_1, x_4\}$, $\{x_2, x_5\}$, $\{x_3, x_6\}$	$\{x_1, x_2, x_3\}$, $\{x_2, x_3, x_4\}$	$\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_6\}$	One variable from each of: $\{x_1\}$,$\{x_2, x_3\}$, $\{x_4, x_5, x_6\}$ $\{x_7, x_8, x_9, x_{10}\}$
Good	—	$\{x_1, x_3, x_5\}$ $\{x_1, x_3, x_6\}$ $\{x_3, x_4, x_5\}$ $\{x_3, x_4, x_6\}$	$\{x_1, x_5, x_6\}$ $\{x_1, x_3, x_6\}$ $\{x_2, x_4, x_6\}$ $\{x_2, x_3, x_4\}$ $\{x_3, x_4, x_5\}$ $\{x_4, x_5, x_6\}$	—
Moderate	—	—	$\{x_1, x_3, x_6\}$, $\{x_1, x_4, x_6\}$	—
Bad	Any other subset			

Table 6.2: Classification rules for selected variable subsets for Models I–IV.

6.6. The classification of variable subsets is detailed in Table 6.2.

For each generated data set, the correlation matrix was calculated and a suite of eleven different variable extraction methods was then performed. These methods are taken from those discussed in Section 6.2 are summarised in Table 6.3. All calculations were run in R for Windows version 2.1.1 [101] on a Pentium IV 2.4GHz PC with 1.5Gb RAM.

The results for the simulation study are presented in Table 6.4 in terms of the suitability of the returned subsets. The results are presented for each individual model, and then an overall performance is given at the bottom of the table.

For Model I, which is composed of three correlated pairs of variables, we can see that the majority of the selection procedures make accurate selections by selecting only one from each pair of variables. The principal components forward selection method **B4** appears to struggle with this simple model, being correct only about three times in four. Both the simple (**H**) and correlation-based (**HC**) selection

	Description	Type
M1	McCabe's first solution to the principal variable problem. An exhaustive search of all subsets of appropriate dimensions to find that subset which satisfied $\min \lvert\boldsymbol{\Sigma}_{22\cdot 1}\rvert$.	Exhaustive
M2	As M1, only returning the subset which satisfied $\min \operatorname{tr}(\boldsymbol{\Sigma}_{22\cdot 1})$.	Exhaustive
M3	As M1, only returning the subset which satisfied $\min \lvert\lvert\boldsymbol{\Sigma}_{22\cdot 1}\rvert\rvert^2$.	Exhaustive
B1	Jolliffe's first backward variable selection procedure using the principal components of $\boldsymbol{\Sigma}$, removing variables associated with the highest loadings in the final component. The principal component analysis is repeated for the remaining variables each time a variable is selected.	Stepwise
B2	Jolliffe's second backward variable selection procedure using the principal components of $\boldsymbol{\Sigma}$. The principal component analysis is performed only once.	Stepwise
B4	Jolliffe's forward variable selection procedure using the principal components of $\boldsymbol{\Sigma}$, associating selected variables with the highest loadings in the first components.	Stepwise
A1	Beale's multiple correlation method. An exhaustive search of all subsets of appropriate dimensions to find that subset which maximised the minimum multiple correlation between the p selected variables and any of the remaining variables.	Exhaustive
A2	Jolliffe's stepwise version of A1.	Stepwise
KP	Krzanowski's Procrustes method.	Stepwise
DF	The graphical Gaussian selection method of De Falguerolles *et al* seeking the set of variables given which the remaining variables are conditionally independent.	Exhaustive
H	The initial simple selection procedure using h statistics.	Stepwise
HC	The correlation-based selection procedure using h statistics.	Stepwise

Table 6.3: Summary of the variable selection methods tested via Monte Carlo simulation.

Model	Subset type	Selection method											
		M1	**M2**	**M3**	**B1**	**B2**	**B4**	**A1**	**A2**	**KP**	**DF**	**HC**	**H**
I	Best	100	100	100	100	99.4	74.2	100	100	100	100	100	100
	Bad	0	0	0	0	0.6	25.8	0	0	0	0	0	0
II	Best	29.8	99.4	99.4	0.2	0.6	97.0	0	0.2	4.6	22.4	24.2	99.2
	Good	70.2	0.6	0.6	99.8	99.4	3.0	0	99.8	95.4	0	0	0.8
	Bad	0	0	0	0	0	0	100	0	0.0	77.6	75.8	0
III	Best	47.6	100	99.8	38.6	76.6	35.2	35.6	38.4	70.8	51.6	15.4	32.8
	Good	52.4	0	0.2	61.4	0.2	4.8	64.4	61.6	29.2	48.4	84.6	67.2
	Moderate	0	0	0	0	20.6	55.6	0	0	0	0	0	0
	Bad	0	0	0	0	2.6	4.4	0	0	0	0	0	0
IV	Best	100	100	100	86	99.8	100	0	100	100	0	0	100
	Bad	0	0	0	14	0.2	0	100	0	0	100	100	0
Overall	Best	69.4	99.9	99.8	56.2	69.1	76.6	33.9	59.7	68.9	43.5	34.9	83.0
	Good	30.4	0.1	0.2	40.3	24.9	2.0	16.1	40.3	31.1	12.1	21.2	17.0
	Moderate	0.0	0.0	0.0	0.0	5.2	13.9	0.0	0.0	0.0	0	0.0	0.0
	Bad	0.0	0.0	0.0	3.5	0.8	7.5	50.0	0.0	0.0	44.4	43.9	0.0

Table 6.4: Table of percentage times various variable selection methods select various types of subsets of variables for simulated data of Models *I*–*IV*.

methods using h statistics attain 100% accuracy on these tests.

Moving on to the second model we can begin to observe greater differences in the performance of the different techniques. McCabe's methods **M2** and **M3** continue to attain a high level of accuracy, whereas the method **M1** based on the determinant suffers a notable drop in performance, though it still never selects a "Bad" subset. Jolliffe's methods **B1** and **B2** predominantly return only "Good" subsets rather than "Best", and the previously unreliable **B4** method outperforms both on this model. The multiple correlation method **A1** appears to suffer a major failure by exclusively returning "Bad" subsets. Further investigation reveals that it excludes the independent variable X_3 from all returned subsets due to its negligible predictive power, thus resulting in a 100% failure rate. Curiously, the purportedly inferior **A2** appears to be less susceptible to these problems. The **DF** method suffers a similar fate, often preferring multiple variables from a group thus explaining its poor performance. **KP** never makes a bad selection with this model, though it seems to prefer the "Good" subsets to the "Best". As expected, the correlation-based stepwise procedure **HC** suffers from a similar problem as **A1** by forcing out the independent variable giving it a low performance, though still performs better than **A1**. The simple selection process **H** has a comparable performance to McCabe's optimal solutions.

The third model appears to prove more of a challenge to many of the methods. Of McCabe's solutions, **M2** and **M4** retain their high level of performance, and **M1** is again performing slightly less well. However, one of the problems in interpreting the results from this model is the blurring between the relative merits of "Good" and "Best" subsets for this model. Whilst "Best" subsets are theoretically the superior, in practical terms the "Good" subsets could be equally valuable. Relaxing the classification criteria would result in a similar situation as in Model I, and so should probably be avoided. Nonetheless, **M1** still never makes an incorrect selection of subset. The exhaustive method **DF** also fares quite well on data from this model as it never makes a bad selection, but its performance is not as good as **M2** or **M3**. The fortunes of the principal components methods are mixed, with **B1** being the best and the other two frequently selecting "Moderate" or "Bad" subsets indicating a

poor consistency of performance. Both the multiple correlation methods performed well on this data set. Of the stepwise methods, **KP** performs the best with 70.8% of subsets being "Best" and none being "Moderate" or worse. Method **HC** performs quite well on these data, though it is still inferior to **H**. However, the performance of **H** on this model is disappointing when compared with its excellent performance elsewhere.

Model *IV* is the largest of the models having ten variables in total, only four of which are key signal variables. All methods except for **HC**, **B1**, **DF** and **A1** attain or are close to a 100% success rate. Method **B1** performs reasonably well, but is not as successful as those previously mentioned. However **A1**, **DF** and **HC** again systematically ignore the single independent variable causing a 100% failure rate.

In conclusion, McCabe's methods **M2** and **M3** are the best - they consistently return a high proportion of "Best" subsets. However, they are exhaustive in nature and hence require significant computation to enumerate and evaluate all subsets. Whilst simple for small toy problems such as this, trials seeking the four McCabe-optimal subsets of size 7 in a 20-variable problem took approximately one hour to simultaneously evaluate. Nonetheless, method **M1** appeared to be the inferior sibling in this family of solutions; whilst still never being incorrect in its selections many of its choices were inferior to those made by **M2** and **M3**. The principal components methods had mixed fortunes - method **B1** shows a poorer overall performance to **B2** due to its failures under Model *IV* and relatively poor performance on Model *III*. This is unexpected as **B1** repeats the principal component analysis at each stage, which is assumed preferable to performing it only once as in **B2**. However, it is likely the case that re-performing the PCA on the reduced subset of variables is not sufficient to accommodate for the removal of each variable. Performing the PCA on the partial variance/correlation matrix may serve to boost the performance here. Both methods however retain subsets that are "Good" or better for more than 90% of the simulations. Method **B4**, the forward selection method, is slightly more inconsistent than **B1** or **B2** with a higher probability of obtaining either a "Best" or "Bad" subset.

Of the multiple correlation methods, the stepwise method **A2** was the better

of the two. The reason for this disparity being that **A1** consistently ignores single independent variables due to their low predictive power and negligible impact on the multiple correlation. The stepwise method seems to dodge this problem, with a generally good performance and since it never selects a subset that was "Moderate" or "Bad" it also demonstrates high consistency. The **KP** method is conceived to determine the subset that best represents the multivariate structure of all the variables, consequently its performance on these tests is quite good with no "Bad" or "Moderate" subsets and 69% of all subsets being "Best" - rather better indeed than originally reported [75]. It is also the best stepwise method for Model *III*.

Of the two novel methods proposed in this chapter, we can see that the simple stepwise procedure (**HC**) based on h statistics performs poorly. This was due to the fact that the method ignores independent variables in much the same way as **A1**, though with slightly better performance. Nonetheless it remains the second poorest selection method in the study. The unscaled method **H** is quite different however, producing a high frequency of "Best" subsets (83%) and never returning a subset that was "Bad" or even "Moderate". This demonstrates both a high level of accuracy and consistency which surpasses that even of McCabe's optimal solution **M1**, though its performance is still not on a par with the other methods **M2** and **M3**. Only its performance on Model *III* was disappointing, with a lower than average number of "Best" subsets. That said, **H** is a stepwise procedure and not exhaustive and so has consequent benefits in terms of computation and speed of execution that are lost when using McCabe's exhaustive methods. McCabe's methods aside, **H** is the best stepwise variable selection method of those studied.

6.7.1.2 Structured Models

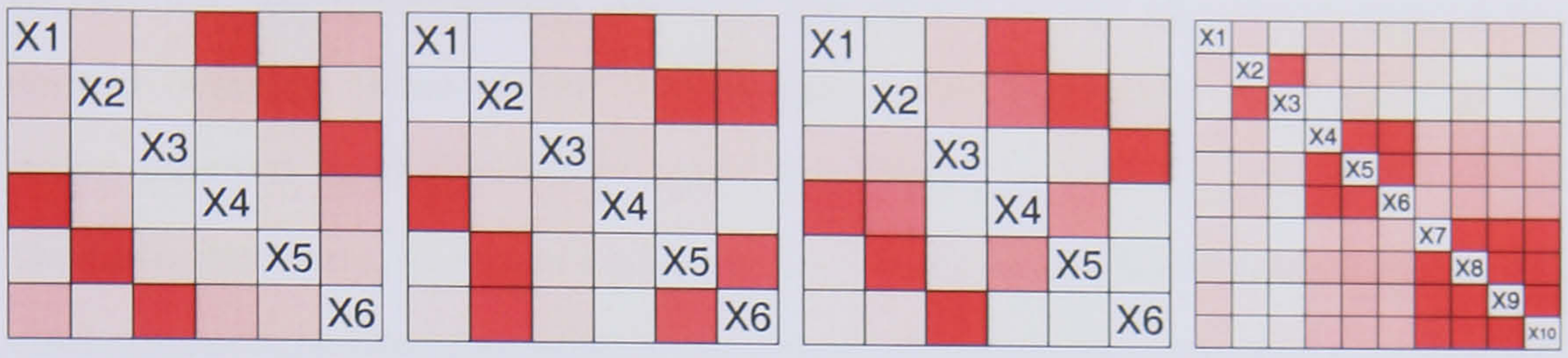
The goal of Krzanowski's Procrustes method for variable selection is to determine a subset which preserves the original structural features of the data. To test this method he performed a modified version of the simulation performed above, arguing that that simulation contained "no inherent structure" with each data set being simply a set of points scattered about zero. Additionally, the selected variables are not tested to see if they preserved the structure of the data, and the classification of

subsets does not reflect a pursuit of such an objective. To achieve this he built additional structure into the data and examined whether the selected subsets contained variables which conveyed this structure. To assess the effectiveness and performance of the selection method proposed in this chapter on such data, the simple selection method **H** will be tested in a simulation following Krzanowski's design.

The data were generated via the models used in the above simulation and were then subsequently modified. Groupings of cases in these data were created by including additional structure into the data via a 3^2 factorial method. The first factor governed the *type* of structure to be created within the data and took three levels: 'single outlier', 'weak groups', or 'strong groups'. The second factor controlled the *amount* of structure present, i.e. the number of variables in which the additional structure was to be found. This factor again had three levels: 'in one variable', 'in two variables' or 'in three variables'.

In the case of structure of type 'single outlier', the value 10 was added to each of the first j variables of the first case in the data, where j was the corresponding level of the 'amount of structure' factor, i.e. 1, 2 or 3. For 'weak groups', the first 25 sample members were unchanged, the next 25 had the value 2 added to the first j variables, the following 25 had the value 4 added to the first j variables and the final 25 had the value 6 added to those j variables. For 'strong groups' the procedure was the same as for 'weak groups' except multiples of 10 were used in place of multiples of 2. Variable selection was then carried out on these modified data, selecting subsets of the appropriate dimension. Each subset was then examined to determine how many of the j structure-bearing variables were present. Running this simulation 100 times for each combination of the type of structure, amount of structure and model type using the unweighted selection method (**H**) generated the results presented in Table 6.5. Krzanowski's reported results are presented alongside for comparison. The **HC** procedure was not tested here due to the obvious shortcomings highlighted in the previous simulation study.

The reasoning behind performing this simulation with structured data is to ascertain whether the 'structure' inserted into several variables within the data is preserved in the reduced subset by returning some of these structure-laden variables.



(a) Model I (b) Model II (c) Model III (d) Model IV

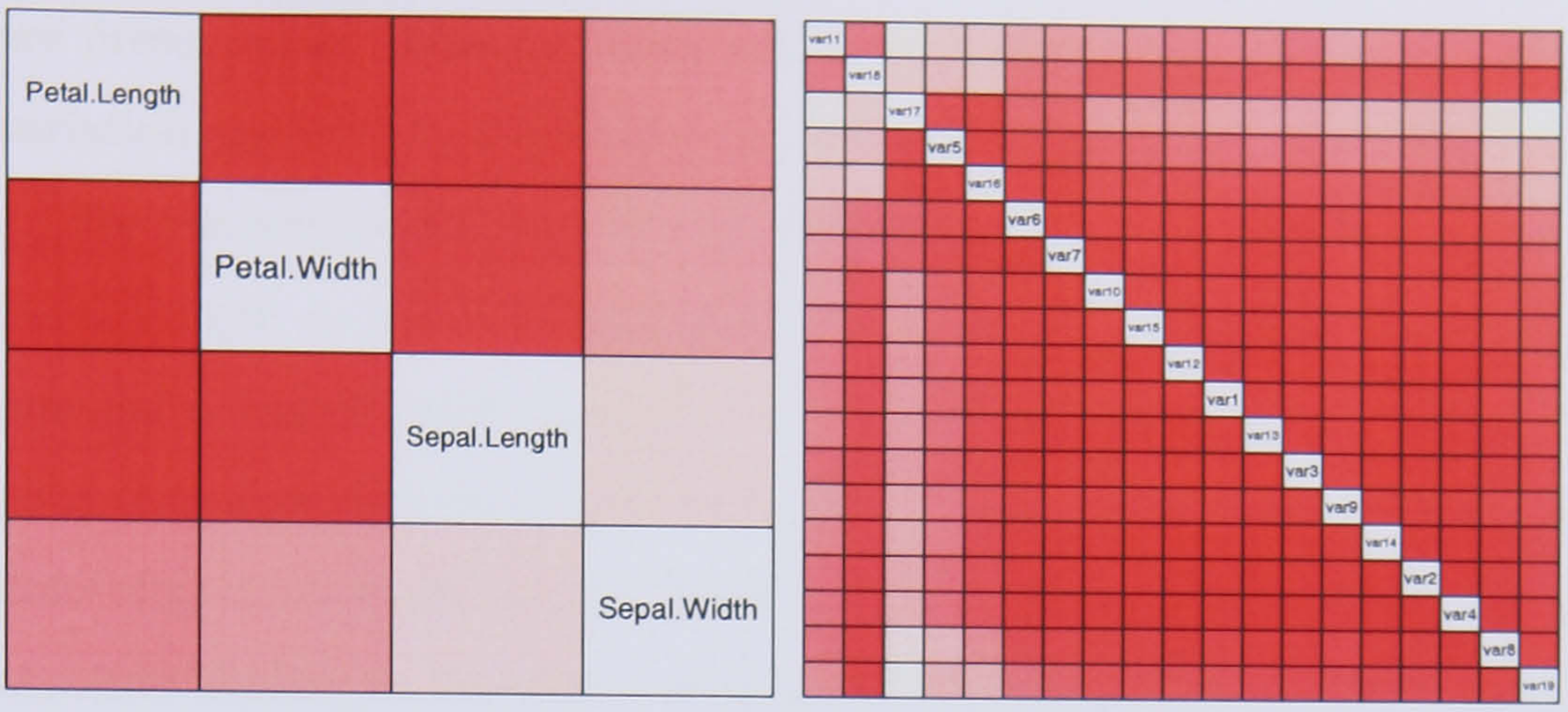
Figure 6.6: Correlation plots for simulated data for Models I–IV.

Structure type	Number of variables exhibiting structure	Number of structure variables selected	Model								Overall	
			I		II		III		IV			
			H	**KP**	**H**	**KP**	**H**	**KP**	**H**	**KP**	**H**	**KP**
Single Outlier	1	0	47	0	48	0	2	9	0	0	24.25	2.25
		1	53	100	52	100	98	91	100	100	75.75	97.75
	2	0	0	0	0	0	3	2	0	0	0.75	0.5
		1	99	6	98	15	83	75	55	97	83.75	48.25
		2	1	94	2	85	14	23	45	3	15.5	51.25
	3	0	0	0	0	0	0	0	0	0	0	0
		1	98	6	44	20	78	6	0	0	55.0	8
		2	2	84	49	72	21	84	100	100	43.0	85
		3	0	10	7	8	1	10	0	0	2.0	7
Weak Groups	1	0	46	0	48	0	0	0	0	0	23.5	0
		1	54	100	52	100	100	100	100	100	76.5	100
	2	0	0	0	0	0	0	0	0	0	0	0
		1	100	0	100	0	100	0	100	0	100	0
		2	0	100	0	100	0	100	0	100	0	100
	3	0	0	0	0	0	0	0	0	0	0	0
		1	100	0	100	0	100	0	93	0	98.25	0
		2	0	1	0	0	0	11	7	100	1.75	28
		3	0	99	0	100	0	89	0	0	0	72
Strong Groups	1	0	42	0	32	0	0	0	0	0	18.5	0
		1	58	100	68	100	100	100	100	100	81.5	100
	2	0	0	0	0	0	0	0	0	0	0	0
		1	100	0	100	0	100	0	100	0	100	0
		2	0	100	0	100	0	100	0	100	0	100
	3	0	0	0	0	0	0	0	0	0	0	0
		1	100	0	100	0	100	0	100	0	100	0
		2	0	0	0	0	0	0	0	0	0	0
		3	0	100	0	100	0	100	0	100	0	100

Table 6.5: Table of number of times a structure-bearing variable is selected for each model with additional structure as used by Krzanowski.

In essence, we are assessing whether the effects of the structure-bearing variables should override those of the underlying model type with the structure variables being selected in preference to other variables. As the strength of the structure in the data increases, it becomes increasingly more likely that Krzanowski's method includes structure-bearing variables in the subset selected. In fact, for weak and strong groups almost all structure variables are selected. However, for the **H** method we find that for both weak and strong groups on two or three variables typically only one of these structure-bearing variables is selected. This is likely since the structure variables form a tightly correlated group with strong inter-correlations. Therefore, once one variable of this group has been selected then the partial variance and hence the h values of the other variables in that group will dramatically fall thereby reducing their desirability to the selection algorithm. Essentially, the effects of adding this structure to the data is equivalent to there being an underlying latent variable representing this structure and the variables themselves being noisy realisations of this variable. It is therefore the case that the procedure detects the addition of the multiples of two and ten as being the underlying signal thus leaving the other structure variables effectively redundant, and reducing their desirability for selection once one of the group has been selected. This behaviour of selecting only one structure variable does, in the case of **H**, override that of the underlying model as this would typically require at least two of these variables to be present to be classed as a 'Best' subset.

The results for **H** on data with a single outlier are more variable than with the other structure types. The subsets selected for this type of structure appear governed by the underlying models rather than the structure. However, the effect of inducing the outlier structure has been to force the procedure to favour the unaffected partners of the structure-bearing variables before the structure was added. This is because the addition of the outlier reduces its correlations with other variables, making its equivalent(s) prior to the addition of structure the more likely choice with higher correlations. It should be noted that the selection of these equivalent variables still yields a 'Best' classification under Jolliffe's models. However it is clear that the procedure sees variables with a single outlier as being less desirable candidates for



(a) Correlation plot of the iris data (b) Correlation plot of the aphids data

Figure 6.7: Correlation plots for the correlation matrices of the iris and aphids data sets.

inclusion in the subset when compared with its performance on the base models. This seems to be the opposite to the behaviour of Krzanowski's method which is evidently highly sensitive to such outliers. It interprets this outlier behaviour as a necessary component of the underlying structure. Similar behaviour is observed for most structure types on single variables.

6.7.2 Real Data

6.7.2.1 Iris Data

Whilst evaluation of the presented methodology on simulated data is informative, it is the analysis of real data that is the focus of most statistical work. The first real data set considered and presented here is the famous (Fisher's or Anderson's) iris data set [3, 47]. The data consist of four size measurements on 150 samples of three species of iris - the correlation plot of the correlation matrix for these data is displayed in Figure 6.7. Whilst only being four-dimensional, this data set is not a realistic candidate for variable selection, however it is such a widely used example data set that familiarity with it is likely. Therefore, a number of the previously tested variable selection methods including the unweighted stepwise method **H** were applied to the data to assess their performance on real data and to draw comparisons between them. The results from selecting one to three variables from this data set

are presented in Table 6.6 along with the corresponding values for the percentage variation captured by that subset as determined by $\text{tr}(\widetilde{\boldsymbol{S}}_{22\cdot 1})$ and $||\widetilde{\boldsymbol{S}}_{22\cdot 1}||^2$ per (6.23).

For the single-variable selection problem, we can see that most methods choose Petal Length as the most representative variable. Only two of the principal components methods differ from this choice with **B1** picking Sepal Width, which is a poor choice as evidenced by low resulting value for the percentage variance. Paradoxically, method **B4** selects Petal Length as being the variable with the largest absolute loading in the first principal component and method **B2** rejects the same variable at the first step for having the largest absolute loading in the last principal component. In addition, it is worth noting that one of the eigenvalues in the data is almost zero implying that one of the variables is redundant.

For the selection of two variables we can see that all methods now include Sepal Width as a key variable, however the choice of a second variable varies among the different techniques. McCabe's methods along with **B4**, **A2**, **DF** and **H** all return Petal Length as their chosen second variable. This agreement between these methods would corroborate the idea the **H** is a stepwise attempt at finding the optimal solution described by **M3**. Jolliffe's method **B1** and Krzanowski's **KP** prefer to retain both width measures with a slightly lower performance. Only method **B1** chose to retain both sepal measurements and it ranked last in terms of $\text{tr}(\boldsymbol{\Sigma})$ and $||\boldsymbol{\Sigma}||^2$. Interestingly, no method chose both the petal measurements despite being highly correlated with one another (0.82) and moderately correlated to the sepal measurements. The strength of this pairing could suggest that one variable could be used as a feasible surrogate for the other - possibly the redundancy indicated by the zero eigenvalue. In fact, once the effects of one of these variables are accounted for the partial correlations of the other variable with the sepal measurements drop dramatically, which indicates a likely overlap in the information contained in these variables.

For the 3-variable case, all methods choose both petal variables and it is the differing choice for the third variable that the variation among the methods is visible. The best subset in terms of $\text{tr}(\boldsymbol{\Sigma})$ and $||\boldsymbol{\Sigma}||^2$ is that chosen by McCabe's methods, **B1**, **B2** and **KP** which prefer Petal Width over Petal Length. All of these return

Number of Variables	Selection Method	Sepal Length	Sepal Width	Petal Length	Petal Width	% tr($\boldsymbol{R}$)	% $\mid \boldsymbol{R} \mid^2$
1	**M2**			x		71.8	90.8
	M3			x		71.8	90.8
	B1		x			61.1	82.7
	B2				x	68.3	89.8
	B4			x		71.8	90.8
	A2			x		71.8	90.8
	KP			x		71.8	90.8
	DF			x		71.8	90.8
	H			x		71.8	90.8
2	**M2**		x	x		94.2	99.6
	M3		x	x		94.2	99.6
	B1	x	x			90.3	98.6
	B2		x		x	91.0	98.8
	B4		x	x		94.2	99.6
	A2		x	x		94.2	99.6
	KP		x		x	91.0	98.8
	DF		x	x	x	94.2	99.6
	H		x	x		94.2	99.6
3	**M2**	x	x		x	99.2	100.0
	M3	x	x		x	99.2	100.0
	B1	x	x		x	99.2	100.0
	B2	x	x		x	99.2	100.0
	B4	x	x	x		98.4	100.0
	A2		x	x	x	96.5	99.8
	KP	x	x		x	99.2	100.0
	DF	x	x		x	99.2	100.0
	H	x	x	x		98.4	100.0

Table 6.6: Selected variables for Fisher's Iris data using various selection methods.

subsets with a retained proportion of $||\mathbf{\Sigma}||^2$ at almost 100%. However, the other two methods (**B4** and **H**) choose Petal Length and have a slightly lower resulting value of the percentage trace. However, the difference is negligible reflecting the fact that the majority of the information was captured on the two variables previously selected. Method **A2** performed relatively poorly with the three-variable problem, its choice of both petal measurements resulted in the lowest values for percentage variation. One point worthy of note is that for all the stepwise methods, we can observe that the 'optimal' 2-variable subset is contained within the corresponding 3-variable group which is not required to be the case for McCabe's methods as these require exhaustive evaluation. In fact, these methods have returned subsets with only a small amount of overlap.

6.7.2.2 Aphids Data

The second real data set to be examined consists of 19 variables measured on 40 *alate adelges* (winged aphids). These data were first examined by Jeffers [65] and have often been subsequently examined in a variable reduction setting [75, 68, 67]. Krzanowski determined via a cross-validatory approach that the minimum number of variables required to adequately describe the data was four. Using this information, the various selection methodologies have been applied and the resultant four-variable subsets are given in Table 6.7 with their associated value of the percentage variance in terms of $\text{tr}(\mathbf{\Sigma})$ and $||\mathbf{\Sigma}||^2$.

We can see immediately that all variables select variable 5 in their reduced subset, and all but **KP** additionally choose variable 11. However, from these subsets we can see that McCabe's methods return the same subsets, as do **B4** and **H**, with the others all returning different groups. Overall, there is little practical difference in the subsets returned by the different methods with each returning a subset that represents 80–90% of the variation. In these terms, we can see that the McCabe solutions are the best with a value of 89.8%. Both **H** and **B4** come a close second with 89.1% - which is close for the 'optimal' exhaustive methods and the sub-optimal stepwise methods. This is particularly meaningful when considering that this is a larger problem than the iris data creating a greater consequent potential for a diver-

Selection Method	Variable												% tr($\boldsymbol{\Sigma}$)	% $\|\boldsymbol{\Sigma}\|^2$
	2	5	6	8	9	11	12	13	14	17	18	19		
M2		×				×		×			×		89.8	99.8
M3		×				×		×			×		89.8	99.8
B1		×			×	×						×	81.7	98.1
B2	×	×		×		×							85.5	99.3
B4		×				×		×		×			89.1	99.7
A2		×	×			×				×			78.2	96.6
KP		×					×		×		×		86.3	99.5
H		×				×		×		×			89.1	99.7

Table 6.7: Selected four-variable subsets for Jeffer's aphid data using various selection methods.

gence between the stepwise and exhaustive methods. This is an encouraging result for **H**. The other methods are slightly inferior in performance, though Krzanowski's method is still returning a good subset. **B1** and **A2** both make poor choices resulting in the two lowest value for the percentage trace and percentage squared norm.

Plotting each of the four possible 'scree' plots from Section 6.5.4 for the aphids data analysed above, we obtain the plots in Figure 6.8. Figure 6.8(a) shows the h scores for each of the variables as it was selected by the **H** procedure. We can see that these weighted scores decrease rapidly, and most of the variables after the third or fourth selected are approximately zero. The exact values of the first four h statistics are 13.3, 1.8, 0.7, and 0.3 which highlights the rapid reduction in magnitude due to the fact that the variables are all moderately to strongly correlated and once the key variables are selected the partial variances all but disappear. A red dashed line is drawn on the plot at $h = 1$ since a variable with an h value below this level conveys less information than a single independent variable. Therefore, in Section 6.6 this was proposed as a threshold for the assessment of the intrinsic dimensionality of the data. In this case it would suggest that the effective dimensionality is 2. This does not agree with Krzanowski's results [75] for determining the minimum number of variables that can adequately represent the data. This could suggest that a threshold

value of 1 is too conservative and a reduced value may be more appropriate, or it may indicate that h values may not be suitable for a direct estimation of intrinsic dimensionality.

The plot in Figure 6.8(b) of the squared norm of the correlation matrix of remaining variables is again decreasing as one would expect and in a similar fashion to the h scores in Figure 6.8(a). However the rate at which the values decrease here is slower and is likely more representative of the true situation. The value for $||\Sigma_{22\cdot 1}||^2$ falls from 7.44 to 1.92, 1.15 and then 0.57 over the first four variables. This reduction is due to the fact that with each variable we select, we are capturing more of the variability in the data. We can see that after the fifth variable the points are all close to zero and are approximately linear suggesting that they represent little more than noise.

The plot of the percentage variability in terms of the trace in Figure 6.8(c) increases slowly, likely due to the fact we are not seeking to optimise the trace of the conditional variance. This plot is curved as predicted and shows the linear trend after the first four or five variables and so would agree with Krzanowski. It also illustrates that we capture 80–90% of the variability of the data in those first four variables with the remainder contributing a negligible and near-constant amount. The plot of the percentage variability in terms of the squared norm explained by a given number of extracted variables is displayed in Figure 6.8(d). This plot is simply a reinterpretation of the information in Figure 6.8(b) relative to the squared norm of the initial correlation matrix, which in this case equals 198. This plot shows that even with one variable in excess of 90% of the information is captured. This increase is sudden and gives an angular appearance to the plot, rather than the expected curve but is due to the high correlations between many of the variables in the data set. It also suggests we obtain nearly 100% of the variability with 2 or more variables. Whilst this would be at odds with Krzanowski's dimension assessment of 4, it does agree more with the eigenvalue-based procedures such as scree plots which suggest the majority of the information (73%) is conveyed on the first principal component alone.

Method	Component/Variable					
	1	2	3	4	5	6
SCA	$\{X_1, \ldots, X_{12}\}$	$\{X_{13}, \ldots, X_{16}\}$	$\{X_{19}, \ldots, X_{20}\}$	$\{X_{17}, \ldots, X_{18}\}$	–	–
H	X_8	X_{15}	X_2	X_{18}	X_9	X_{20}

Table 6.8: Block simple components and selected principal variables for the neuromotor data.

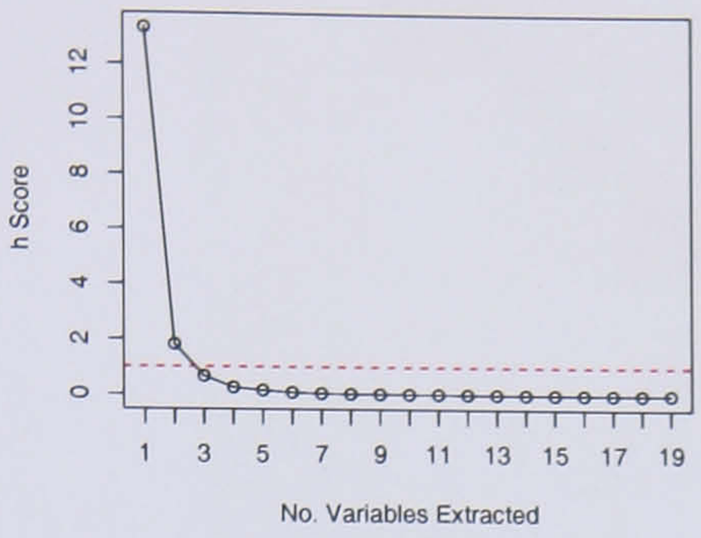
6.7.2.3 Neuromotor Data

The third real data set to be investigated consists of 20 variables measured on 467 children. The data was investigated by Rousson and Gasser [106] and Largo *et al* [78], and concerns the assessment of the development of neuromotor functions in children and adolescents. The correlation plot for these data is shown in Figure 6.9. The first 12 tasks concern fine motor tasks involving the feet $(X_1, \ldots, X_4)$, the hands $(X_5, \ldots, X_8)$, or the fingers $(X_9 \ldots, X_{12})$. These tasks can be separated into pairs where the same task is performed on the dominant and non-dominant side, thus accounting for the strong pairings in the data. Variables $X_{13}, \ldots, X_{16}$ correspond to pegboard tasks, and variables $X_{17}, \ldots, X_{20}$ consist of gross motor tasks.

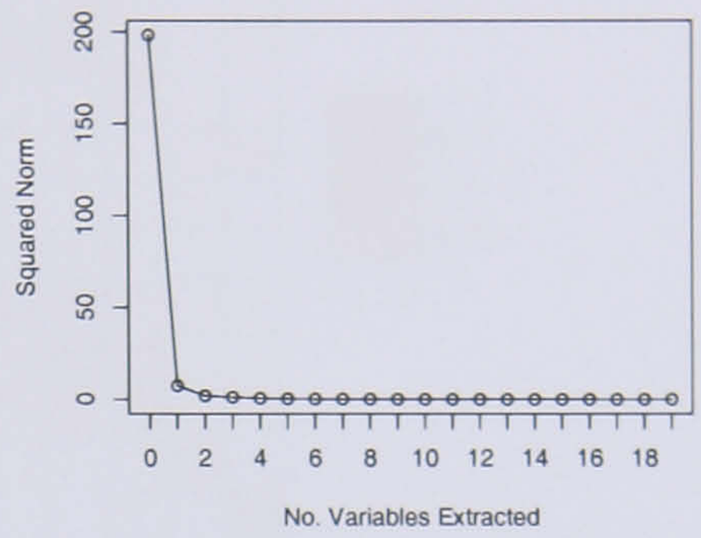
It is clear that these correlations are structured with many pairings of variables, a moderately correlated block of the first 12 variables, a more strongly correlated group for $X_{13}, \ldots, X_{16}$ and two final pairs of variables. This would suggest that dimension or variable reduction would be quite effective here.

These data were analysed in [106] to illustrate the method of Simple Components Analysis, where it was determined that the data could be expressed in terms of 4 block components and 2 difference components. The four block components were composed of variables $X_1, \ldots, X_{12}$, $X_{13}, \ldots, X_{16}$, $X_{19}, \ldots, X_{20}$, and $X_{17}, \ldots, X_{18}$ respectively. The first difference component contrasted the pairs $\{(X_1,X_2), (X_5,X_6), (X_9,X_{10})\}$ with $\{(X_3,X_4), (X_7,X_8), (X_{11},X_{12})\}$; the second difference component contrasted $X_1, \ldots, X_4$ with $X_9, \ldots, X_{12}$.

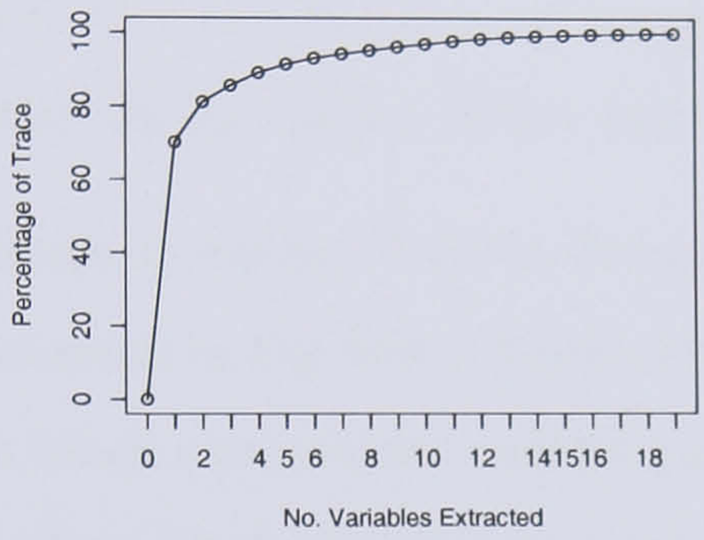
The simple block components will be compared with the results from applying **H** in order to determine whether there is any similarity between the methods. The resulting simple components and principal variables are given in Table 6.8, where



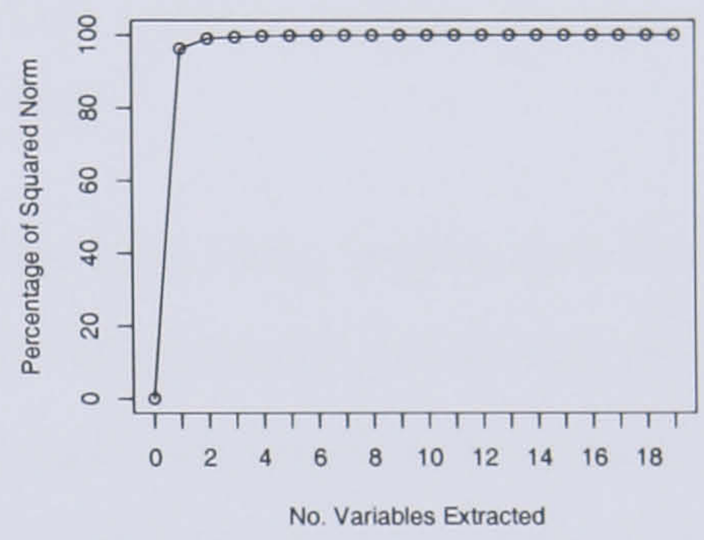
(a) h scores



(b) $||\widetilde{\boldsymbol{S}}_{22\cdot 1}||^2$



(c) Percentage $\mathrm{tr}(\widetilde{\boldsymbol{S}}_{22\cdot 1})$ Explained



(d) Percentage $||\widetilde{\boldsymbol{S}}_{22\cdot 1}||^2$ Explained

Figure 6.8: Four Scree-type plots for the aphids data.

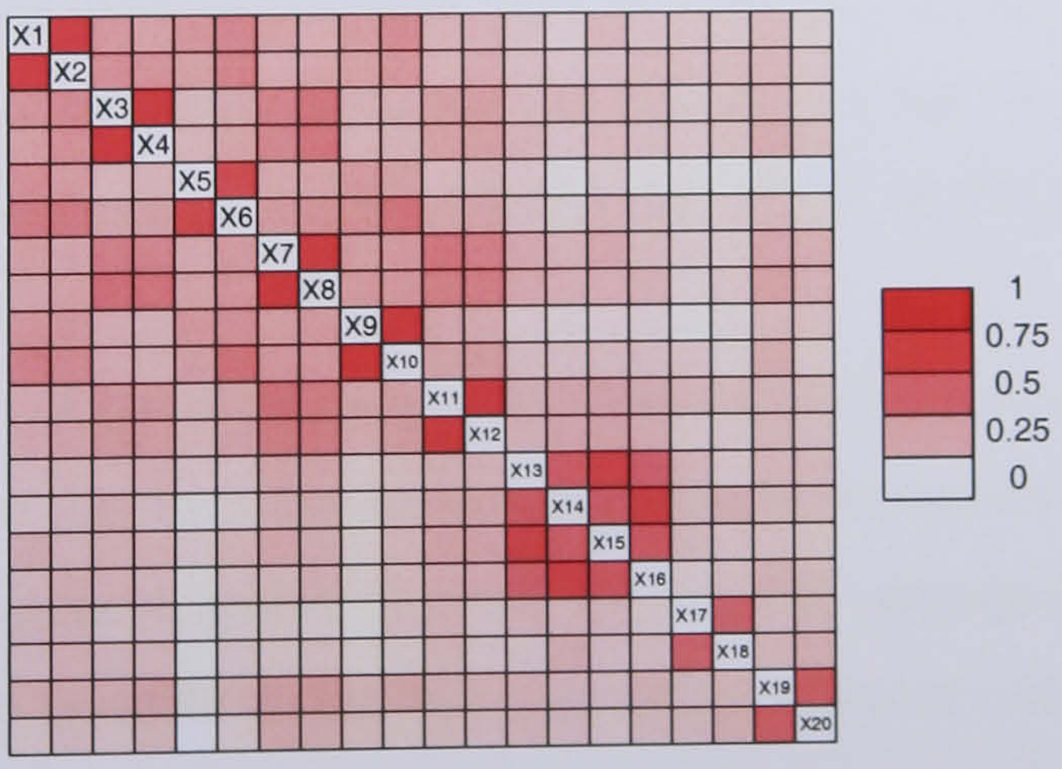


Figure 6.9: Correlation plots for the correlation matrix of the neuromotor data set.

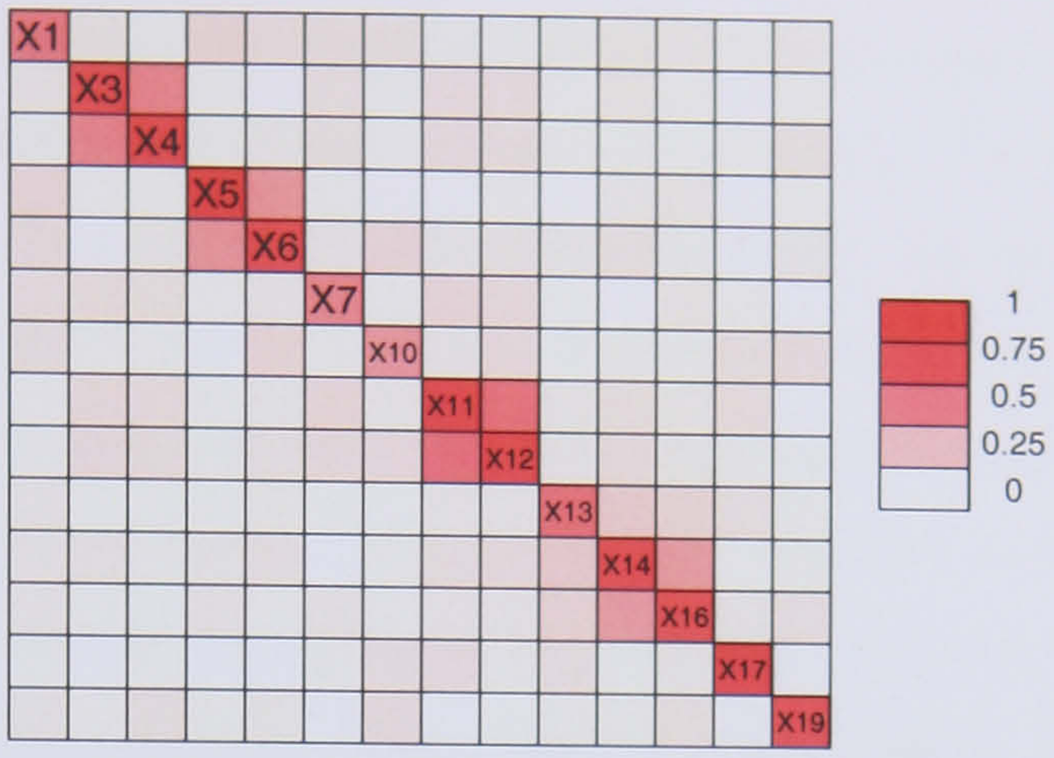


Figure 6.10: Correlation plots of the partial variance matrix for the neuromotor data set after the extraction of the first 6 PVs.

we have chosen to extract 6 PVs. We can see from these results that the first 2 PVs are are contained in the first 2 block simple components, and furthermore the third and fourth block components contain principal variables 6 and 4 respectively. Thus each of the four block simple components is represented by at least one of the first six principal variables.

These first 6 PVs provide 52.7% of $\mathrm{tr}(\boldsymbol{\Sigma})$ and 81.1% of $||\boldsymbol{\Sigma}||^2$, which is quite good performance using only 30% of the variables. Looking at the remaining variation in the data after the extraction of these six PVs in Figure 6.10 we can see that the majority of the off-diagonal correlation has been eliminated and the only remaining residual associations exist within four pairs of variables. In fact, if we increase the number of principal variables to 10 we can eliminate these remaining associations and increase performance to 74.1% of $\mathrm{tr}(\boldsymbol{\Sigma})$ and 93.7% of $||\boldsymbol{\Sigma}||^2$ using only 50% of the data.

The calculation of statistics such as $\mathrm{tr}(\boldsymbol{\Sigma})$ and $||\boldsymbol{\Sigma}||^2$ to compare these two methods is difficult since the raw data are not available so we cannot calculate the values of the simple component scores in order to find the appropriate values of $\boldsymbol{R}_{22\cdot 1}$. However, it will be possible to use the percentage variance criterion introduced by Rousson and Gasser to evaluate both methods. Their criterion is defined as:

$$\mathrm{Opt}_2(P) = \frac{\mathrm{tr}(\boldsymbol{\Sigma}\boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{\Sigma}\boldsymbol{P})^{-1}\boldsymbol{P}^T\boldsymbol{\Sigma})}{\mathrm{tr}(\boldsymbol{\Lambda}_q)},$$

where $\boldsymbol{P}$ is the matrix of loadings for the simple components, $\boldsymbol{\Sigma}$ is the variance

matrix and $\mathbf{\Lambda}_q$ is the $(q \times q)$ diagonal matrix of the q largest eigenvalues. When evaluating this criterion for **H** each column of the matrix P will contain a single 1 in the position of the selected PV and zeros elsewhere. On this statistic the SCA method captures 99.4% of the variation, whereas **H** captures only 34.0%. These results are not surprising and are really only indicative of the number of variables retained by the two methods. SCA uses all 20 variables giving almost 100% of the variance, whereas **H** uses 30% of the variables and thus captures a similar amount of variability. Despite this obviously poor relative performance, the **H** method does appear to mirror the simplicity of the simple components by extracting representative variables of the block components.

Chapter 7

Reducing the Orthopædic Data

7.1 Introduction

Having established in Chapter 5 that having a large number of variables in a graphical model causes some difficult analytic and computational problems, reducing the size of the orthopædic data sets was determined to be necessary. Chapter 6 discussed many existing variable reduction strategies and developed some novel methods and extensions that accommodate both longitudinal data and utilities for individual variables. The focus of this chapter is to apply these variable reduction techniques to the repeated measurements within the orthopædic data sets, and then to analyse and discuss the results.

This chapter is divided into two main sections looking at reducing the size of the orthopædic data sets. Section 7.2 considers reducing the size of the knees data set, and Section 7.3 moves on to consider the hips data. Restricting attention to the pre-operative data for brevity, each section addresses the issue of the intrinsic dimensionality of the data via a principal component analysis and the application of several methods previously discussed. Using that information an appropriate number of variables are chosen by applying several of the dimension reduction techniques existing in the literature. Then the stepwise selection methods proposed in Chapter 6 are applied to the data and compared to the results from the other methods. Finally, a subset of variables are selected over all time points in the data by using the temporal extension to the variable selection process proposed in Section

6.5. Subsets are also obtained for the knees data using information from subjective variable utilities.

7.2 The Knees Data

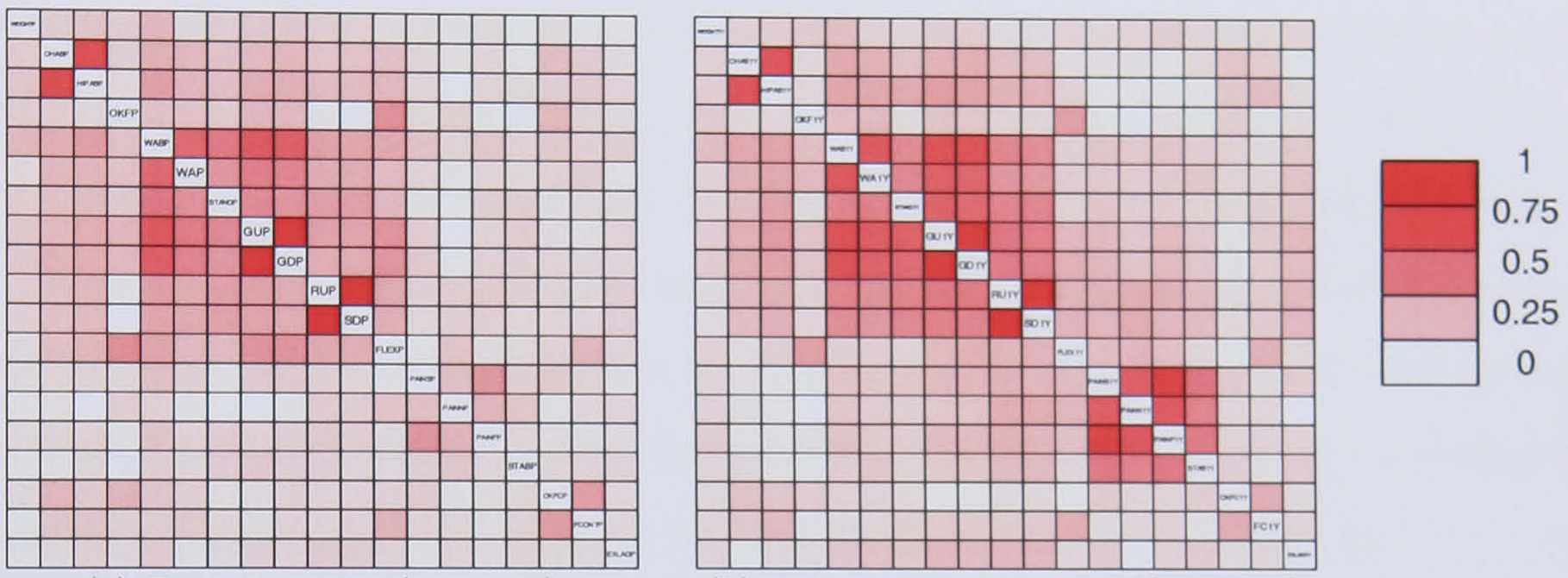
7.2.1 Data Structure

The knees data were fully discussed in Section 2.2.1 and contain a total of 20 repeated measurement variables representing the patient's status, each recorded at four time points. The times at which data is recorded are pre-operatively and at 1, 5 and 10-year follow-up consultations. Not all patients remained in the study up to the 10-year point, so there is a consequent drop in the sample sizes at each point.

To gain insight into the possible structure of the data, correlation plots for the data at each time point are presented in Figure 7.1. One variable (*Coronal Tibio-Femoral Angle*) exhibited a zero variance in the 5 and 10-year data and so has been removed from those plots. Looking at these plots we can see that the correlation structure is quite similar over the four time points, though it becomes noisier at the later time points due to the reduction in the sample size. Nonetheless, this suggests a degree of constancy to the correlation structure between these measurements which persists despite the passage of time.

It is immediately clear from the plots that there is a central block of correlated variables. This block corresponds to the walking ability scores with the exception of *Stability*, which exhibits little association with these scores and is more closely related to pain measures. The average correlation in this group is approximately 0.43 pre-operatively, with correlations increasing at later times. It also appears that *Walking Ability* is associated to *Hip Abduction* and *Other Hip Abduction* as shown by the horizontal and vertical bars above and to the left of the central block, which are most visible in the 5-year plot. Within this larger group however, we can see that there are two pairs of variables which are particularly strongly associated with $r > 0.9$ in both cases. These are, first, *Sitting Down* and *Rising Up* (labelled SD and RU) and then *Going Up Stairs* and *Going Down Stairs* (labelled GU and GD).

We can also see a second block of correlated variables towards the bottom right



(a) Pre-operative ($n = 599$) (b) 1-year post-operative ($n = 559$)

(c) 5-year post-operative ($n = 239$) (d) 10-years post-operative ($n = 86$)

Figure 7.1: Correlation plots of the repeated measurements in the knees data observed at each of the four time points.

of the three post-operative plots. This block corresponds to the three pain scores plus *Stability*. Whilst the associations within this group are fairly weak and not immediately noticeable pre-operatively ($r \simeq 0.28$), they strengthen post-operatively ($r \simeq 0.68$). The correlation between the pair *Pain Frequency* and *Pain Severity* is especially strong at 0.83.

In addition to these two groups of variables, there are a number of pairings among the remaining variables that are highlighted by the plots. For example, *Other Hip Abduction* (labelled OHAB in the upper left corner) is noticeably associated to *Hip Abduction*. There are also similar, though weaker, pairings between both *Flexion* and *Other Knee Flexion* and *Fixed Contracture* and *Other Knee Fixed Contracture*.

Thus we can see that there is significant structure underlying these knees data. Furthermore, the presence of groups of correlated variables as well as pairs of strongly associated variables would suggest the possibility of redundancy among the observed measurements making the data an ideal candidate for variable selection.

7.2.2 Principal Component Analysis and Dimension Assessment

The use of principal component analysis is common when seeking to reduce dimensionality. Whilst not directly applicable to variable reduction, it is an effective tool for dimension reduction and is used in several methods for ascertaining the intrinsic dimensionality of the data. Therefore the results of the principal component analysis are discussed here, before moving on to the assessment of dimensionality.

The principal components of the pre-operative knees data were calculated via the singular value decomposition of the correlation matrix. The loadings for the first five principal components are given in Table 7.1 and their associated standard deviations in Table 7.2. The first principal component accounts for 22% of the variation of the data and is essentially an average of the patients walking ability scores combined with *Flexion* and both *Hip Abductions*. This component would appear to correspond with the large central group of variables shown in the correlation mosaics in Figure 7.1. The second component represents only 9% of the total variation and conveys information on *Sitting Down* and *Rising Up* as well as several of the anatomical

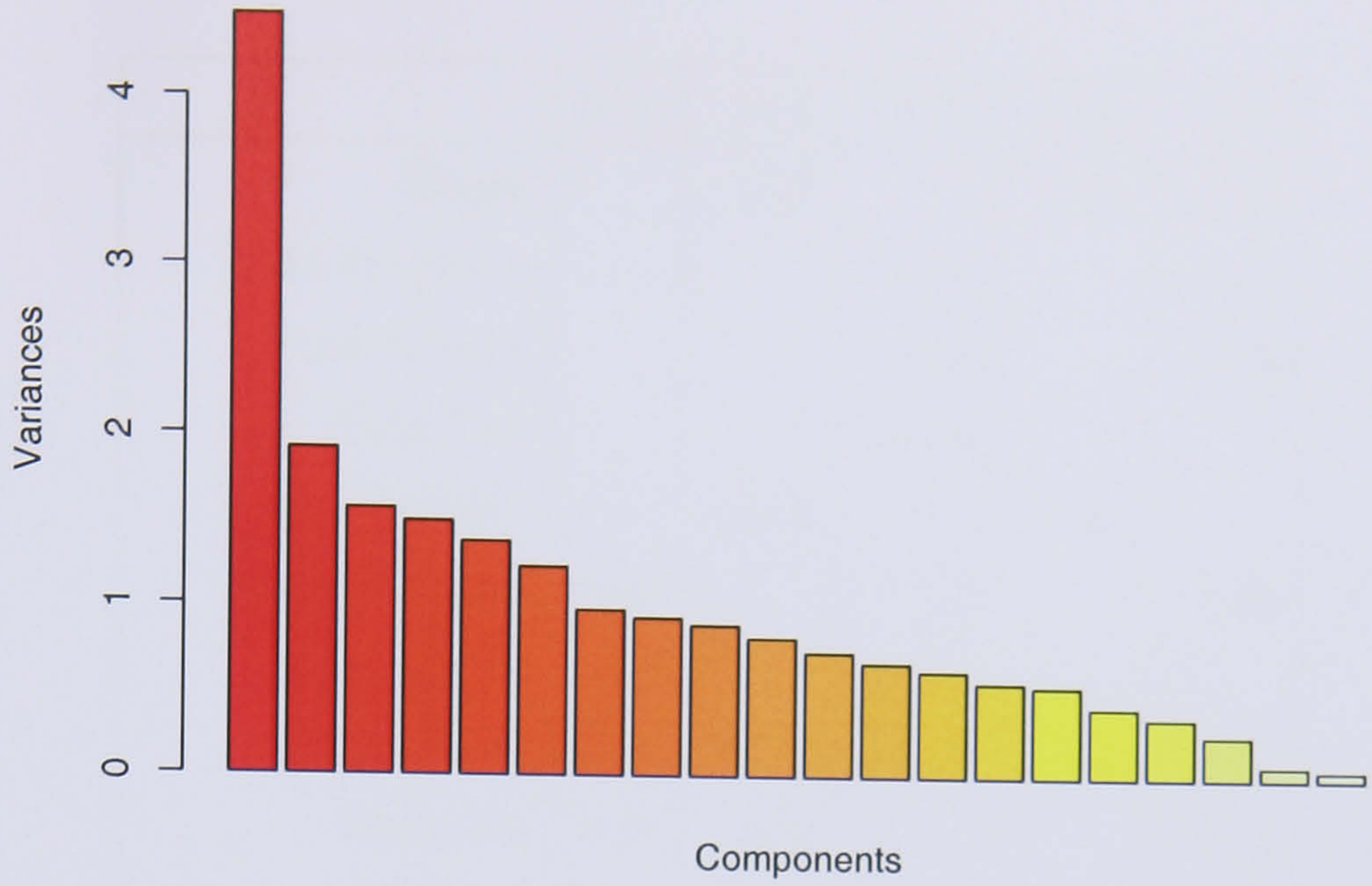


Figure 7.2: Scree plot for the principal components of the pre-operative knees data.

angles, with the exception of the hip measures. It also noticeably contrasts *Fixed Contracture* with *Other Knee Fixed Contracture*. The third component encapsulates the variation in the three pain scores suggesting they commonly vary in step with one another. The fourth component governs the level of *Hip Abduction* on both hips.

Having calculated the principal components we can use the results to assess the dimensionality of the pre-operative knees data set. The methods used are those discussed in Section 6.2.3 with the exception of Eastment and Krzanowski's cross-validatory method. The reason for this exclusion is that it required systematically predicting each element of the data matrix after its row and column have been deleted by using different numbers of the principal components which are calculated on this reduced data set. This would require 11980 calculations of the principal components, which was deemed to be excessive.

Kaiser's rule using eigenvalue thresholds of 1 and 0.7 gave dimensions of 7 and 16 respectively. By further examining the eigenvalues, we learn that to capture a total of 75% of the variation of the data would require 10 of the principal components. The scree plot of the variances of the principal components (λ^2) is shown in Figure 7.2 which illustrates that the first component accounts for a large share of the variance,

	PC1	PC2	PC3	PC4	PC5
Weight		0.17		0.23	−0.28
Pain Frequency			0.54		−0.19
Pain Severity			0.55		−0.24
Night Pain			0.48		
Stability		0.16		−0.17	
Walking Ability	0.36				−0.21
Walking Aids	0.29				
Sitting Down	0.29	0.34			0.43
Rising Up	0.29	0.35			0.41
Standing	0.27				
Going Up Stairs	0.37		−0.18	−0.24	−0.23
Going Down Stairs	0.36			−0.21	−0.23
Coronal T-F Angle		0.36		0.20	
Fixed Contracture		0.32	−0.17	−0.16	−0.32
Flexion	0.26				0.33
Extension Lag		−0.20			
Hip Abduction	0.26			0.54	
OK Fixed Cont.		−0.48	−0.18		
OK Flexion		−0.34		−0.28	0.21
Oth. Hip Abduction	0.23	−0.19		0.55	

Table 7.1: The first five principal components of the pre-operative knees data. Loadings of value < 0.15 have been omitted for clarity.

	PC1	PC2	PC3	PC4	PC5
Standard Deviation	2.1182	1.3863	1.2536	1.2238	1.1745
Proportion of Variance	0.2243	0.0961	0.0786	0.0749	0.0690
Cumulative Proportion	0.2243	0.3204	0.3990	0.4739	0.5429

Table 7.2: The importance of the first five principal components of the pre-operative knees data.

Method	Dimensionality
Kaiser's rule $\lambda < 1$	7
Kaiser's rule $\lambda < 0.7$	16
75% of variation	10
Scree Plot	6
Broken Stick	7
Velicer	13

Table 7.3: Table of estimates for the intrinsic dimensionality of the pre-operative knees data.

but also shows a linear trend starting at the 6th component. Using the 'broken stick' method, the proportion of variation attributed to each component is compared to its expected value which suggests an intrinsic dimensionality of 7. Applying Velicer's method of finding the number of principal components which minimises the average partial correlation given those components suggests a value of 13. These results are summarised in Table 7.3.

The minimum number of dimensions in terms of principal components appears to be 7. However, this may be an oversimplification as several of the other methods suggest somewhat larger values, and the $\lambda < 0.7$ method suggests a value more than double that value. Nonetheless, it is not surprising that the value for the dimensionality of the data is this large relative to the number of variables since the knees data are quite sparsely correlated with small groups exhibiting moderate correlations within these groups and only minimal correlations between them. Additionally, these estimates are only considering the number of principal components, in practice the number of variables will be larger than this.

7.2.3 Reducing the Pre-operative Data

Whilst the knees data consist of four separate time points, most variable selection procedures can only operate on data from one time point. Therefore, the pre-operative data is first considered as an extended example and illustration of the application of the variable selection procedure. The first question to be answered

is how many of the original variables should be retained in order to maintain an adequate description of the original data. The estimates obtained above are far from unanimous and range from 7 to 16. On consideration, we shall favour a cut-off point for the proportion of variation expressed by the variables in terms of the trace or squared norm of the partial variance matrix of the remaining variables. This is natural as the variable selection procedures proposed in the previous chapter operate in terms of the partial variance, so providing such a threshold value would be well accommodated into the selection process. Therefore, the number of variables to be selected will be the smallest that is sufficient to express at least 75% of $||\Sigma||^2$. For these data this corresponds to a subset of seven variables.

Having determined a size for the subsets, the various selection procedures investigated in Chapter 6 were applied to the data. The resulting variable subsets are displayed in Table 7.4 with variable names displayed in reduced form. Additionally, for each subset in the table the corresponding percentages of $\mathrm{tr}(\Sigma)$ and $||\Sigma||^2$ that are explained by that subset are given, as is the length of time taken to arrive at this subset. All calculations were run in R for Windows version 2.1.1 [101] on a Pentium IV 2.4GHz PC. It should be noted that for the stepwise inclusion methods **B4**, **HC** and **H** the variables are listed in the order of their selection and hence are in descending order of importance according to the corresponding selection criteria.

Looking first at the performance time of the different methods, we can see the marked difference between the exhaustive McCabe methods **M1**–**M3** and the other stepwise procedures. The McCabe methods required the enumeration and evaluation of all 77 520 possible seven-variable subsets which took approximately three minutes. The other procedures returned subsets almost instantaneously, with the exception of the multiple correlation method **A2** and Krzanowski's method **KP**. The reason for **KP**s slow performance was that it requires the calculation of several $(n \times n)$ matrices, which in this case has over 350 000 elements.

If we look at the general merit of the subsets returned by the different routines, we can see that most subsets convey more than 75% of $||\Sigma||^2$ and 55% of $\mathrm{tr}(\Sigma)$. The exceptions to this are methods **M1**, **B1** and **A2** - it was also seen in the previous chapter that these methods returned subsets that were slightly inferior

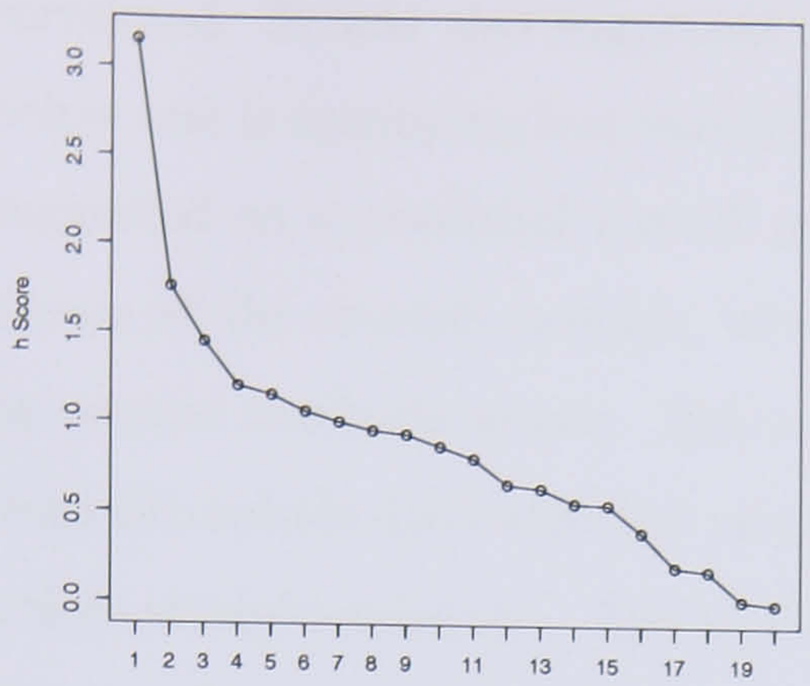
	Variable									
Method	1	2	3	4	5	6	7	% tr($\mathbf{\Sigma}$)	% $\|\|\mathbf{\Sigma}\|\|^2$	Time (s)
M1	Weight	PainF	Stab	SD	FCont	ExLag	OKF	48.8	64.0	184.35
M2	PainF	RU	GU	CTFAng	FCont	OKF	OHAb	56.1	80.4	180.97
M3	PainF	RU	GU	CTFAng	FCont	OKF	OHAb	56.1	80.4	158.15
B1	Weight	PainF	Stab	SD	FCont	ExLag	OKF	48.8	64.0	0.03
B2	PainN	SD	GD	FCont	OKFC	OKF	OHAb	54.3	77.8	0.02
B4	GU	OKFC	PainS	OHAB	SD	CTFAng	Weight	55.2	78.9	0.00
A2	PainS	PainN	Stab	RU	CTFAng	HipAb	OKF	50.7	69.6	4.05
DF	PainS	WAb	SD	GU	CTFAng	Flex	OKFC	54.7	78.9	127.85
KP	Weight	PainS	PainN	SD	GU	FCont	OHAb	55.6	79.2	17.55
HC	GU	RU	OHAb	WAb	OKFC	Flex	PainS	54.1	78.9	0.15
H	GU	RU	OHAb	PainS	FCont	OKF	ExLag	56.0	79.9	0.16

Table 7.4: Table of selected 7-variable subsets of the pre-operative knees data using various selection methods.

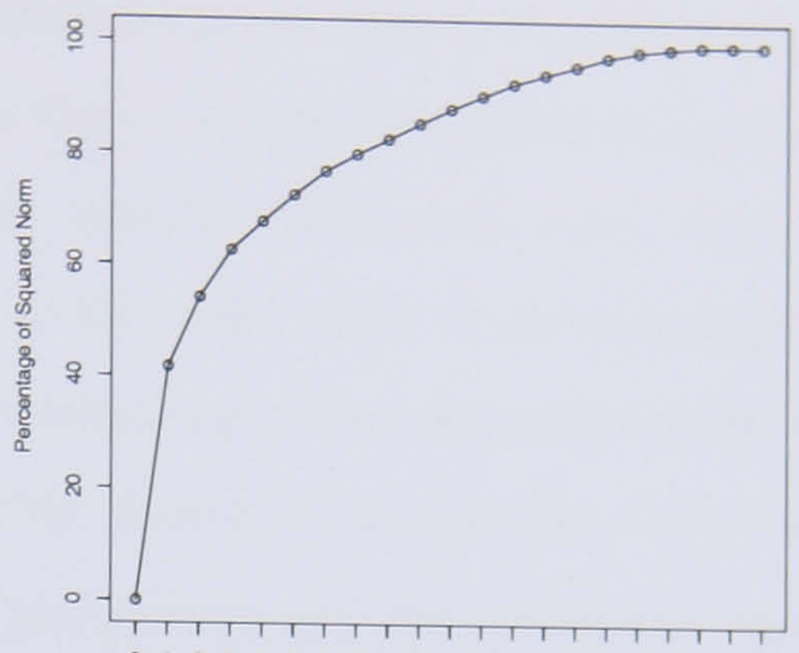
to those returned by the other methods. This disparity in performance aside, one feature worthy of note is the absence of any gulf in terms of the performance of the returned variable subsets between the exhaustive optimal methods and the stepwise non-optimal methods. Whilst the subsets returned by methods **M2** and **M3** do perform the best, the difference between these and the other subsets appears to be no more than 2% of $\text{tr}(\boldsymbol{\Sigma})$ and 3% of $||\boldsymbol{\Sigma}||^2$. This shows that the subsets returned by the stepwise selection methods are close to the optimal solutions. In practical terms, the significance of the differences between these two groups of solutions may be negligible.

If we now consider the composition of the subsets returned by the individual selection procedures we can see that there is a great deal of overlap. The similar performance of the subsets is likely a consequence of this. If we examine the subsets in terms of the variable groupings that we saw in the correlation plots in Section 7.2.1, we can firstly observe that all subsets contain one variable from the tightly correlated pair (*Sitting Down*, *Rising Up*). Furthermore, all methods but **M1**, **B1** and **A2** return a subset containing one variable from the pair (*Going Up Stairs*, *Going Down Stairs*) - this may be a reason for their slightly poorer performance. Similarly, all methods return at least one of the three pain score measurements with methods **A2** and **KP** both returning two. Further similarity can be seen in that all subsets except that of **A2** contain one of the *Fixed Contracture* measurements. This strong degree of overlap among the subsets selected suggests that there is a clear and definite structure underlying the data that is being systematically extracted.

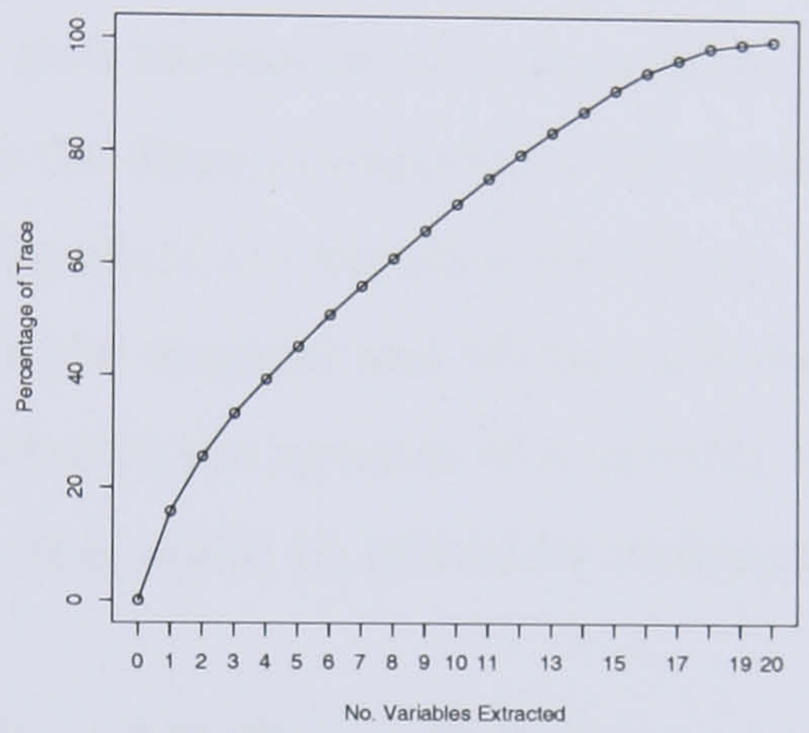
Focusing now on the performance of the methods **HC** and **H**, we can see that both the subsets returned agree in the choice and order of the three most important variables - *Going Up Stairs*, *Rising Up*, and *Other Hip Abduction*. Beyond this both subsets contain *Pain Severity*, albeit in a different position; and both contain one of the *Fixed Contracture* measurements and one of the *Flexion* measurements. The only differing choice in variable between the subsets is that **HC** chooses *Walking Ability*, whereas **H** chooses *Extension Lag*. This is significant as *Walking Ability* is part of the group of mobility scores and is associated with *Going Up Stairs* and *Rising Up*, whereas *Extension Lag* only has weak associations with other variables.



(a) h scores



(b) Percentage of $||\boldsymbol{\Sigma}||^2$ explained



(c) Percentage of $\text{tr}(\boldsymbol{\Sigma})$ explained

Figure 7.3: Scree plots produced from application of variable selection procedure **H** to the pre-operative knees data.

This suggests that the **HC** procedure is dismissing relatively uncorrelated variables in favour of those which are associated with variables already selected. This is as predicted in Chapter 6, and it is remedied by the modifications made in **H**.

Producing the scree-type plots discussed in Section 6.5.4 for the pre-operative knees data using method **H** yields the graphs in Figure 7.3. The first plot displays the h value for each selected variable in the sequence in which they were selected. That is to say the first point corresponds to the value for *Going Up Stairs*, the second for *Rising Up*, etc. From this we can see a steep decline for the first four variables, which then straightens out into a roughly linear pattern. We can notice that the h values do not decrease as rapidly as they did in the plot for the aphids data in Figure 7.3(a) which reflects the fact that the variables are not as tightly and homogeneously

correlated. It was also suggested in Section 6.5.4 that a variable with an h value below one is conveying less variability than a single independent variable. This was suggested as a potential cut-off point, similar to Kaiser's rule. In this case this occurs at the seventh variable, which is the same value of dimensionality suggested by several methods above. The percentage variation plots also differ significantly from the aphids data with the plot of the squared norm having a more visible curve before straightening out. The plot of the percentage trace is almost linear after the third or fourth variable has been chosen. This may suggest that the data are, to a degree, conditionally independent given these first few variables. An alternative explanation for the slow increase of the trace relative to the squared norm is that the trace considers only the diagonal elements of the conditional variance matrix, that is the conditional variances of the variables themselves. Conversely, the squared norm is composed of both the diagonal and off-diagonal elements accounting for variance and covariances. Thus if the removal of a variable had a significant effect on the partial covariances, this would be missed by looking at the value of the trace.

7.2.4 Reducing All Time Points

7.2.4.1 Individual Reduction

The stepwise variable selection procedure **H** was applied to all four time points individually to identify a reduced subset of candidate variables. However, this procedure was complicated by the fact that the variable *Coronal Tibio-Femoral Angle* has a zero variance at the 5- and 10-year time points and so a matrix of correlations could not be constructed for all variables. Therefore, for these individual analyses this variable was not considered as a possibility, leaving a group of 19 variables. The results of these variable selections are presented in the top four rows of Table 7.5.

Examination of these results shows that there is a significant degree of agreement between these subsets. For example, all time points choose one of *Going Up Stairs* (GU) or *Going Down Stairs* (GD) as the first chosen variable which is reasonable due to their good levels of correlation with other variables. All three post-operative time points select a pain variable as their second choice - this is not surprising as

the pain scores form a group of moderately correlated variables post-operatively, but only have weak associations with one another pre-operatively. The ordering of the remaining variables is not fixed across the time points, but the many of the variables selected are common to all four subsets. For example, *Other Hip Abduction* (OHAb) and *Other Knee Flexion* (OKF) are selected at all time points. Additionally, all subsets include one of *Fixed Contracture* (FCont) or *Other Knee Fixed Contracture* (OKFC). One variable from the pair *Sitting Down* (SD) and *Rising Up* (RU) can be found in all subsets except at five-years. The remaining variables then consist of at least one of the more independent variables, such as *Extension Lag* (EXLAG) or *Weight*.

This overlap and collaboration between the time points is encouraging and implies that whilst there may be changes in the ordering (and hence importance) of the principal variables, the principal variables do appear to be similar despite the passage of time, with the exception of the pain variables. Another point that is illustrated here is that typically only one variable from a subgroup of tightly inter-correlated variables is selected, e.g. {GU, GD}. This may imply that one of these variable may act as a surrogate for the other. This is also illustrated with the pain scores and the other subgroups mentioned above. The application of the procedure to determine the overall 'best' subset of variables will be able to eliminate this problem of potentially "equivalent" variables creating superficially different principal subsets. Another point of note is that due to the decrease in sample sizes over time, the second two time points will increasingly suffer from the effects of the sample variation causing a greater uncertainty over their results.

The merit of these subsets, as defined by the percentage trace or squared norm of the original correlation matrix, is listed in the final two columns of Table 7.5. We can see that the subsets become better at later time points as they explain a greater percentage of the variation. The reasons for this are unclear, but the jump in performance from the pre- to post-operative data is likely due to the fact that the pain scores are more correlated post-operatively. This would mean that the value of including a pain score in the subset is greater at later time points thereby boosting the percentage variation.

	Selected Variables									
Time	1	2	3	4	5	6	7		%tr($\boldsymbol{\Sigma}$)	%$\|\|\boldsymbol{\Sigma}\|\|^2$
Pre-op	GU	RU	OHAb	PainS	FCont	OKF	ExLag		56.0	79.9
1-year	GU	PainF	SD	OHAb	OKF	FCont	CTFAng		62.8	87.9
5-years	GD	PainS	OKFC	OKF	ExLag	OHAb	Weight		67.5	91.0
10-years	GU	PainF	FCont	RU	OHAb	ExLag	OKF		70.0	92.0
								Pre-op:	57.7	80.8
								1-year:	65.2	89.2
All	GU	PainS	OKFC	SD	OHAb	ExLag	OKF	5-years:	69.4	92.8
								10-years:	67.6	90.7

Table 7.5: Table of selected variables for the different time points of the knees data using method **H** and overall longitudinal variables selected using method **HT**.

7.2.4.2 Simultaneous Reduction

Before applying the selection method **HT** developed in Section 6.5.1 to extract longitudinal PVs, it is necessary to perform the nonparametric time-dependent PCA [100] on each time point. After examination of many possible values for the smoothing bandwidth, the value of $\sigma = 1$ was chosen. Briefly, the reasons for this choice were that the plots for larger bandwidths became homogeneous with little variation over time and that those for smaller bandwidths were overly distinct for the different times and displayed more variation. Two plots along the lines of those discussed in [100] for the knees data are displayed in Figure 7.4.

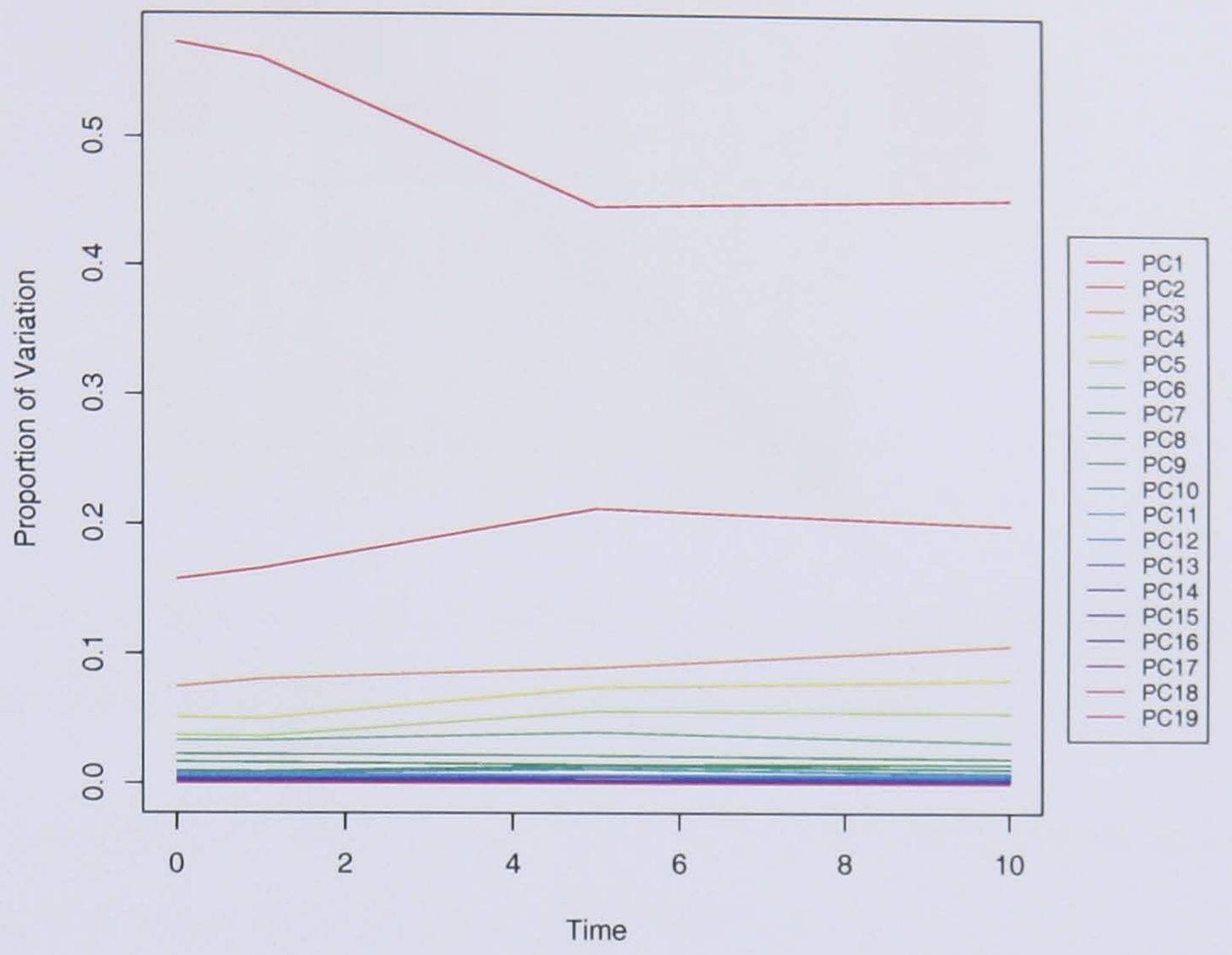
The first plot in Figure 7.4(a) is of the proportion of variation explained by the individual PCs over time and displays that the majority of the variation is explained by the first few principal components, though the importance of the first PC appears to drop at the 5-year point, whereas the importance of the other components seem to increase over this period. This may suggest a change in structure at this point or could be indicative of more noise in the data. The second plot is a plot of the loadings of the variables of the knees data in the first PC over time. This plot is more difficult to interpret, but we can see that the first component serves to contrast the pain scores (oranges and yellows), the walking ability measures (greens and cyans) and other variables with *Weight*, both *Fixed Contractures* and *Extension Lag*. We can see that there is an abrupt shift in the size of the loadings from the 1-year to the 5-year time point which might indicate a change in structure at this point. However, only one of the loadings changes sign which suggests that whilst the strength of associations may change, their directions remain the same. This variable is *Other Hip Abduction* and its initial loading was positive but close to zero. This may suggest that at earlier times this variable was merely unimportant rather than indicating a fundamental reversal of relationships. Similar conclusions can be drawn from the plots for the 18 other PCs.

If the principal components are not calculated in a temporal fashion, then slight variations in the data will likely lead to quite different PC solutions at each of the different time points. Another possible method for investigating possible structural changes in the components would be to calculate the correlations between the com-

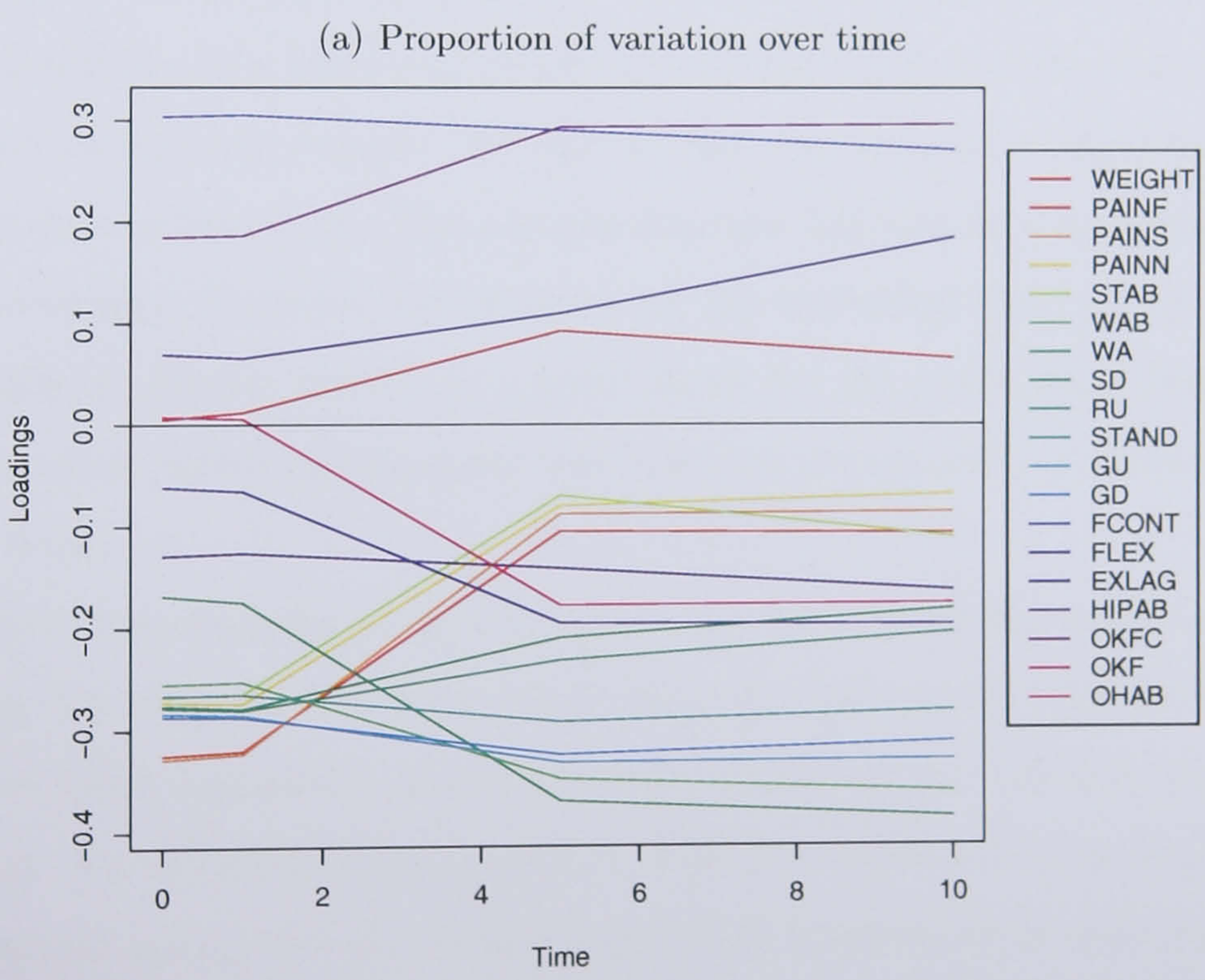
ponents obtained at two different time points. Excepting changes in order, if the structure were the same then we would expect to see each component at the first time point strongly correlated with exactly one of the components at the later time point. Performing such analysis with the knees data did not corroborate the hypothesis of no structural change. However, it is not clear whether this is evidence of genuinely different structure at different times, or whether these changes are attributable to small amounts of variation affecting the ordering and composition of PCs. Investigating potential structural changes over time would be an interesting area for future research, but is beyond the scope of this thesis.

Having applied the **H** method to each separate time point and obtained four distinct subsets, the temporal selection method **HT** can now be applied to all four time points simultaneously using the smoothing bandwidth of $\sigma = 1$ year to obtain a single subset for all the data. The results from this process are listed in the bottom row of Table 7.5. The subset reflects the common choices made when considering time points individually containing variables from each of the groups mentioned above. The performance of this subset is listed in terms of the percentage variation of the *original* (not smoothed) correlation matrix of the data at each time point. The performance on the data from each time point is encouraging. Indeed, the performance actually slightly exceeds that of the individual subsets for all but the 10-year data. This is likely due to certain combinations of variables yielding better results than would be expected when running a simple stepwise procedure. Performing an exhaustive search would, of course, identify these best combinations but would suffer from the problems involved with exhaustive methods. Nevertheless, the performance is still similar to the performance of the individual subsets.

The correlation plot of the partial variances matrix of the remaining variables of the 1-year knees data given the seven chosen variables is shown in Figure 7.5. We can see that there is little off-diagonal activity reflecting the fact that most of the covariances have been removed with the selected variables leaving the matrix close to diagonal. This may suggest that these variables are approximately conditionally independent given the variables we have selected, which may explain the high performance of the extracted subsets in terms of percentage trace and squared norm.



(a) Proportion of variation over time



(b) Loadings of PC1 over time

Figure 7.4: Plots of the results of the nonparametric time-dependent PCA on the knees data.

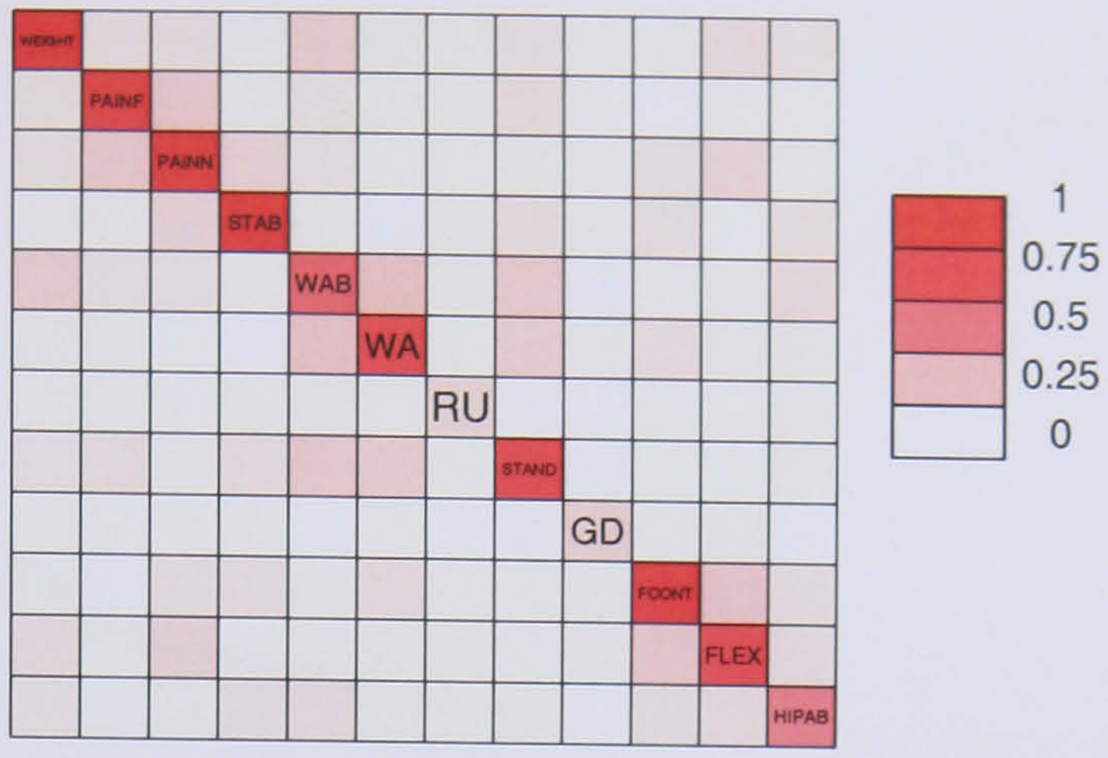
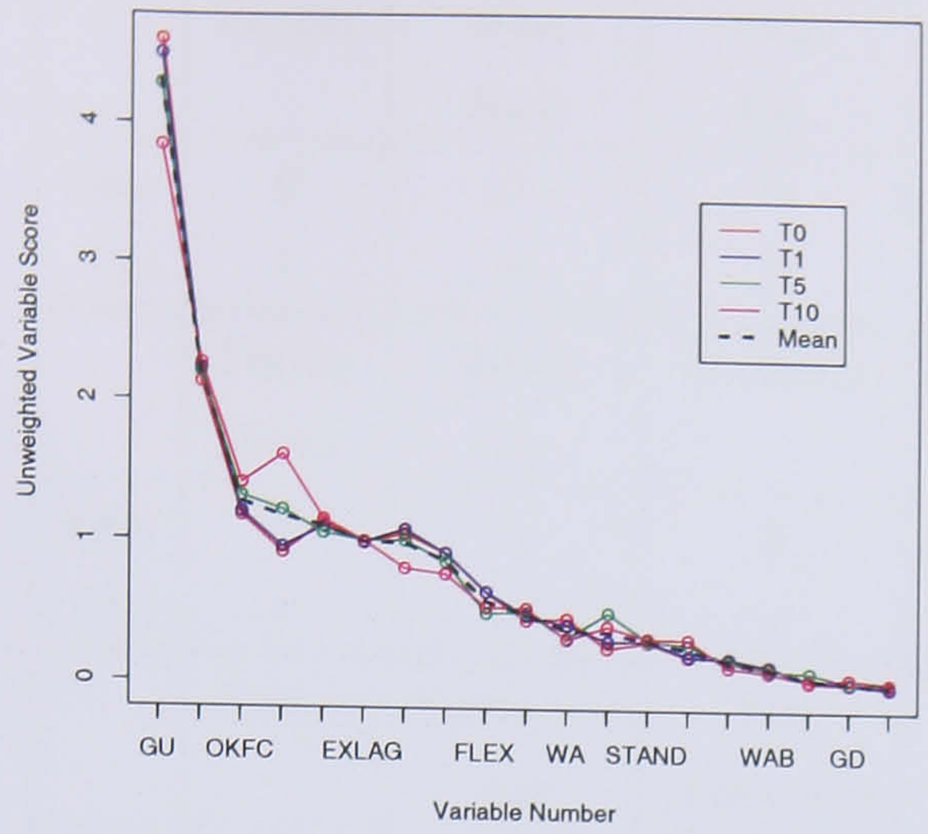


Figure 7.5: Correlation plot of the partial variance matrix of the remaining variables of the 1-year knees data given the seven chosen variables.

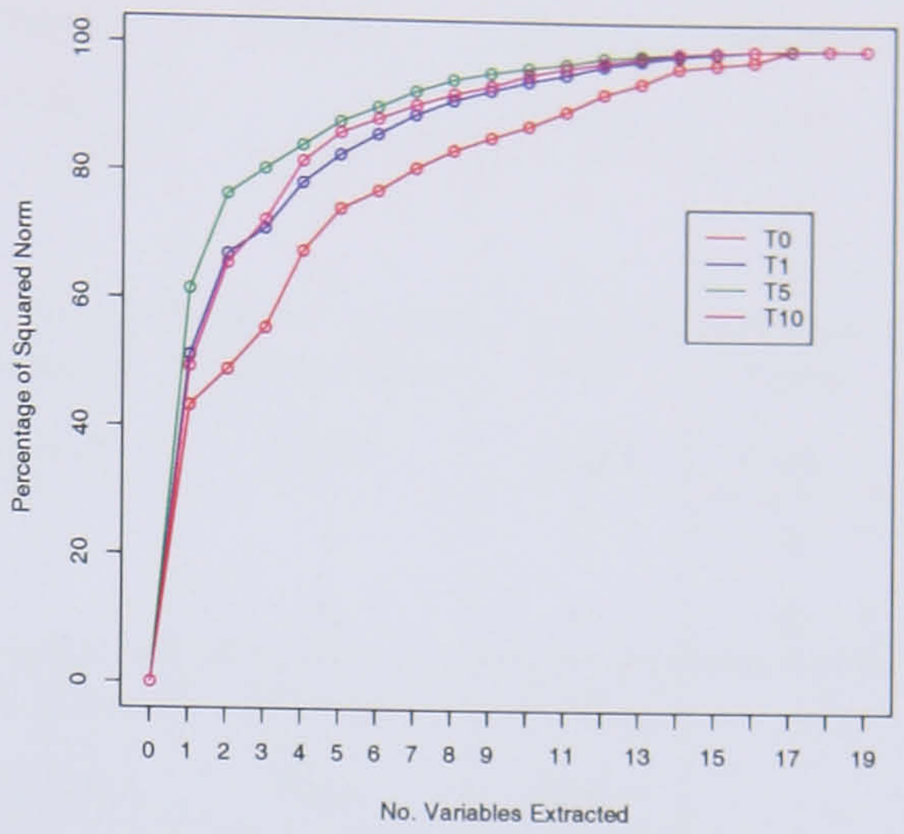
The lower score for the trace is due to the (relatively) high values for the partial variances of the variables displayed on the diagonal.

Scree-plots illustrating the temporal selection process are given in Figure 7.6. The first graph plots the score of each selected variable, with the different time points represented by different colours. We can see that the progress of the average score (the black dashed line) is of a sharp initial decrease followed by a straightening out, as seen previously. However, the progress for the individual time points is noisier. For example the fourth variable is a good choice for the 1-year data (showed by a peak in the pink curve) and a poorer choice for the pre-op and 1-year data (the red and blue lines).

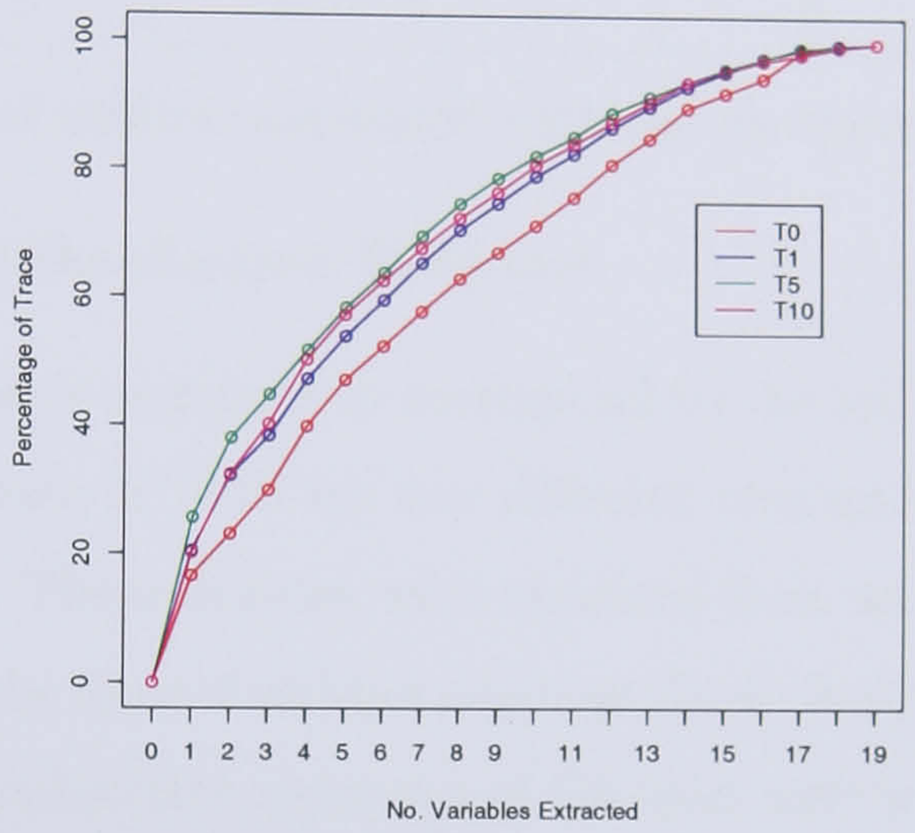
The percentage squared norm and trace plots are constructed using the original correlation matrices for the data rather than the temporally smoothed matrices in order to more adequately assess the performance at the different time points. Looking at the cumulative proportion of variation explained at each time point by the squared norm, we can see that this subset of variables is most suitable for the post-operative data. This is likely due to the fact that the structure of the data changes due to the intervention of treatment making the post-operative data relatively similar. This is shown by the inclusion of a pain score as the second variable - this is a sensible choice for the post-operative data as the pain scores form an obvious group in the correlation matrices. However, this is not an ideal choice for



(a) h scores



(b) Percentage of $||\mathbf{\Sigma}||^2$ explained



(c) Percentage of $\text{tr}(\mathbf{\Sigma})$ explained

Figure 7.6: Scree plots produced from application of temporal variable selection procedure **HT** to the knees data.

the pre-operative data as relationships between pain scores are rather looser. This choice results in only a small improvement in the percentage squared norm for the pre-operative data and a large improvement for the post-operative. The differences between the time points is also shown on the plot of the percentage trace with the pre-operative time point explaining the lowest percentage of the squared norm of its correlation matrix.

	Weight	Pain Freq.	Pain Sev.	Night Pain	Stability	Walking Ability	Walking Aids
Ease	9	7	6	7	5	8	8
Use	7	8	9	7	6	5	3

	Sitting Down	Rising Up	Standing	Going Up Stairs	Going Down Stairs	CTF Angle	Fixed Cont.
Ease	6	6	5	4	4	2	4
Use	7	8	4	7	5	6	5

	Flexion	Extension Lag	Hip Abduction	O. Knee F.Cont.	O. Knee Flex.	O. Hip Abd.
Ease	4	4	4	4	4	4
Use	7	5	4	4	6	4

Table 7.6: Table of utilities associated with the variables of the knees data.

7.2.4.3 Utility-based Simultaneous Reduction

Two sets of utilities were obtained to correspond to the knees variables by scoring each variable on a scale of 0–10 for two different characteristics associated with these measurements. These utilities were obtained from an expert and so can be considered typical of the form of utilities one may obtain in practice. These are given in Table 7.6 and are subjective measures of the ease with which the measurement is collected and the perceived clinical usefulness of recording and examining this measurement. Both utilities are rated on a scale of 0 to 10 where 0 corresponds to most difficult to measure or least useful and 10 is easiest to measure and most useful. No variables are scored with a 0 or a 10, so none of the variables will be forced into or out of the selection set.

Application of temporal variable selection procedure to the knees data using these utilities individually and together yields the variable sets given in Table 7.7. Comparison with the temporal results from Table 7.5 demonstrates that there are significant differences between these variable subsets and those obtained without utility information. The subset obtained using the ease of measurement utility shares only two of its variables with the principal subset determined without using utilities. The variable *Going Up Stairs* is also notably absent from this set due to its low utility compared to *Walking Ability* and its moderate correlation with this

variable, suggesting *Walking Ability* was chosen in its place. Another point of note is that we are starting to introduce potential redundancies in the reduced subset as two pain scores, *Pain Frequency* (PainF) and *Night Pain* (PainN), are included within the group. This incorporation of redundant elements will be due to the use of utility scores overriding the information of the data and the normal execution of the variable selection process. It should also be noted that the variable *Weight* has become more important according to the utility-based subset due to its high utility value boosting its perceived merit. In fact, six of the seven variables returned (including *Weight*) have a utility score above 5.

The variable subset obtained from using the 'clinical usefulness' utility bears slightly more similarity to the original results sharing three variables between the two subsets of seven. This suggests that the utility here is not so strongly at odds with the information from the data. In fact, if we consider the possible relationships within the data then the two subsets become more closely related. For example, *Rising Up* in the utility-based subset is closely related to *Sitting Down*, which is present in the original set. There are also similarities in the fact that both sets incorporate a *Flexion* measure. Combining the two utility measures by simple addition yields a subset that appears to incorporate the main features of both the individual subgroups with strong correspondence in the first few variables. The act of summing the two utilities to combine them appears to be producing sensible and appropriate results.

In terms of the performance of these subsets, we find that there is a lower percentage trace and squared norm explained by these variables when compared to the original longitudinal subset. Typically percentage trace is reduced by between 1 and 8% and the squared norm by between 2 and 6%. This difference is expected as we are no longer choosing the best variables just according to the h values alone. The act of using the utilities to modify the selections made forces us to choose non-optimal variables which have better utility values.

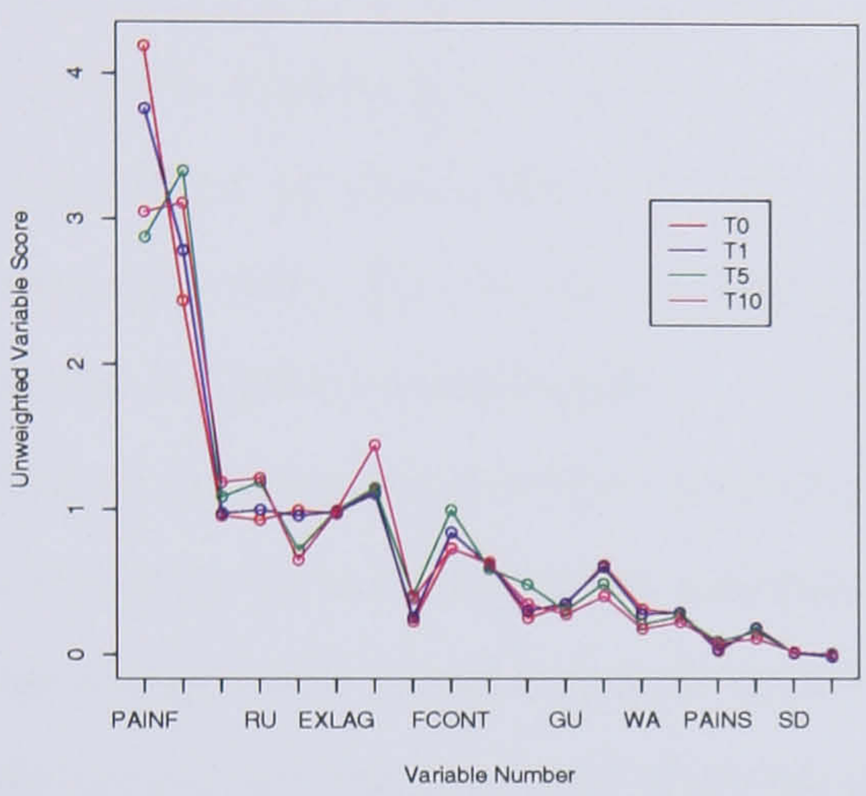
Scree plots for this application of the variable selection procedure using the combined utility are given in Figure 7.7. Figure 7.7(a) shows the h scores for the selected variable at each stage of the selection process. We can see from the peaks and troughs

	Selected Variables									
Utility	1	2	3	4	5	6	7		%tr($\boldsymbol{\Sigma}$)	%$\|\|\boldsymbol{\Sigma}\|\|^2$
Ease	WAb	Weight	PainF	SD	WA	OHAb	PainN	Pre-op:	53.1	75.3
								1-year:	58.6	83.4
								5-years:	61.7	87.5
								10-years:	63.6	88.5
Use	PainS	RU	GU	Weight	Flex	FCont	OKF	Pre-op:	54.7	76.4
								1-year:	63.2	86.9
								5-years:	66.8	90.5
								10-years:	65.0	88.5
Ease+Use	PainF	WAb	Weight	RU	Flex	ExLag	OHAb	Pre-op:	55.2	78.2
								1-year:	62.6	87.0
								5-years:	64.3	88.7
								10-years:	66.4	89.2

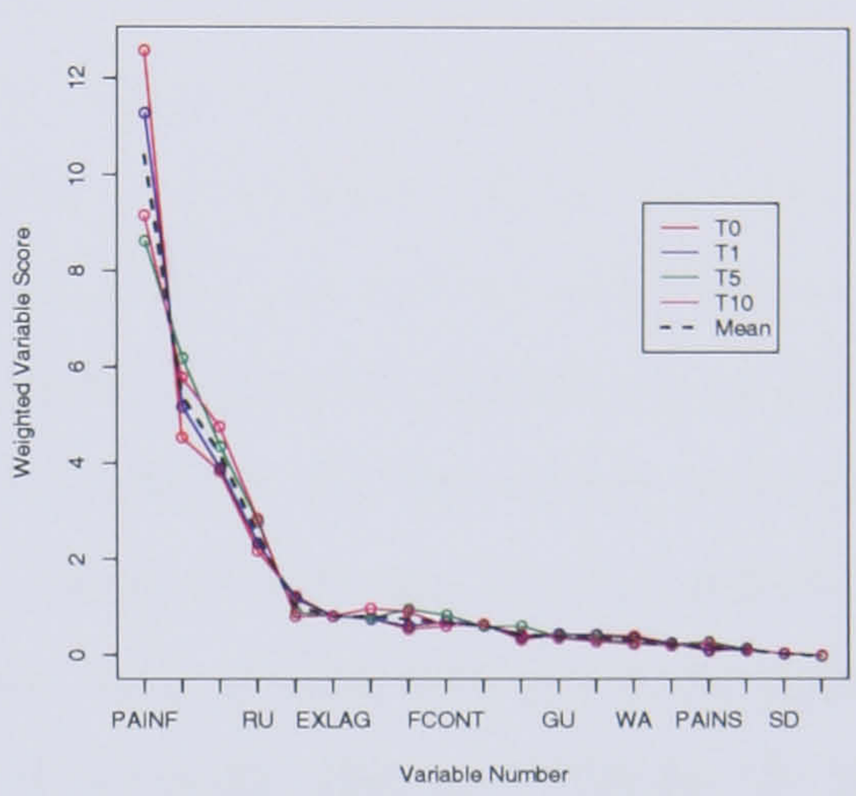
Table 7.7: Table of selected variables for the different time points of the hips data using method **HT** incorporating utilities.

in this plot that the selection procedure is not choosing the best variable in terms of h scores. The utility information is, to an extent, overriding this information allowing the procedure to select what would otherwise be a poor variable if it has a high utility. Figure 7.7(b) displays the utility-weighted variable scores for the selected variables at each stage for each time point. We can see that the variable scores decrease rapidly, they then straighten out and head towards zero. The shape of this plot is what we would usually expect from a plot of the h values from the standard procedure. The fact that the utility-based method selects variables based on the utility-weighted h values means that a plot of these scores will indeed resemble the plot of the h values for the standard procedure.

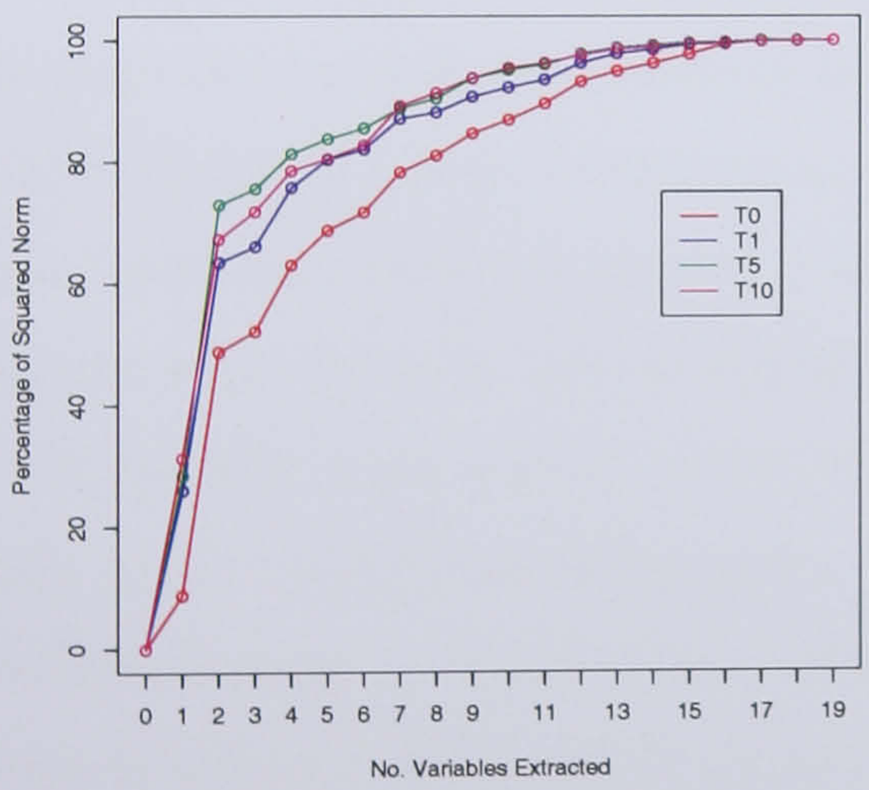
The percentage squared norm plot in Figure 7.7(c) shows a pattern similar to that of Figure 7.6(b), although with a shallower and more rugged ascent. This is to be expected as we are using the utilities to adjust our selections and so we will not select the 'best' variable according to the data so the amount of variability we capture will be less than in the utility-free situation. Again, we can see that the subsets selected are best for the post-operative time points. The plot in Figure 7.7(d) displays the difference between the percentage of the squared norm we explain by the temporal subset in Table 7.5 and the utility-weighted subset from Table 7.7. This is essentially the "loss" of information due to using utilities to adjust the selection process. This shows that after the first variable we typically only lose less than 10% of the variability using the utility methods. The choice of *Pain Frequency* for the first variable is a poor decision in terms of the data, resulting in a loss of 15–40% of the information we would have captured using the first variable from the standard method. That said, however, once we introduce a second variable the loss of information becomes more tolerable and more manageable. It is also evident that the loss of information decreases as we include more variables, with only a negligible loss after the 12th variable indicating that from this point both methods have similar performance.



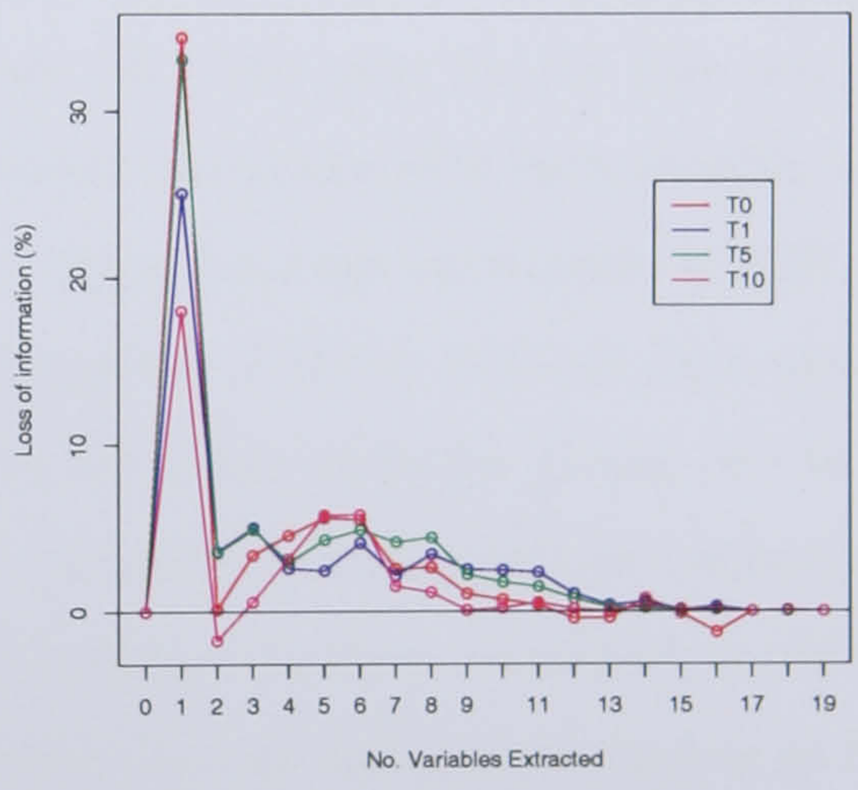
(a) h Scores



(b) Utility-weighted h Scores



(c) Percentage Variability Explained



(d) Loss of Variability

Figure 7.7: Four plots for the application of utilities in variable selection from the knees data.

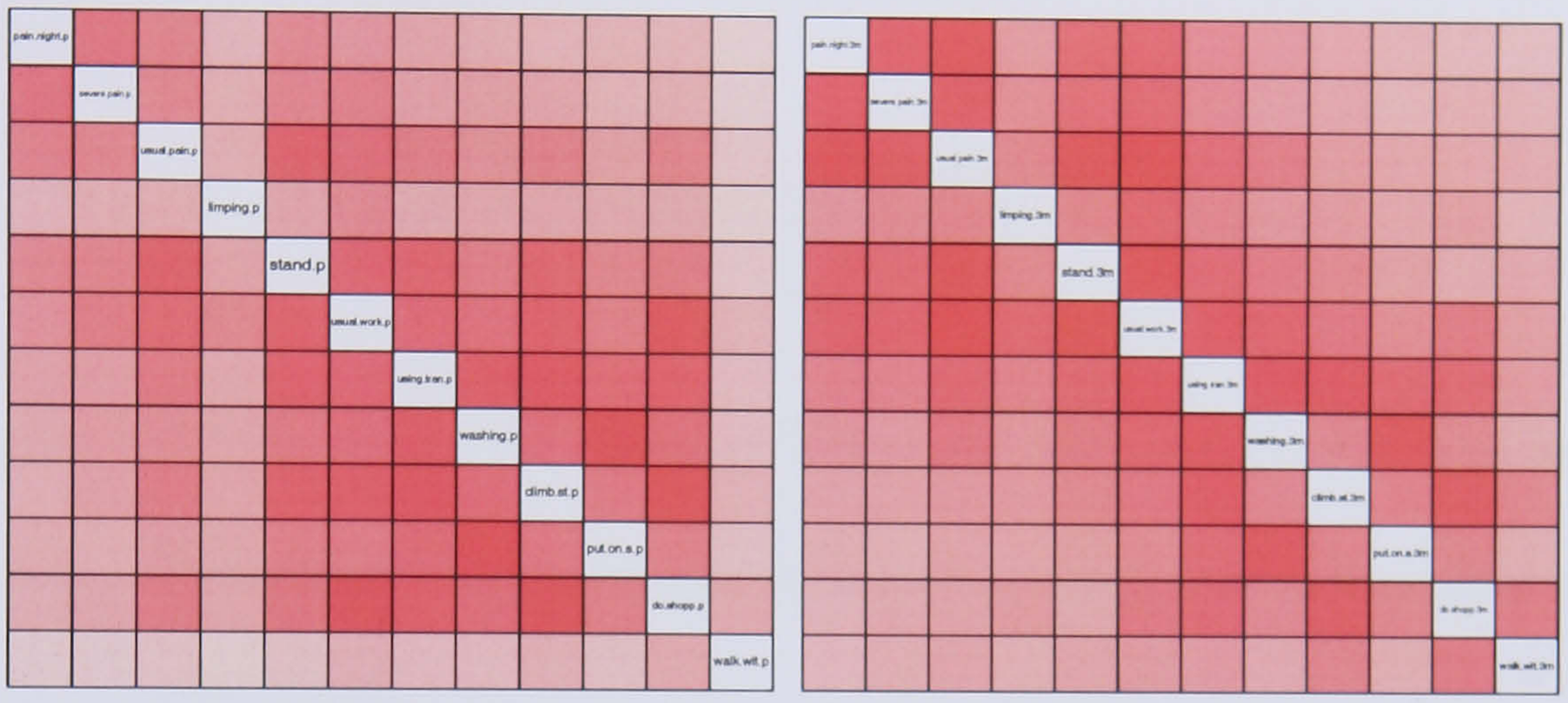
7.3 The Hips Data

7.3.1 Data Structure

The hips data are discussed in Section 2.2.2 and contain eleven repeatedly-measured patient status variables observed at three distinct time points. All eleven variables are measured on a five-point Likert scale and are observed pre-operatively and at 3 and 12 months post-operation. As with the knees data the sample size is not constant as patients may not have completed or returned all of the corresponding questionnaires. To this end, attention is restricted to all patients who had a completed surgeon's questionnaire - this contained important information on patient demographics as well as their pathology and treatment. Further to this, cases where a majority of variables were missing were excluded with any missing data in the remaining cases being imputed by the mean value at that time point. In terms of the actual sample size, this gives samples of size 4634, 2488 and 2338 for the three time points respectively.

Correlation plots of the correlations between the variables observed at the individual time points are presented in Figure 7.8. The plots for the first two time points illustrate a fairly homogeneous correlation structure with each variable being associated to a moderate degree with any other (average correlation is 0.38 pre-operatively). However, pre-operatively the variable *Walking Without Pain* appears to be slightly more weakly correlated with the other variables giving rise to the pale bands on the correlation plot. At 12 months, the associations between the variables appear to remain fairly similar in intensity to those displayed at the previous two time points. Indeed it would appear that the associations appear to have strengthened over time with the level of correlation becoming slightly stronger at each successive time point.

It is clear that the hips data exhibit a strong degree of association between variables suggesting a strong correspondence between the information conveyed by the variables. This implies a notable level of overlap in information and therefore redundancy within the measurements. Therefore, one would expect variable reduction to be particularly effective for these data.



(a) Pre-operative ($n = 4364$) (b) 3-months post-operative ($n = 2488$)

(c) 12-months post-operative ($n = 2338$)

Figure 7.8: Correlation plots of the repeated measurements in the hips data observed at each of the three time points.

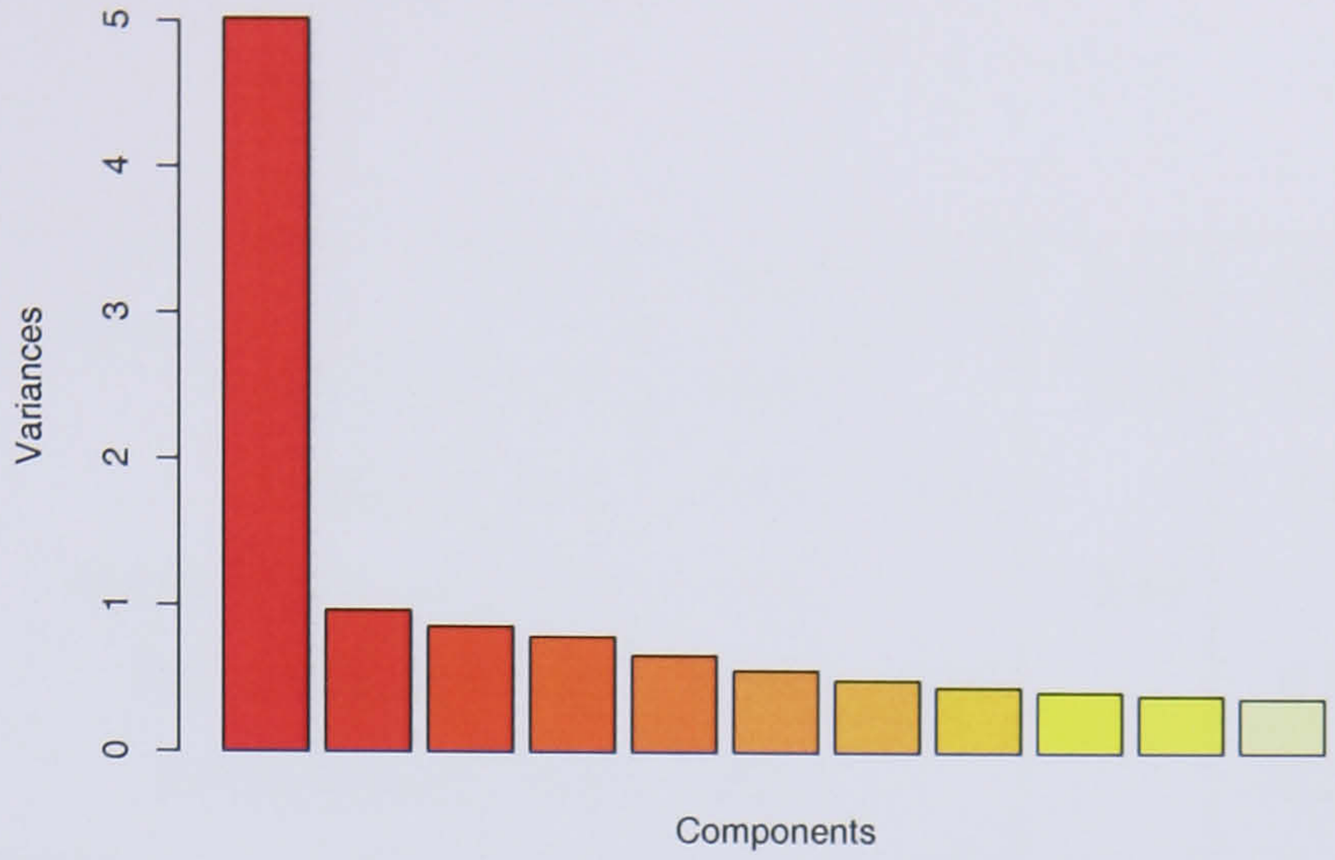


Figure 7.9: Scree plot for the principal components of the pre-operative hips data.

7.3.2 Principal Component Analysis and Dimensionality Assessment

A principal component analysis of the pre-operative hips data yields the loadings and variances for the first five components that are presented in Tables 7.8 and 7.9. Examining the principal components themselves shows that the first component is largely a weighted average of all elements of the data set which represents 44% of the variation of the whole data set. This component would correspond to an optimal univalue hip score that would express the most variation of any single composition of these individual elements. The second component only accounts for 9% of the variation of the data set, and contrasts the three pain scores with six of the ability scores. The third and fourth components principally represent the single variables *Walking Without Pain* and *Limping* respectively. The fact that a single component is dominated by *Walking Without Pain* is not surprising due to its lower than average correlations to other variables. The fourth component contrasts *Putting On Socks* and *Washing* with several general ability measures.

Assessing the dimensionality of the hips data using the same methods as for the knees data we obtain the estimates in Table 7.10. Most of the methods suggest a dimensionality of 1 or 2 which is likely due to homogeneous nature of the correlation matrix with moderate correlations between all the variables. However, this is likely

	PC1	PC2	PC3	PC4	PC5
Usual Pain	0.30	−0.25		0.19	−0.25
Washing	0.31	0.21	0.23		0.44
Using Transport	0.32	0.18		0.19	
Put On Socks	0.29	0.24	0.30		0.54
Do Shopping	0.32	0.27			−0.32
Walking w/out Pain	0.16	0.36	−0.86		0.21
Climb Stairs	0.32	0.23			−0.31
Stand From a Chair	0.32			0.17	−0.17
Limping	0.22			−0.92	
Severe Pain	0.26	−0.49	−0.17		
Usual Work	0.34				−0.24
Night Pain	0.25	−0.52	−0.17		0.31

Table 7.8: The first five principal components of the pre-operative hips data. Loadings of value < 0.15 have been omitted for clarity.

	PC1	PC2	PC3	PC4	PC5
Standard Deviation	2.3003	1.0529	0.9393	0.8961	0.8363
Proportion of Variance	0.4410	0.0924	0.0735	0.0669	0.0583
Cumulative Proportion	0.4410	0.5333	0.6069	0.6738	0.7321

Table 7.9: The importance of the first five principal components of the pre-operative hips data.

Method	Dimensionality
Kaiser's rule $\lambda < 1$	2
Kaiser's rule $\lambda < 0.7$	7
75% of variation	6
Scree Plot	1
Broken Stick	1
Velicer	1

Table 7.10: Table of estimates for the intrinsic dimensionality of the pre-operative hips data.

an oversimplification and would likely be an inappropriate estimate in terms of the number of variables. The other methods suggest values of 6 or 7, which is markedly different from an estimate of 1. It is difficult to suggest which of these values is the most appropriate - using a single dimension gives 44% of the variability, but doing so loses all of the detail about the individual scores. However, retaining 7 dimensions in this case is likely over-conservative. The true value of the intrinsic dimensionality probably lies somewhere in between.

7.3.3 Reducing the Pre-operative Data

Using the 75% $||\Sigma||^2$ rule suggests a four-variable subset would be adequate to represent the pre-operative data. Performing the variable selection as in the previous section yields the results given in Table 7.11. Krzanowski's Procrustes method could not be performed as it required calculation of many 4634×4634 matrices, each having in excess of 21 million entries which was well beyond the confines of available computer memory.

Looking at the speed of the various selection methods, we can see that all methods returned in reasonable time. Even the exhaustive methods were quick to evaluate as only 495 subsets needed to be considered in this case. Turning to performance of the returned subsets, we see again that all methods returned subsets of generally similar quality. McCabe's **M3** method was the best, followed by M2 and then the new methods **H** and **HC**. Again the determinant method M1 and Jolliffe's **B1** per-

	Variable						
Method	1	2	3	4	% $\text{tr}(\boldsymbol{\Sigma})$	% $\|\|\boldsymbol{\Sigma}\|\|^2$	Time (s)
M1	Put On Socks	Walk W/out Pain	Limping	Pain at Night	53.1	81.0	0.89
M2	Washing	Walk W/out Pain	Climb Stairs	Pain at Night	56.1	87.2	1.02
M3	Washing	Do Shopping	Stand From Chair	Severe Pain	55.3	87.8	0.72
B1	Put On Socks	Walk W/out Pain	Limping	Severe Pain	53.2	81.6	0.03
B2	Walk W/out Pain	Stand From Chair	Limping	Usual Work	54.9	84.8	0.01
B4	Usual Work	Pain at Night	Walk W/out Pain	Limping	53.1	82.3	0.00
A2	Put On Socks	Limping	Usual Work	Pain at Night	53.1	81.0	3.05
KP	—	—	—		—	—	—
DF	Washing	Climb Stairs	Usual Work	Pain at Night	55.3	87.8	1.18
HC	Usual Work	Using Transp	Stand From Chair	Washing	53.3	86.6	0.06
H	Usual Work	Put On Socks	Walk W/out Pain	Pain at Night	55.7	86.7	0.07

Table 7.11: Table of selected 4-variable subsets of the pre-operative hips data using various selection methods.

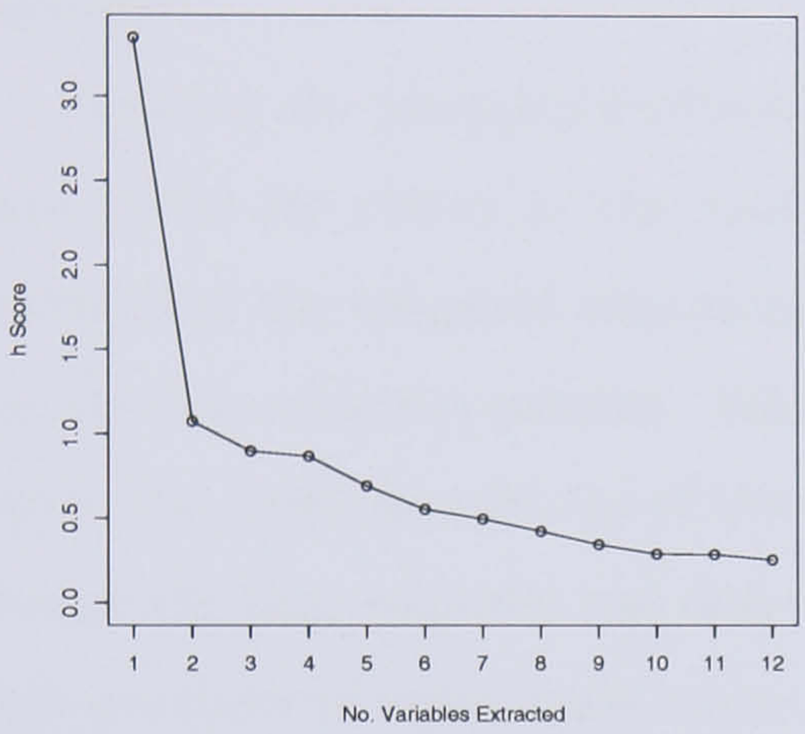
formed relatively poorly, this time with the addition of the forward selection PCA method **B4**. The multiple correlation method **A2** performed surprisingly well on these data, perhaps indicating it to be more successful on homogeneously correlated data as it typically ignores independent uncorrelated variables. **DF** also works well on these data as there appears, from the correlation plots, to be one underlying dimension so the corresponding graphical model would have a single focal variable.

In the absence of any strong and obvious structure to the correlation matrix of the hips data it is hard to divine what variables may or may not be appropriate selections. Nonetheless, most methods have chosen at least one of the pain scores (*Pain at Night*, *Severe Pain*). Many methods have also selected the variable *Walking Without Pain* which exhibited lower correlations than other variables. *Usual Work*, *Put On Socks* and *Limping* were also popular choices.

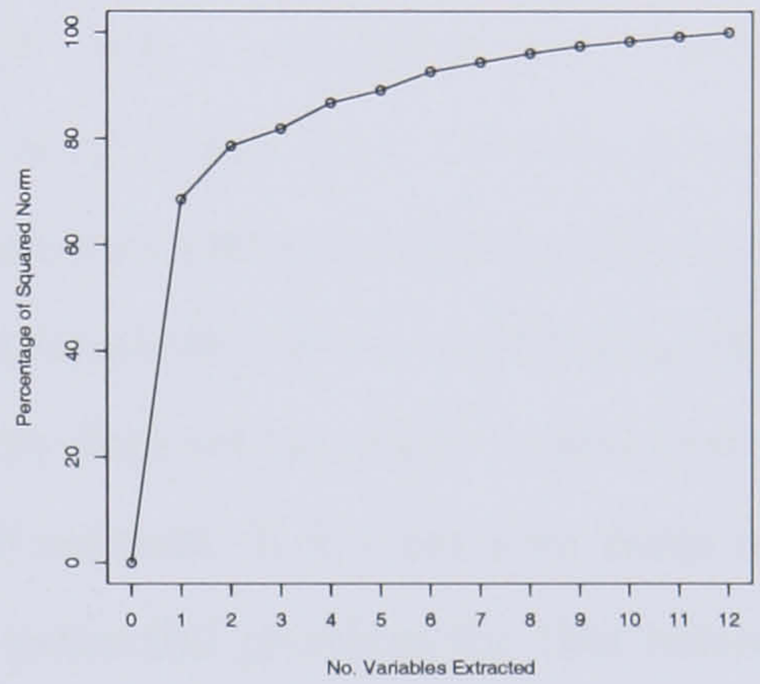
The scree plots for the hips data are shown in Figure 7.10. The first plot shows that the h values for the variables in the data have a steeper initial drop than the knees data and this is followed almost immediately by an approximately linear pattern which may suggest that only one variable is required. The h value first falls below unity for the third variable selected, suggesting 3 as a possible dimensionality. Both of these values are reasonable and tally with the values obtained previously by other methods. The plot of the cumulative proportion of variation in terms of squared norm has a steeper initial ascent than the knees data, probably due to the fact that since all variables are moderately correlated then once the first variable is selected the remainder convey relatively little novel information. The plot of the percentage trace increases far more slowly, as with the knees data. This is representing a different story to the percentage squared norm, suggesting both of the subsets for the hips and the knees data have performed in an almost linear fashion despite the data sets being quite different in terms of the underlying structure.

7.3.4 Reducing All Time Points

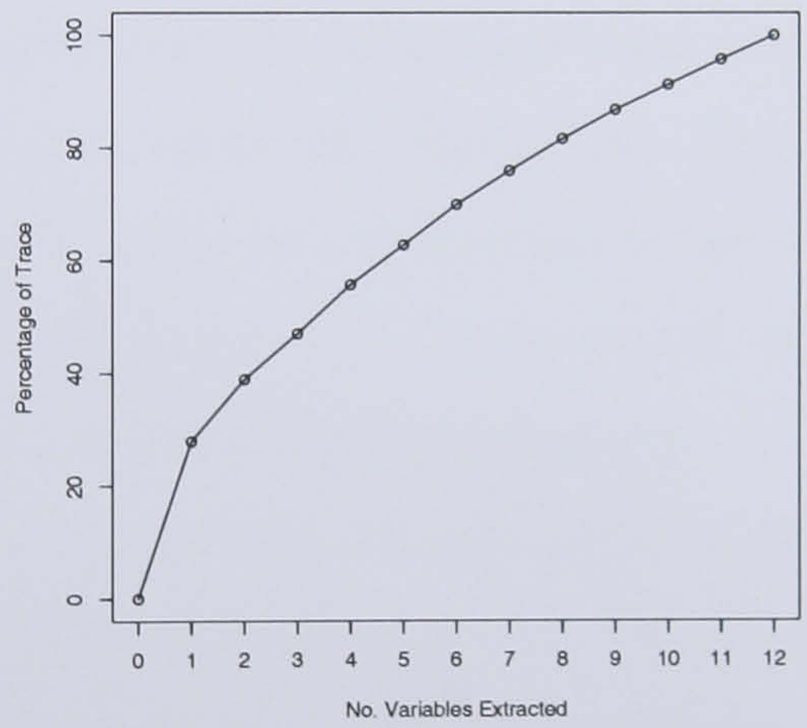
Applying the **H** method to each of the three time points in the data yields the results in the first three rows of Table 7.12. Looking at the subsets, we can see that there is a strong degree of agreement between the subsets with all subsets containing *Usual*



(a) h scores



(b) Percentage of $||\boldsymbol{\Sigma}||^2$ explained



(c) Percentage of $\mathrm{tr}(\boldsymbol{\Sigma})$ explained

Figure 7.10: Three scree plots produced from application of variable selection procedure **H** to the pre-operative hips data.

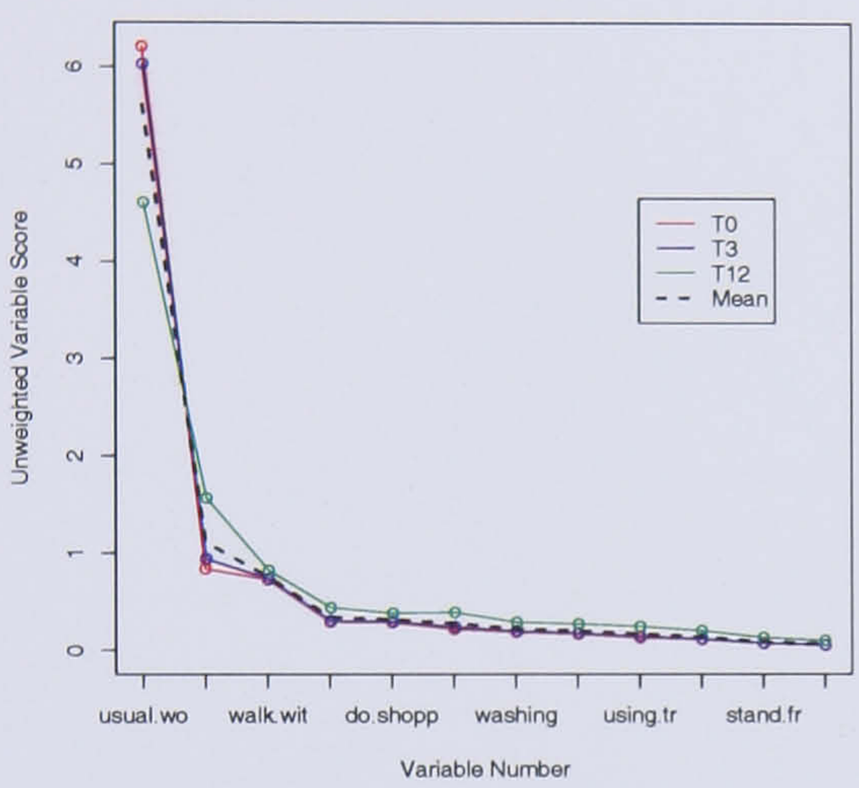
Work, *Walking Without Pain*, *Put On Socks* and one of *Night Pain* or *Severe Pain*. In fact, the 4-variable subsets obtained for the post-operative time points contain exactly the same variables reflecting their similarity of correlation structure. All four subsets perform well, representing at least 55% of the trace of the original correlation matrix and at least 86% of its squared norm. This good performance on a small group of variables is likely due to the strong correlations exhibited by all variables.

Applying the temporal method **HT** with a bandwidth of $\sigma = 6$ months to the data yields the subset in the final row of Table 7.12. At this point, it is worth noting that the temporal smoothing process increases the speed of computation in the variable selection process. Whilst negligible when considering the knees data, due to the larger sample size of the hips data set the effect is more pronounced here though the time required was only 2.6 seconds. It is clear that large data sets with high numbers of cases could cause a potential problem for this temporal method. Nevertheless, the longitudinally obtained subset of four variables is equivalent to that obtained at the 3-month and 12-month time point and is similar to the pre-operative subset with *Severe Pain* replacing *Night Pain*. The fact that we have obtained a temporal subset which is the best individual subset for both of the post-operative time points, is obviously due to the fact that both time points have identical four-variable subsets and so when averaging over time the scores for these subsets will be high. Looking at the performance of the subsets we can see that performance on this subset is identical to the performance of the individual subsets post-operatively, and pre-operatively we observe a slightly poorer level of captured variation in terms of the trace. This relatively low value for the percentage trace reflects the fact that pre-operatively we preferred *Night Pain* over *Severe Pain*.

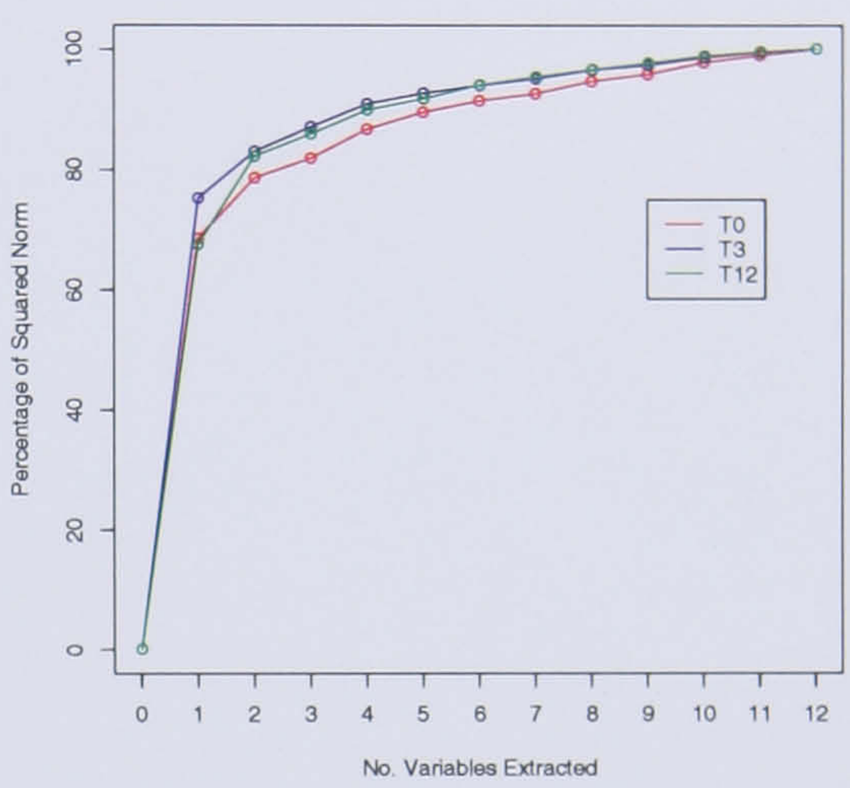
The scree-plots for the temporal selection process are shown in Figure 7.6. The graph of the change in h scores shows a steep initial drop representing that a great deal of the variation is captured in the first few variables, with the plot levelling off after the fourth variable. The four time points are typically in agreement over the choice of variable with little disagreement after the first two variables have been selected. The major initial change is also reflected in the plot of the percentage squared

	Selected Variables						
Time	1	2	3	4		%tr($\boldsymbol{\Sigma}$)	%$\|\|\boldsymbol{\Sigma}\|\|^2$
Pre-op	Usual Work	Put On Socks	Walking W/out Pain	Night Pain		62.8	86.7
3-months	Usual Work	Put On Socks	Walking W/out Pain	Severe Pain		60.0	91.0
12-months	Usual Work	Put On Socks	Walking W/out Pain	Severe Pain		64.5	93.7
All	Usual Work	Put On Socks	Walking W/out Pain	Severe Pain	Pre-op: 3-months: 12-months	55.5 60.0 64.5	86.7 91.0 93.7

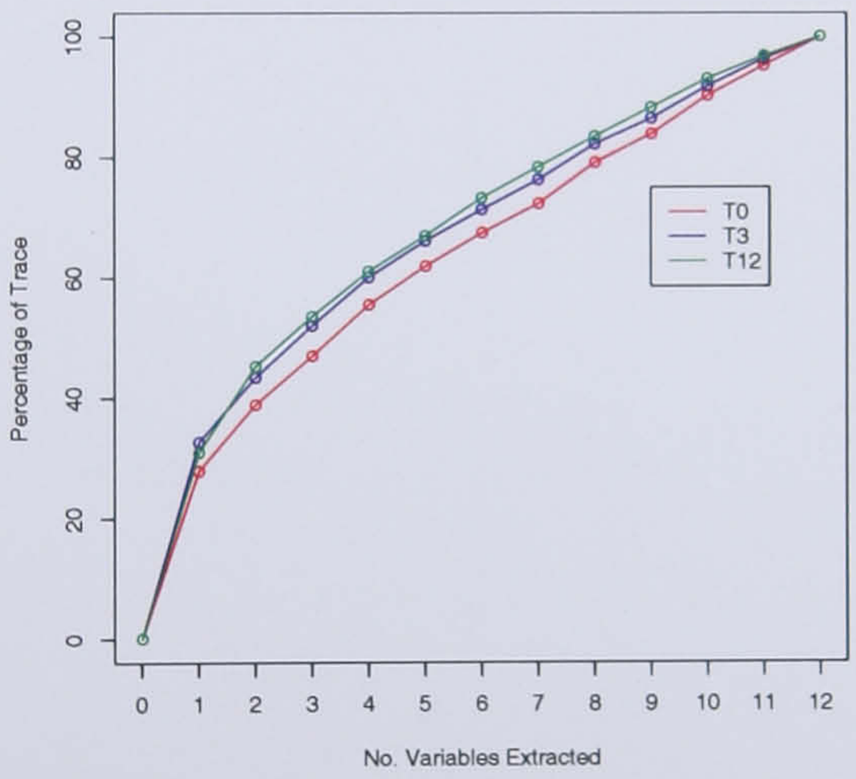
Table 7.12: Table of selected variables for the different time points of the hips data using method **H** and overall longitudinal variables selected using method **HT**.



(a) h scores



(b) Percentage of $||\boldsymbol{\Sigma}||^2$ explained



(c) Percentage of $\text{tr}(\boldsymbol{\Sigma})$ explained

Figure 7.11: Three scree plots produced from application of temporal variable selection procedure **HT** to the hips data.

norm. It also illustrates that, as with the knees data, the two post-operative time points have a similar high-level of performance on the selected variables, whereas the pre-operative data (red line) fares slightly worse. A corresponding result is again shown in the plot of the percentage trace.

Chapter 8

Chain Graphs and Prediction

Having generated some graphical models for the orthopædic data in Chapter 5, it was discovered that there were a number of limitations that hampered the development of larger models. Therefore the previous two chapters have covered, in detail, variable reduction strategies and the results of their application to the data. Having accomplished this, we now return to the issue of constructing viable models for these data using graphical modelling techniques. However, rather than continue to apply the standard graphical modelling approach, chain graphs [124, 52, 22] shall be used instead. Chain graphs are an extension to the previously discussed graphical models and allow for the incorporation of the temporal structure of our data as well as providing some computational benefits.

This chapter is therefore organised as follows. Section 8.1 introduces the concept of chain graph modelling. It begins with a review of the theory behind chain graphs and their associated graphical models - the majority of the material presented in this section is extracted from the texts of Edwards [38] and Whittaker [125]. Following on from this, in Section 8.1.2 several applications of chain graph modelling which are available in the literature are reviewed. Having established the general basis of the methodology, in Section 8.2 the techniques are applied to the orthopædic data to build suitable models for multiple time points. The models are presented, and their conditional independence implications and the nature of some of the modelled relationships are also discussed. The chapter concludes in Section 8.3 with a more detailed investigation into the predictive capabilities of these models, their validation

and an assessment of their goodness of fit.

8.1 Chain Graphs and Other Preliminaries

8.1.1 Chain Graph Theory

8.1.1.1 Block Structure

When constructing a statistical model it is commonplace to partition the variables in the model into a number of groups, for example in a simple regression setting we split the variables into two groups - the covariates and the responses. These two groups represent a partial ordering of our variables with our covariate variables all being contemporaneous and prior to our responses. The graphical models constructed in Chapter 5 have an implicit assumption that the constituent variables are all concurrent and therefore it is reasonable to introduce symmetric associations between them. This is not the case when we have an ordering over the variables with some being observed prior to the others, and so those models ignored this structure to the data. Nonetheless, having a partial ordering over the variables is particularly useful when there is this temporal structure to our data. This is particularly true with the abstract data model discussed in Section 2.1 where, for example, the patient's pre-operative status variables are antecedent to the treatment variables. So we could partition the variables within our data into a series of groups each representing variables observed at a particular time point; we can then exploit this *block-recursive* structure to our data.

In general, we assume our set of variables, V, satisfies a particular type of partial ordering, $\preceq$. This ordering is derived from the condition that the variables can be partitioned into subsets $B_1 \cup B_2 \cdots \cup B_k$, called *blocks*, which are completely ordered - and hence the blocks form a *chain*. This induced partial order on the variables V is such that $x \prec y$, whenever $x \in B_r$, and $y \in B_s$ and $r < s$; and $x \simeq y$ whenever $x, y \in B_r$. Furthermore, we consider variables within the same block to be concurrent and hence assume their association structure to be symmetric. For variables from different blocks we introduce a direction to the association allowing only associations

from earlier to later blocks. Corresponding to this block-recursive structure for the data, it is assumed that the joint density of our data can be factorised as follows:

$$f(B_1, \ldots, B_n) = f(B_1)f(B_2|B_1)\ldots f(B_k|B_1 \cup B_2 \cdots \cup B_{k-1}). \qquad (8.1)$$

This factorisation of the density formalises the notion that the density each block of variables depends only on the variables within that block and those variables that have preceded it. This is a key concept of this block structure.

8.1.1.2 Chain Graphs

To capture this block structure in the form of a graph it is necessary to attach direction to some edges since the undirected graph framework is no longer applicable. As before, we write a graph as $\mathcal{G} = (V, E)$ where V is our set of nodes, E is our set of edges, and we identify an edge with an *ordered* pair of vertices. Whenever we have both $(x, y) \in E$ and $(y, x) \in E$, then we interpret this as an undirected edge and draw a line between x and y. This is written as $x \leftrightarrow y$. Whenever $(x, y) \in E$ and $(y, x) \notin E$, then this is a directed edge and we say $x \rightarrow y$ and draw an arrow from x to y. If $x \rightarrow y$, $y \rightarrow x$ or $x \leftrightarrow y$ then we say that x and y are adjacent $(x \sim y)$.

These graphs are known as *block-recursive* or *chain* graphs and their properties are well documented in the literature [83, 124, 52, 22, 116, 9, 82]. The class of chain graphs includes both undirected graphs and directed acyclic graphs (DAGs) as special cases when all edges are undirected or directed respectively.

One of the restrictions of this class of graphs is that we prohibit graphs that contain directed cycles and graphs which contain cycles with at least one directed edge. Such graphs violate the partial ordering assumption of our variables as they would allow variables to belong to multiple blocks, and so they cannot be chain graphs.

As with the undirected graphs in Chapter 5, we attach conditional independence statements to the chain graph to represent the association structure between the variables in the model. Associations between variables within the same block are represented by undirected edges which are drawn as lines; and associations between

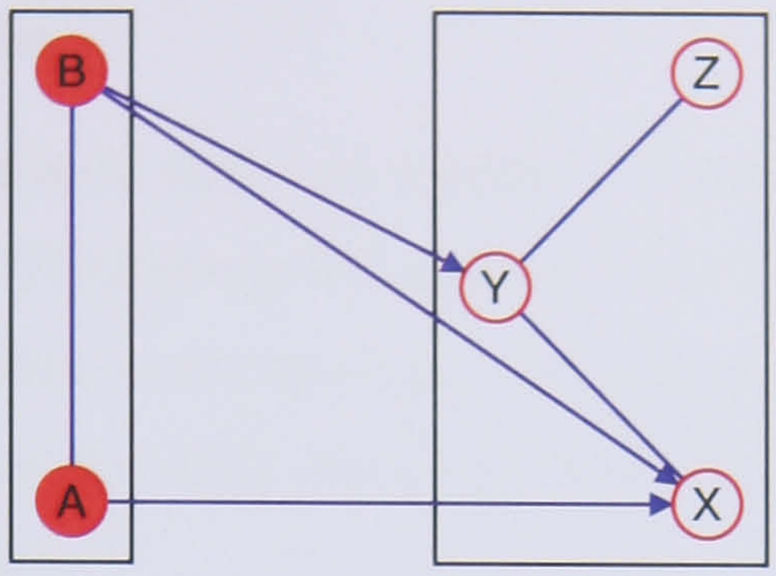


Figure 8.1: A simple chain graph.

variables of different blocks are shown via directed edges (arrows) from the earlier block to the later. An example of a chain graph is given in Figure 8.1. The two blocks of variables, $\{A, B\}$ and $\{X, Y, Z\}$, are emphasised in the graph by being enclosed in boxes.

As with the undirected independence graphs, the statements of conditional independence are defined in terms of the absence of specific edges from the graph. So, if a line is missing between two variables x and y from the same block B_i, or an arrow is missing from $x \in B_j$ to $y \in B_i$, for $j < i$, then this means that:

$$x \perp\!\!\!\perp y | B_1 \cup B_2 \cdots \cup B_i, \tag{8.2}$$

which is a version of the pairwise Markov property for chain graphs. In other words, the interpretation of a missing edges between a pair of variables is that those variables are conditionally independent given all other prior and concurrent variables, where we define prior and concurrent relative to the later of the two variables.

The chain graph of variables $V = \{X_1, \ldots, X_m\}$ has then been defined by Whittaker [125] as the graph $\mathcal{G} = (V, E)$, where $b(x)$ corresponds to the block B_i such that $x \in B_i$, $V(x) = \bigcup_{l \leq b(x)} B_l$ and the edge (x, y) with $x \preceq y$, is *not* in the edge set E if and only if $y \perp\!\!\!\perp x | V(y) \setminus \{x, y\}$. If this condition fails, and $x \prec y$, then the edge is directed and only $(x, y) \in E$; otherwise it is undirected and both $(x, y) \in E$ and $(y, x) \in E$.

8.1.1.3 Markov Properties

As with the undirected graphs there are a series of Markov properties; these have been studied in depth by Frydenberg [52] and are useful for interpreting the underlying conditional independence relationships of a chain graph. However, the Markov properties for chain graphs are often somewhat less transparent than their equivalents for undirected graphs.

In order to discuss the Markov properties, we must first define some particular graph theoretic quantities. Define the *neighbours* of a node x as the set of nodes that are joined to x with a line, and let the *parents* of x be all those nodes which are the origin of an arrow pointing to x. Then the *boundary* of x is the set of nodes which are parents or neighbours of x i.e. $\mathrm{bd}(x) = \{y \in V : x \sim y \text{ or } y \rightarrow x\}$. The *descendants* of node x, $\mathrm{de}(x)$, are the vertices y such that $x \rightarrow y$ and not $y \rightarrow x$. The *non-descendants* are then defined to be $\mathrm{nd}(x) = V \setminus (\mathrm{de}(x) \cup \{x\})$.

In formal terms, the *pairwise chain Markov property* states that for any pair (x, y) of non-adjacent variables with $y \in \mathrm{nd}(x)$ that $x \perp\!\!\!\perp y | \mathrm{nd}(x) \setminus \{y\}$. This expression can be simplified to the form given in (8.2) by simply observing that $\mathrm{nd}(x) = B_1 \cup B_2 \cdots \cup B_i \setminus \{x\}$. Thus the absence of an edge between a pair of variables means that those variables are conditionally independent given all other prior and concurrent variables, where prior and concurrent are defined relative to the later of the two variables.

The *local chain Markov property* states that for any variable $x \in V$ we have that $x \perp\!\!\!\perp \mathrm{nd}(x) | \mathrm{bd}(x)$. Again, we can simplify this relationship by expressing the non-descendants of x in terms of the blocks of the model, giving:

$$x \perp\!\!\!\perp (B_1 \cup B_2 \cdots \cup B_i) \setminus \{x\} | \mathrm{bd}(x),$$

which says, in other words, that each variable x is conditionally independent of all other concurrent and prior variables given its immediate neighbours and parents.

For the undirected independence graphs, the global Markov property made a conditional independence statement about a two general subsets of variables given a third such subset. To make such a similar statement from a chain graph is far less clear and such conclusions cannot be read directly from the chain graph. Therefore,

the global Markov chain property shall not be discussed here - details can be found in Section 3.2.3 of Lauritzen [81], Section 7.2.1 of Edwards [38] or Section 3.6 of Whittaker [125].

For an example of the application of the Markov properties, consider the chain graph in Figure 8.1 from which we can infer several conditional independence relationships. Looking at the second block of the graph we observe that there is no line present joining the variables X and Z indicating some form of conditional independence of these variables. We can hence apply the pairwise Markov property to learn that this conditional independence is of the form $Z \perp\!\!\!\perp X | \{A, B, Y\}$. If we were interested in a single variable alone, Y say, then the local chain Markov property applies and we can determine that $Y \perp\!\!\!\perp A | \{B, X, Z\}$. We can also consider the relationship between variables from different blocks, such as variables B and Z which are non-adjacent. We can apply the pairwise property to learn that $Z \perp\!\!\!\perp B | \{A, X, Y\}$. To refine this property into $Z \perp\!\!\!\perp B | Y$ requires application of the global chain Markov property, which has not been discussed here.

8.1.1.4 Modelling

In order to construct a statistical model that correctly represents the association structure represented in a given chain graph, we need a multivariate response model for each B_i given $B_1 \cup \cdots \cup B_{i-1}$ in which arbitrary sets of conditional independence relations like those in (8.2) hold. To do these we use undirected conditional Gaussian regression (CG-regression) models.

We know from Chapter 5 that an undirected graphical model is modelled by fitting the joint distribution of all variables in the model using a CG distribution. Consider, for example, partitioning the set of variables into two disjoint subsets, a (the covariates) and b (the responses). Then the joint CG model $\mathcal{M}_{ab}$ induces a conditional model $\mathcal{M}_{b|a}$ that describes the distribution of the responses given the covariates. Since our joint model is of the form of a CG distribution, the conditional models (known as CG-regression models) include simple linear regression and logistic regression models as special cases and also generalise to incorporate multivariate responses and covariates which can be discrete, continuous or a combination of both.

It should be noted that the situation where the conditional models have discrete responses and continuous covariates is a more complicated situation as the conditional model cannot be simply determined directly from the joint model; maximum likelihood estimation in these circumstances is computationally more difficult.

In order to determine a suitable choice of model for each block of variables, one can apply a useful simplifying principle: the choice of model for each block is independent of the models chosen for the other blocks. Hence we can consider modelling each block of the chain graph separately. Furthermore, if we restrict ourselves to graphical models, then we can say that the decision to include an arrow from $x \in B_j$ to $y \in B_i$ with $j < i$ depends only on which other arrows pointing to variables in B_i are present, and on which lines between variables in B_i are present, and only on these. Thus the structure of our covariate variables is of no relevance to us when considering the relationships between covariates and responses, and within the responses themselves.

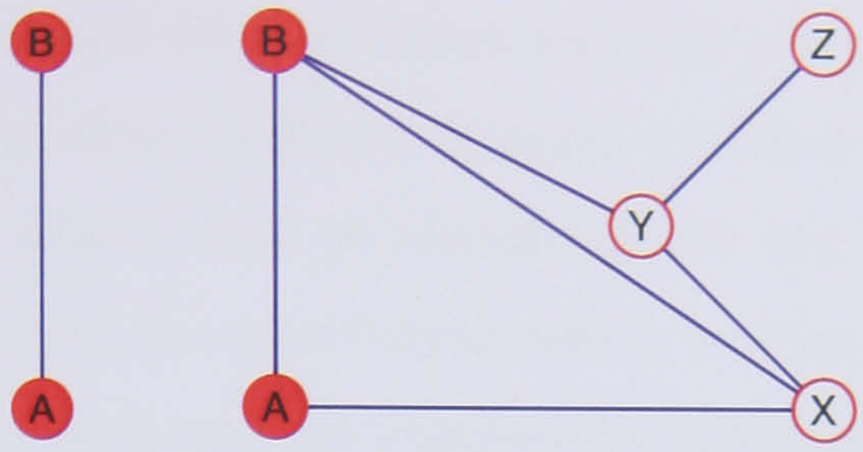
Thus to model a chain graph, we must obtain a sequence of undirected graphical models $\mathcal{M}_i$ representing the distribution of B_i given $B_1 \cup \cdots \cup B_{i-1}$. Since the conditional model asserts no relationships between the covariates, it is appropriate that the associated conditional should contain all interactions between the covariates. This has no effect on the final model choice since the nature of the relationships amongst the prior variables is of no relevance.

Thus the simple chain graph in Figure 8.1 is modelled using two separate undirected graphical models. This is represented in the model formula for a chain graph which gives the formula for each of the component models. So for Figure 8.1, we have that the chain graph formula is:

1. AB,

2. $AB/ABX, BY, Z/ABX, BXY, YZ$.

The corresponding independence graphs for these component models of the chain graph are given in Figure 8.2. We can see that the first component model is the same as the first block in the chain graph. The second undirected model contains the same relationships between $\{X, Y, Z\}$ and the arrows in the chain graph are preserved,



(a) Block 1 (b) Block 2

Figure 8.2: The component blocks of the chain graph in Figure 8.1.

though they are replaced by lines. However, the relationships between the covariates have changed as all interactions between the covariates are now allowed.

8.1.2 Chain Graph Applications

There have been several applications of a graphical modelling strategy to real data that have been published in the literature. Several of these consider medical data with a view to identifying how certain patient factors or environmental factors impact on a particular outcome. Neil-Dwyer *et al* [92] sought to study the associations between patient demographics and operative information with the outcome after aneurysmal subarachnoid haemorrhage. Using a graphical model they sought to assess the associations between the variables with a view to discovering any potential causal pathways between both the demographics and operative information and the patient's outcome. Their study effectively identified that three of these variables had a direct impact on the patient's ultimate outcome and thus could potentially be the causal pathways that they were seeking.

Several other medical studies were performed along these lines, for example Mohamed *et al* [90] utilised a similar methodology to identify the determinants of infant mortality in Malaysia among a set of demographic, environmental and medical variables. They discovered many variables were significantly associated with infant mortality including the prematurity of the infant, the quality of available drinking water, and the level of maternal education. A further medical example is that of the study by Klein *et al* of the occurrence of heart disease and its potential relationships to a range of patient factors including their cholesterol level and blood pressure.

Ruggeri *et al* [110] assessed the short-term outcome of mental health care for 194 patients. Unlike the previous methods, Ruggeri *et al* also included information on the costs of the care with a view to identifying the key determinants of this quantity. They also considered the associations within and between their predictors and outcomes. They also extracted the regression coefficients from the graphical model and exploited the results by Roverato and Whittaker [108] to obtain standard errors for these quantities.

Pigeot *et al* [98] also use this familiar formula of assessing the effects of a number of potentially informative factors on a response to attempt to assess the determinants of that response. The area of their study was the analysis of data on the occupational careers of sociology students. Taking basic demographic information, and details about the students' performance at university coupled with information on the sociologists current occupation, they sought to ascertain the key factors useful for determining the sociologists' current job satisfaction, job adequacy and level of earnings. Rather than fit a graphical model via standard techniques, they instead considered each variable in the model as a univariate regression and then linked the corresponding node in the graph to any variables that, if included, could improve the prognostic power of the regression. They thus built a graphical model on the basis of edge inclusion indicating variables having strong predictive relationships for one another. However, such a model could not be interpreted in terms of the conditional independence statements normally associated with a standard graphical model.

8.1.3 Bootstrapping

Bootstrapping is a technique that can be used to combat a deficiency of chain graph modelling methodology, i.e. the absence of expressions for the standard errors of the model parameters. Bootstrapping has been well documented in the literature [41, 43, 42, 30] and can be applied to a wide variety of situations.

Whilst Efron's original paper discusses the application of the bootstrap in a number of contexts, we shall focus on the basic non-parametric version. Suppose that we have data $y_1, \ldots, y_n$ which are a random sample taken from an unknown distribution, and an estimator of interest $t = t(y_1, \ldots, t_n)$. We now draw R 'bootstrap

samples' from the data - this consists of taking a random sample *with replacement* of size n from the original data. We then compute the estimator for each of these new samples, such that $t_i^* = t(Y_1^*, \ldots, Y_n^*)$ where $Y_1^*, \ldots, Y_n^*$ are the values in the i-th bootstrap sample.

The collection of bootstrap estimates t_i^* can then used to learn about the estimate obtained from the original data. For example, from [42] we can obtain an estimate of the standard error of t, using the result that as $R \to \infty$ the sample standard deviation of t_i^* tends to the standard error of the original estimate t. Furthermore, when R is sufficiently large we can deduce approximate confidence intervals for the parameter t based on quantiles of the bootstrap estimates [30]. Two methods for approximate 95% confidence intervals are:

- Efron's percentile method:

$$[t^*_{0.025}, t^*_{0.975}] \tag{8.3}$$

- Hall's 'basic' method:

$$[2t - t^*_{0.975}, 2t - t^*_{0.025}] \tag{8.4}$$

where t^*_α is the 100α empirical percentile of the bootstrapped estimates t^*. Note that both intervals are of the same width and it is only their location that differs.

Thus we can obtain bootstrap estimates for the standard errors of the chain graph model parameters by fitting the chain graph model to each of the bootstrap samples. The different parameter values obtained by fitting the model to these different data sets will provide us with the bootstrap estimates, t_i^*. Using the results above we can then determine the bootstrap estimate of the standard error and also derive approximate confidence intervals for the parameters.

8.1.4 Regression Evaluation

Having obtained a suitable chain graph model for the data, we will consequently obtain conditional regression models for each of our response variables. We would therefore be interested in investigating the adequacy of these conditional models as they will inform us about suitability of the chain graph model to the prediction of the individual responses. Methods for the evaluation of a multiple regression

are well documented in many statistics textbooks [104]. Since we obtain many regression models for each chain graph model we construct there will be a large number of models we may wish to analyse. Therefore the methods use shall focus on straightforward techniques such as assessment of goodness of fit via the coefficient of determination and residual analysis.

The *coefficient of determination* for a regression model (sometimes called the 'coefficient of multiple determination' for multiple regression) is defined to be the proportion of the variation in our response variable that is explained by the regression model, i.e.

$$R^2 = 1 - \frac{s^2_{\hat{e}}}{s^2_y}, \tag{8.5}$$

where $s_{\hat{e}}$ is the residual standard deviation and s_y is the standard deviation of the response variable. The square root of the coefficient of determination, R, is the multiple correlation between the responses and the covariates. By its definition, R^2 can be used as an indicator for the performance of a regression model, whereby a model with R^2 close to 1 would indicate that the information in the response is almost totally explained by the covariates, and an R^2 close to 0 indicates that the response is very poorly explained by the covariates in the model. Thus R^2 can be used as a summary statistic to indicate the goodness-of-fit of the regression model.

However, the formulation of R^2 as in (8.5) suffers from a key weakness. Each additional variable included into the model cannot increase the value of $s^2_{\hat{e}}$. Since s^2_y is fixed and $s^2_{\hat{e}}$ can only ever decrease, including additional terms will give higher values of R^2, even when the new variables cause the equation to become less efficient. Theoretically, using an infinite number of covariates to explain the response would yield an R^2 value of 1. To compensate for this problem, we use the *adjusted* R^2:

$$\overline{R}^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right), \tag{8.6}$$

where n is the sample size and k is the number of degrees of freedom. By compensating for the sample size and degrees of freedom, $\overline{R}^2$ can decline in value if the contribution to the explained variation by an additional variable is less than its impact on the degrees of freedom.

To investigate the assumption of independent Normal errors we can examine the

distribution of the model residuals. The Normality of the errors can be assessed via inspection of histograms and Normal quantile plots of the residuals - deviation from Normality would indicate a violation of the Normal assumption and would constitute a potential cause for concern over the model's validity. To assess the independence of the errors, a scatterplot of model residuals and the covariate values can be used. Under the assumption of independence, the plot should show random scatter of points about the origin - evidence of trend in the plot would indicate that important information and features of the response were not being captured by the model.

It is also usual in an assessment of a model's adequacy to investigate the importance of the individual coefficients in the model. Typically this is assessed using the idea that $\hat{\beta}/\hat{se}(\beta)$ follows a t distribution, where $\hat{\beta}$ is our parameter estimate. This allows for the construction of significance probabilities for each coefficient. However, since we neither know the standard errors for the parameters nor their distribution we cannot apply this result. Therefore to assess the coefficients we must rely on the bootstrapping methods discussed in Section 8.1.3.

8.2 Construction of Chain Graph Models

8.2.1 Methodology

8.2.1.1 Block Structure

In order to obtain a chain graph model for an orthopædic data set, we must first obtain a partition over the variables in the data set into the sequence of blocks, as discussed in Section 8.1.1. Since we have data with a temporal aspect that follows the schema of the data abstraction in Section 2.1, it is sensible to order our variables based on their temporal sequence, thus reflecting the potential causal direction. Using the temporal structure of the data generalisation in Figure 2.3, we would obtain a chain graph model with at least five blocks. However, conceptually, it may be better to further sub-divide the variables in these blocks into smaller blocks if there still remains an element of temporal ordering to the variables within

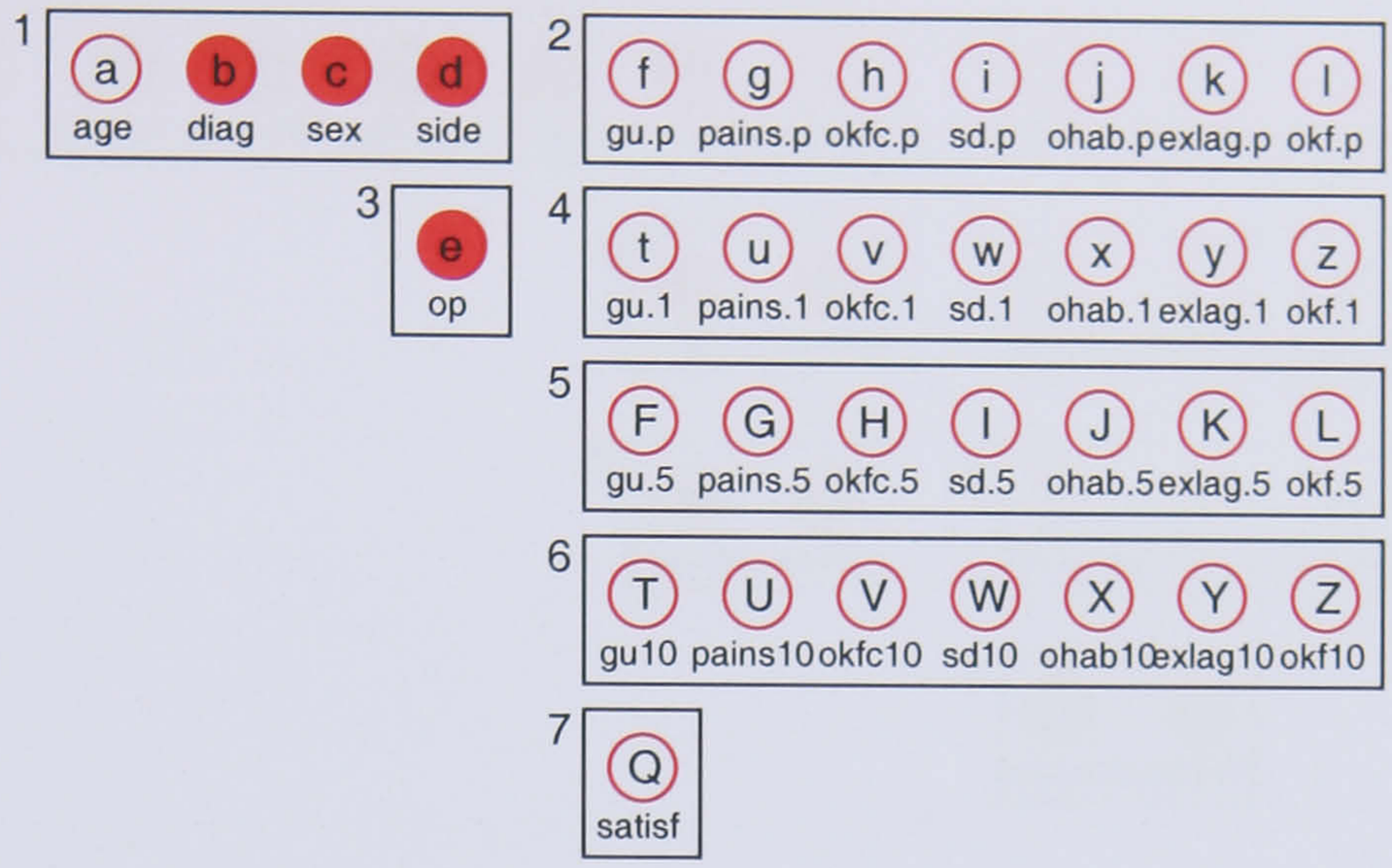


Figure 8.3: Block structure of the knees data.

a single block. For example, patient status measures and patient satisfaction may be recorded at one instance, leading one to place all variables within the same block. If we consider the quantities being measured, we may expect patient satisfaction to be dependent on the patients status and thus to occur after those variables in the temporal ordering of the data. In this case, we would split the block in two with the patient status variables preceding satisfaction to preserve this covariate/response relationship.

The data sets have now been reduced using the methods of Chapter 6 and so we now consider only a subset of the patient status variables at each time point. For both data sets we shall use the subsets returned by the **HT** procedure without using utility information. For the knees data, we obtain a partition into seven blocks. This framework is presented in Figure 8.3. The first block contains four variables, three of which are patient demographics: *Age*, *Sex* and *Side* (i.e. left or right knee). The fourth variable here is *Diagnosis* which is considered to be a known, fixed quantity by clinicians as the distinction between the two conditions is such that the identification of the pathology is assumed to be completely accurate. Thus the prediction of the diagnosis from the pre-operative variables is not an interesting problem in this case, and so is considered to be a pre-determined quantity.

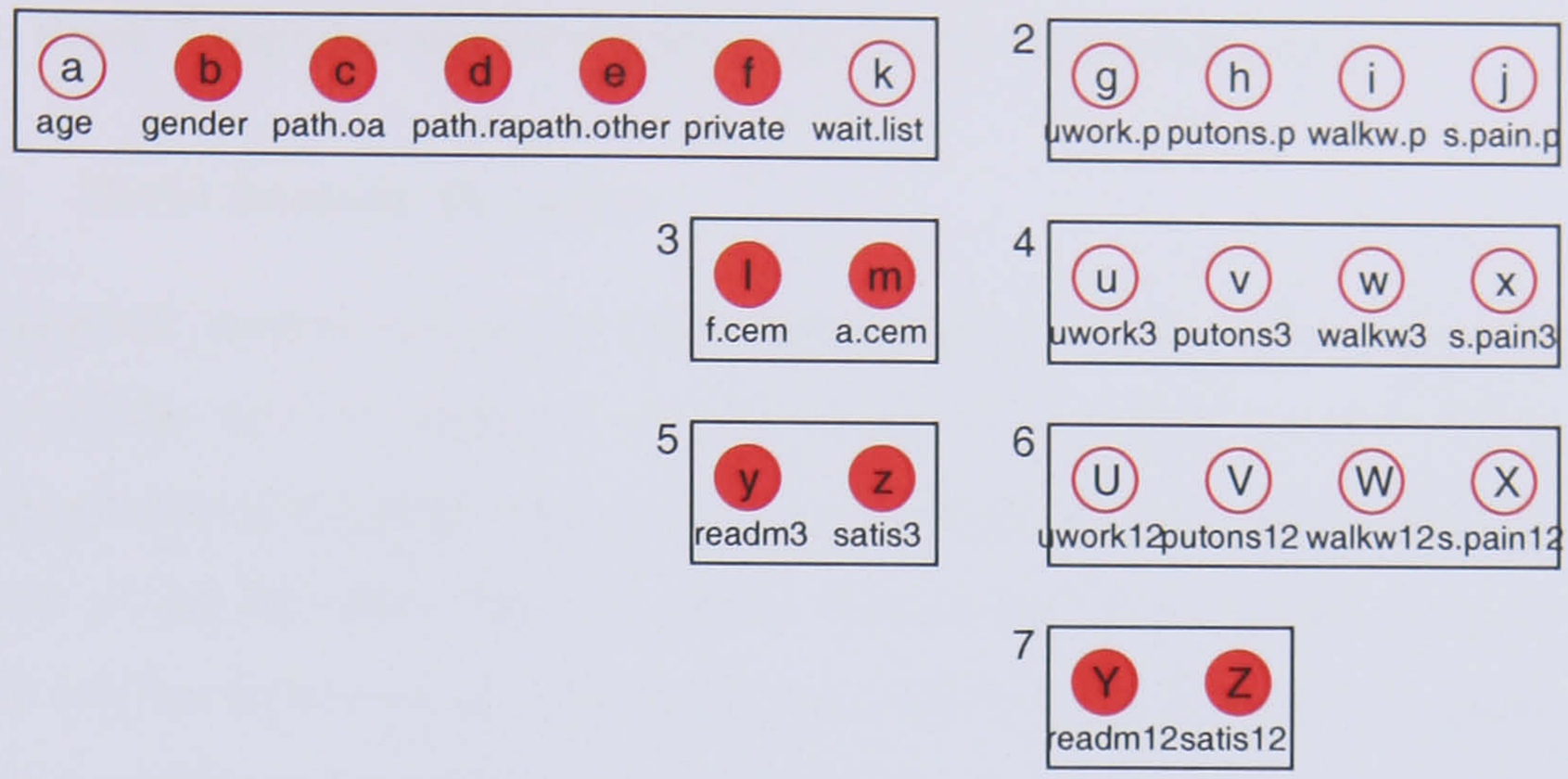


Figure 8.4: Block structure of the hips data.

The second block of the knees data contains the 7-variable subset of the patient status variables that was obtained using the variable selection procedures in Chapter 7. The variables in block 2 correspond to the pre-operative values. The same variables were recorded at later time points and so appear in blocks 4, 5, and 6. The third block is the treatment variable *Operation*, denoting whether cement was used during the joint replacement procedure. This block is placed after the pre-operative patient status variables and before the 1-year post-operative variables. The final block in the framework contains a single variable which denotes patient satisfaction at the 10-year stage. This variable was placed after the 10-year patient status variables rather than in the same block for the reasons discussed above.

Repeating this procedure on the hips data set, yields the block framework given in Figure 8.4. The first block is again composed primarily of demographics (*Age* and *Gender*) and diagnosis details. This block also includes the variable *Private* indicating whether a patient was private or NHS, and the variable *Waiting List* encoding the length of time spent on a waiting list. Blocks 2, 4 and 6 contain the reduced subset of the four patient status measurements recorded pre-operatively, at 3-months and at 12-months post-operatively. Block 3 contains the operation information with two binary variables recording the use of cement during the procedure. Block 5 contains 2 variables the first of which ($\boldsymbol{y}$) denotes whether the patient was

readmitted at 3-months, and the second (z) encodes patient satisfaction at this stage. Block 7 contains similar variables for the 12-month time point.

8.2.1.2 Model Selection Procedure

The approach used to obtain the final chain graph models for the orthopædic data will be similar to that employed with the undirected models. Models will be chosen using forward stepwise selection from the independence model, using the BIC criterion to test for edges that are worthy of inclusion in the model. As previously, models will be chosen via the `stepwise` command in MIM. Models were fitted using the `fit` or `cgfit` commands where appropriate.

It should be noted that now we are working with chain graphs, we must build a model for each of the constituent blocks of the model. To achieve this we consider the variables within a particular block B_i to be the responses and the prior variables in the blocks $B_1, \ldots, B_{i-1}$ to be covariates. As previously discussed, the only factors which affect whether to include an arc in block B_i or an arrow pointing to a variable in B_i are those arcs already within and arrows already pointing into block B_i, the structure of the covariates is irrelevant to the construction of the model. Thus to simplify matters we add all pairwise interactions within the continuous covariates and all pairwise interactions within the discrete covariates but no discrete/continuous interactions. The reasons for omitting these interactions is that to include all of them would render the model unnecessarily over-complex, and to include only some such edges would have implications in terms of the models decomposability as it could introduce 'forbidden paths' into the model (see Section 5.1.5.2 for details). Having thus fixed a structure for the covariate variables, the stepwise selection then considers adding to the model the most eligible arrows from covariate to response or arc between responses. This process is applied to each block in the model and the combination of these undirected models gives us the final chain graph model.

8.2.1.3 Changing Sample Size

As has been previously mentioned, the data sets under investigation have a different sample size at each time point. The number of cases decreases over time due a number of factors. Chain graph models are typically constructed using the complete cases of the data where we have patients with observed data at all time points. To restrict ourselves to working with only complete cases of the data would sorely limit the amount of data available and would discard large amounts of information about the earlier time points.

To combat this problem, we can exploit the fact that the variables in block B_i only depend on themselves and the variables in blocks $B_1, \ldots, B_{i-1}$. Information contained in variables in later blocks is irrelevant at that stage. If we consider building the model for the variables in the first block, B_1, then the only information we need to achieve this is a data set composed of complete observations over the variables in B_1. We can therefore use all cases in the data set to achieve this. It may then be the case that some cases drop out of the data set by the next time point, and so are missing observations for the variables in B_2 and onwards. However, this does not affect the construction of the model over B_1. This method exploits the fact that patients only ever disappear from the data set and never re-appear at a later date. So in order to build the model for B_2, we use only those cases where we have complete observations for the variables in $B_1 \cup B_2$. We can then continue fitting each successive block, B_i, using all cases in data that are complete for the variables up to and including that block, i.e. $B_1 \cup \cdots \cup B_i$. By using this method we ensure that none of the data are wasted and that we exploit all the information that we have available in order to build better models in the earlier blocks.

8.2.1.4 Predictive vs. Explanatory Models

When performing a standard stepwise forward selection process we consider each block in turn, and for each block consider all eligible arcs between variables within that block and arrows pointing to a variable in the block. We then include that arc or arrow which is the most significant, has the greatest negative value of BIC, or is deemed to be the best choice via some other criterion. This approach is sensible when

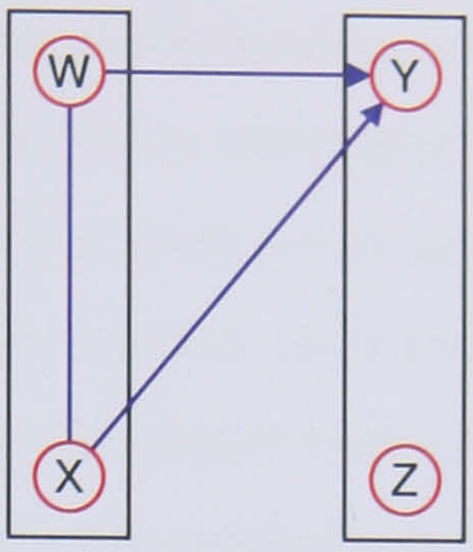


Figure 8.5: A simple chain graph.

we are equally interested in learning about the structure of our response variables and their relationships to the covariates. In these circumstances, it is equally valuable to include an arc between two responses as it is to include an arrow from a covariate to a response.

If, however, we seek to predict our responses from the covariates then the structure amongst the responses may be of less interest to us. For example, consider the simple chain graph in Figure 8.5. Suppose we have arrived at this model through forward selection and at the next stage we determine the arcs $[XZ]$ and $[YZ]$ to be eligible for inclusion, with $[YZ]$ being favoured over $[XZ]$. Under a standard selection procedure, we would add $[YZ]$ as it is the most valuable to us according to our criteria. Having done so, assume now that the significance $[XZ]$ drops and so stepwise selection terminates.

If we were in a situation where the value of Y were missing and we sought to predict it using W, X and Z, then having included this edge $[YZ]$ in the model is valuable to us. If however, the situation was that we knew only the covariates, W and X, and we were seeking to predict the responses, Y and Z, then the inclusion of $[YZ]$ is of little use to us, since we know neither quantity and we seek knowledge of relationships between covariates and responses. In this case, we would have far rather preferred to include $[XZ]$ as it helps us to explain the responses in terms of the covariates. Having an arc between Y and Z would be useful if one of the variables were known and we knew the relationship between these variables. For

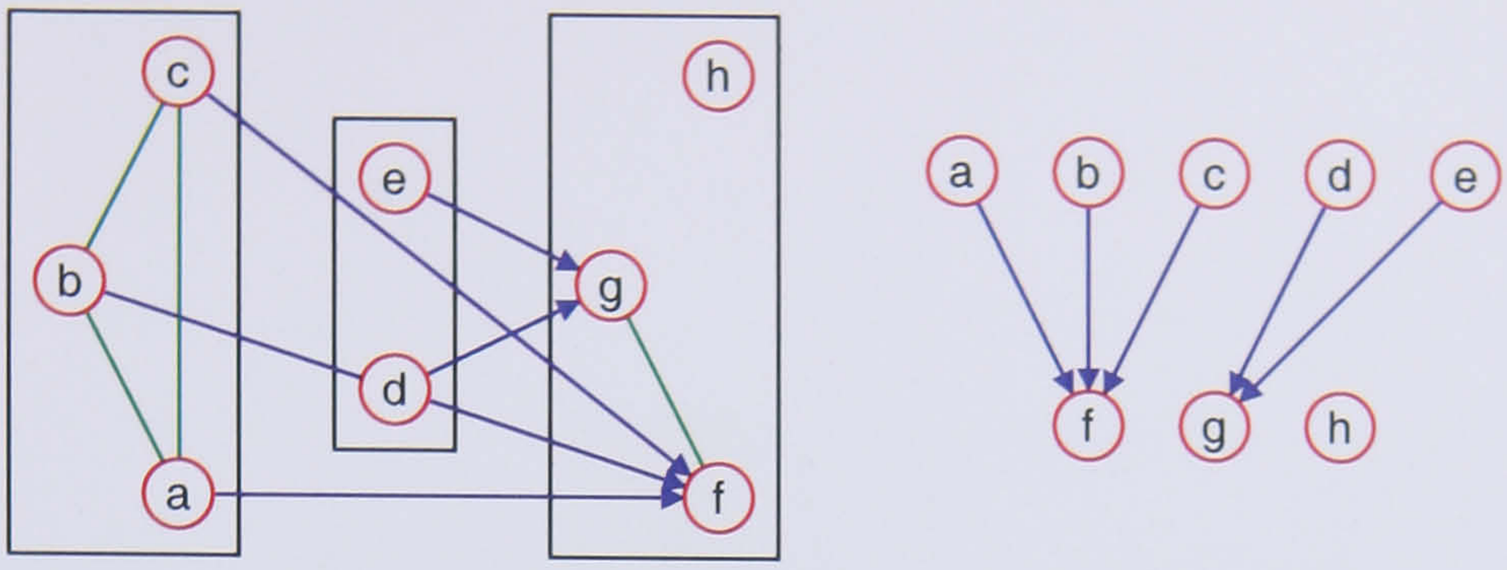
example, if Y and Z were perfectly correlated we could immediately determine one quantity from the other. However, since the data are obtained at different times the variables Y and Z will represent future information as yet unobserved. Therefore, for the purposes of the prediction of these future quantities, the focus is best placed on determining the relationships between the covariates and the responses rather than learning of the associations within the responses.

Indeed, if we were seeking to make predictions for the orthopædic data then we will likely be in this case, as typically all variables within a block will either be known or missing. Thus to make chain graph models that are useful for such prediction scenarios, we can prevent the selection procedure from choosing arcs within the block of response variables and include as many significant arrows as possible. Once no more covariate/response associations can be added, we could then turn our attention to the response structure of the model. Chain graphs constructed in this manner shall be referred to as *predictive chain graphs* or *predictive models* to indicate the differences in the manner of their construction and the focus on this particular prediction setting.

To emphasise this fact that we are only interested in the relationships between covariates and responses we can illustrate this by collapsing the prior blocks into a single line and omitting any undirected edges, and doing the same for the responses. Thus we transform a graph such as that in Figure 8.6(a) into the form in Figure 8.6(b). The top row of variables are the covariates and were split across two blocks, however since their internal structure is of no interest at this time we collapse the blocks and omit their structure thus giving a two-tiered graph with the covariates in the topmost row and the responses beneath. This covariate-response layout for the graph greatly improves its interpretability and emphasises the associations between these two groups of variables.

8.2.1.5 Forbidden Edges and Model Selection

In Chapter 5, it was noted that there were some problems with the model selection procedure ignoring edges that represented significant relationships as their inclusion into the model would render it non-decomposable. The chain graph situation will



(a) Chain graph (b) Covariate-response layout

Figure 8.6: Covariate-response layout for a chain graph.

also suffer from these problems, but to a lesser extent. The reason for this is the structure imposed onto the prior variables of a particular block in the model keeps the discrete and continuous variables separate. This will minimise some of the problems due to the introduction of 'forbidden paths' into the model and may allow the introduction of some edges that were not previously allowed, though it by no means resolves the problem.

8.2.2 Results - Knees Data

8.2.2.1 Complete 10-year Model

The methods described above were applied to the knees data in order to construct a chain graph model for all of the time points in the data set. As the sample size decreased over time, the technique discussed in Section 8.2.1.3 was applied to obtain this model. The chain independence graph for this model is presented in Figure 8.7. Due to the complexity of this model and its associated graph, undirected edges in the model have been coloured green to enable their discrimination from the blue arrows.

Initial impressions of this model are that it is highly complex and that there are a great deal of notable relationships present between variables in the data set. Despite this formidable complexity we can still learn a great deal about the structure of the data. Firstly let us consider the within-block structure, i.e. the structure of variables within the same block as indicated by the green lines. The first block contains

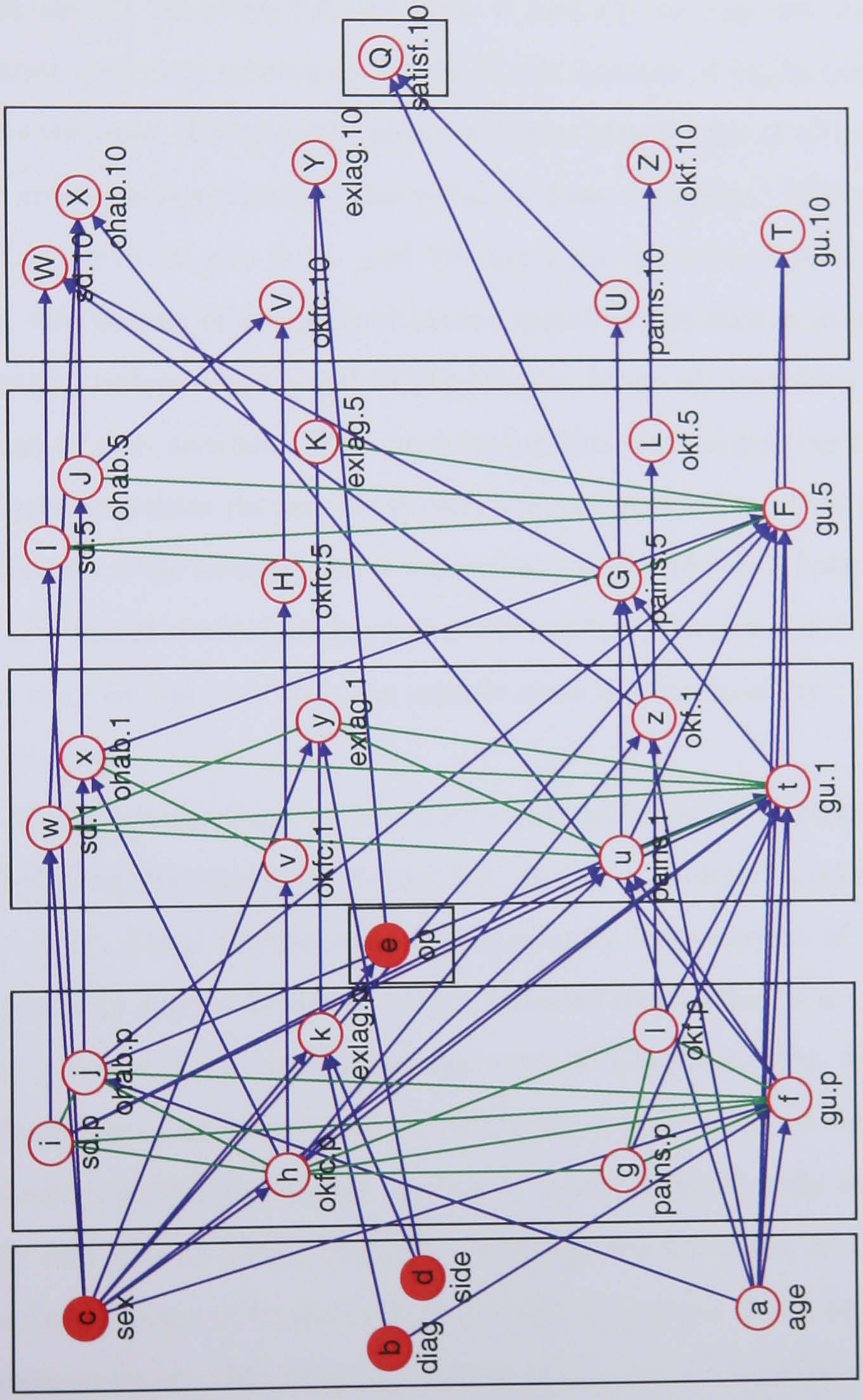


Figure 8.7: The chain graph model for the reduced 10-year knees data.

demographic variables and has no internal structure at all, suggesting marginal independence of these four quantities. This echoes the results from the undirected model in Figure 5.6.

Turning our attention now to the four blocks containing the reduced subset of the patient status variables (blocks 2, 4, 5 and 6), we can see that the amount of association structure present in terms of the number of edges between variables diminishes over time. The pre-operative variables have a total of 10 arcs representing important relationships between the pairs of these variables. This number of arcs drops to 7, 4 and 0 in the 1-, 5- and 10-year post-operative blocks. Thus as time progresses, this subset of the patient status variables appears to grow progressively more mutually independent until at the 10-year stage all variables are marginally independent of one another. The reasons for this are unclear as the correlations between these variables do not drastically change over time. This is likely due to the fact that since the sample size is somewhat smaller at these later time points we would require larger deviations from the independence hypothesis in order to reject it. Thus it may be the case that the sample sizes are too small to detect significant relationships here.

Some of the relationships within the blocks are notably different than what we might expect from the results of Figure 5.6. In the undirected model, *Pain Severity* and *Other Knee Fixed Contracture* are marginally independent of the main block of walking ability scores, however in this reduced model this is no longer the case as they are associated to one another and several other variables. Additionally, the variable *Extension Lag* which was conditionally independent of the other walking ability variables in the undirected models is, again, conditionally independent pre-operatively and also at 5 and 10 years. However, interestingly, at the 1 year point *Extension Lag* appears to be dependent on both *Going Up Stairs* and *Sitting Down.* This appears to be at odds with the undirected model we obtained in Figure 5.7.

Despite the changing structure within the blocks, there are still some common edges to the pre-operative, 1-year and 5-year blocks. These edges are: (*Going Up Stairs*, *Sitting Down*), (*Going Up Stairs*,*Pain Severity*), and (*Going Up Stairs*,*Other Hip Abduction*). The commonality of these relationships suggests that there may

be some fundamental relationships between these quantities that remain over time and despite the intervention of the operation.

It may be desirable in these cases to force the blocks of patient status variables to have the same internal association structure at each time point. This may be sensible from a clinical perspective or if the data display no evidence of changes in their correlation structure over time. In this case we could build a single model over the patient status variables at, say the first time point. We could then fix this structure into the model and replicate it across the various blocks at the appropriate stage in the model building instead of performing a stepwise selection over the internal arcs of the block. Applying this methodology to the knees data set resulted in a notably different model to that shown above. Aside from the internal block structure, the primary distinction was that the number of between-block arcs was less than when we do not impose the common within-block model. This could be particularly problematic as it is the between-block arcs that are predictively useful to us, rather than the within-block arcs. The reduction in the number of these arcs is because their possible inclusion into the model graph depends on the arrows feeding into and the edges within the response block. Thus fixing a particular structure over the responses can cause a reduction in the significance of some covariate/response edges as some of the information can be captured within the responses, as well as making many such edges forbidden under the constraints of decomposability. For these reasons, this common-block approach will not be investigated further.

The variable *Going Up Stairs* shown on the diagram as ($\boldsymbol{f}$, $\boldsymbol{t}$, $\boldsymbol{F}$, $\boldsymbol{T}$) and is a focal point of the first three of the patient status variable blocks indicating that it is associated with many of the other variables in those blocks. This variable was identified in the variable selection process as the most informative about the data, and the fact that we have such relationships present in the model graph reinforces this. Nonetheless, the level of structure within the blocks is quite low - this is not surprising as one of the goals of the variable selection process was to identify a subgroup of variables that explains the structure of the data but has minimal overlap of information between the variables.

Having discussed the relationships within the blocks of the chain graphs, we

can now turn our attention to the relationships between blocks. These are more interesting as they correspond to relationships that exist between time points and so are both informative for prediction of later quantities and could, under the correct circumstances, indicate potential causal paths.

The first block in the chain graph contains the demographic variables, and it is clear from Figure 8.7 that these variables have a notable association with the data observed at later time points since there are many arrows emanating from that block. The variable *Age* (***a***) is associated with patient status variables at each time point, and is associated with all of the *Going Up Stairs* variables (***f***,***t***,***F***,***T***). The variable *Sex* (***c***) has the most associations with patient status variables of all the demographics, this suggests a large number of significant sex differences in terms of the patient status. We can also see that the variable *Side* (***d***), which was previously seen to be marginally independent of all other variables, is now apparently associated with *Extension Lag* (***k***) pre-operatively and at 1-year. However, this relationship might be spurious due to the problems previously noticed with *Extension Lag*. This association is quite interesting as it was not detected with the undirected models. Finally, the variable *Diagnosis* (***b***), which encodes whether the patient has rheumatoid or osteoarthritis, has three associations to later patient status variables indicating significant differences on these quantities. In the undirected model, we observed that the graph was missing a significant arc joining *Diagnosis* with pre-operative *Going Up Stairs* (***f***); this arc is now present in this chain graph model. In fact, the predicted mean value of pre-operative *Going Up Stairs* for the osteoarthritis group is 2.677, whereas the rheumatoid group has a value of 1.931 indicating that patients in the rheumatoid group have a poorer condition pre-operatively. Since there are many other patient status variables associated to *Going Up Stairs*, then these differences due to *Diagnosis* will be evident on these variables via *Going Up Stairs* as an intermediate variable.

Looking at the temporal associations due to the pre-operative variables, we can see that there are many associations between pairs of the same variables separated by time, e.g. from *Sitting Down* pre-operatively (***i***) to *Sitting Down* 1-year post-operatively (***w***). This suggests that the post-operative state of these variables is

dependent on the patient's prior state, which is quite sensible. However, the variables *Pain Severity* (***g***) and *Extension Lag* (***k***) do not have such associations to their later counterparts. The reasons for the lack of associations between the *Pain Severity* variables could be attributable to the intervening operation affecting pain in a way that is irrespective of the prior state; an alternative justification could be that *Pain Severity* is best expressed in terms of the post-operative variables rather than the pre-operative quantities and so a pre-to-post arrow is absent. There are also some arrows beginning in the pre-operative block that point to variables later than 1-year post-operation, such as the arrow from *Going Up Stairs* (***f***) to *Other Hip Abduction* at 10-years (***X***).

The third block contains only the *Operation* (***e***) variable which is associated only with *Extension Lag* at 10 years post-treatment. In terms of the other patient status variables and the other time points, there are no relationships. This suggests that treatment type is not associated to the prior state of the patient (since this is a randomised study), and also that the treatment types are indistinguishable in terms of patient status at 1, 5 and 10 years after the operation (with the exception of *Extension Lag* in the latter case).

At the 1-year time point, we observe that there are are many arrows feeding into *Going Up Stairs* (***t***) and *Pain Severity* (***u***) indicating that they are the variables that are most associated with the patient's pre-operative state. We can obtain the regression equations corresponding to these relationships by performing an appropriate marginalisation of the fitted joint density function. In the case of *Going Up Stairs*, for a male patient with osteoarthritis, we obtain :

$$t = 2.919 - 0.023a + 0.249f + 0.231g + 0.031h + 0.034i + 0.009j + 0.011l,$$

where the variable letters correspond directly to the labelled nodes in the chain graph model. Thus we can see that an increase in age (***a***) is associated with lower levels of *Going Up Stairs* at 1 year (***t***). Further, we can see that better levels of, for example, pre-operative *Going Up Stairs* (***f***) and *Pain Severity* (***g***) contribute to an improvement in the state of *Going Up Stairs* at 1 year. We also obtain the estimate for the standard deviation of t which is $\sigma_t = 0.733$. This standard deviation is rather large considering that t was originally measured on a five-point scale. Such aspects

of the predictive capabilities of the chain graph, and the adequacy and validity of the fitted model are discussed in Section 8.3. In fact there are many of these univariate regression models that we might choose to display and examine - such examination is best suited to computer interaction with the modelling software.

Most of the measurements at 1-year are related to those at 5-years in a similar way as before with many arrows between pairs of the same quantities separated by time. However, we now see that *Pain Severity* at 1 year is associated with itself at 5 years - an association that was not evident between the pre-operative and 1-year values suggesting that the absence of an association previously was due to either high levels of noise or due to the intervention of the operation disrupting such a relationship. A similar story is evident when considering the 5-year block, though we should observe that as with the internal structure of the block the number of arrows feeding into this group of variables has also diminished. We can see that the variable *Going Up Stairs*, $\boldsymbol{F}$, is still the best predicted by prior values as indicated by the relatively large number of arrows feeding into it. We can again consider the regression formula for this variable, $\boldsymbol{F}$, where we restrict ourselves to a male patient with osteoarthritis:

$$F = 3.900 - 0.010a + 0.120f + 0.021j + 0.392t - 0.207u - 0.013x.$$

Now we observe that the intercept in this equation is approximately 1 point larger than the equation for t at 1 year, perhaps suggesting a consistent improvement in patient condition over this time. Again, we can see that the patient's age has a negative association with more elderly patients having reduced levels of *Going Up Stairs*. We can also observe that $\boldsymbol{F}$ is dependent on its prior values both pre-operatively and at 1 year as shown by the presence of terms in $\boldsymbol{f}$ and $\boldsymbol{t}$. Interestingly, we also have a negative coefficient for 1-year *Pain Severity* ($\boldsymbol{u}$), which would seem to suggest that patients with high levels of $\boldsymbol{u}$ (i.e. less severe pain) have lower values of $\boldsymbol{F}$ at 5 years, which seems counter-intuitive

The temporal relationships are again similar when considering the 10-year block with the majority of arrows in the graph linking variables to their prior observations. However, the number of associations is again reduced with still fewer arrows pointing to variables in the 10-year block, and those few that do typically come from a

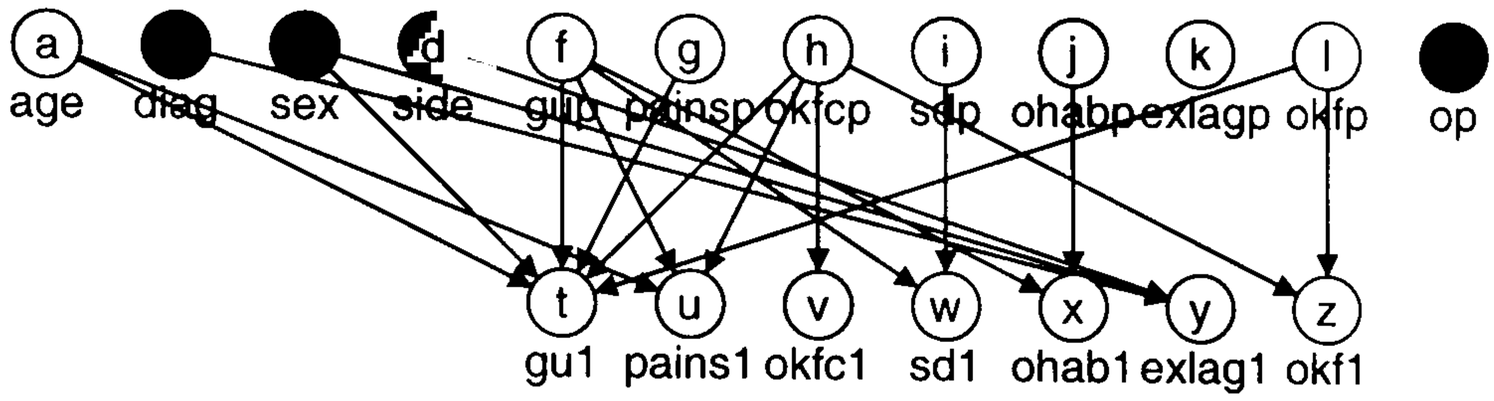


Figure 8.8: The predictive chain graph model for the reduced 1-year knees data.

variable's prior observation at 5 years.

The final block in the graph contains a single variable recording the patient's satisfaction at 10 years after their operation. We observe that satisfaction is only associated to the patient's *Pain Severity* at 5 and 10 years after treatment ($\boldsymbol{G}$ and $\boldsymbol{U}$). The fitted relationship obtained via the model is:

$$Q = 2.937 + 0.216G + 0.212U,$$

which indicates patients with better scores for *Pain Severity* have a higher level of satisfaction. This seems to be intuitively reasonable, and suggests that the best route to patient satisfaction is via an improvement in their pain severity.

8.2.2.2 1-year Predictive Model

As discussed in Section 8.2.1.4, the goal of the obtaining a predictive model is to favour the inclusion into the model of predictively important relationships between the model covariates and the responses. Application of this version of the selection procedure when considering the 1-year post-operative patient status variables as the responses yields the model graph as shown in Figure 8.8.

This model is substantially similar to that obtained in the full joint model as our covariates and responses are the same. However, there are some slight differences and the re-presentation of the model graph does highlight some of the previously discussed features. For example, we can see from the vertical arrows that four of the 1-year variables are directly associated to their pre-operative observations. It is also clearer to see that the variable *Going Up Stairs* ($\boldsymbol{t}$) has the most arrows feeding into it from the prior variables, reflecting its strong associations with many variables. The relationships between the demographic variables *Age* ($\boldsymbol{a}$), *Sex* ($\boldsymbol{b}$)

and *Diagnosis* (***c***) are also easier to read, as is the absence of any relationships to or from *Operation* (***e***). Notably there is no arrow joining *Diagnosis* to any of the response variables. Whilst we saw an arrow in the joint model relating diagnosis to the pre-operative value of *Going Up Stairs*, it would appear that post-operatively there is no difference due to patient pathology.

There were however some slight differences between this model and that obtained in the full joint model. For example, the arc joining pre-op *Other Hip Abduction* (***j***) and 1-year *Going Up Stairs* (***t***) has vanished in the prediction model. The precise reasons for this absence are unclear, however if we refer back to the regression equation obtained for t we observe that the coefficient of j is small (0.009), suggesting that it was not as predictively important as expected from the joint model. Similarly, we now include an arc from pre-op *Going Up Stairs* (***f***) to 1-year *Sitting Down* (***w***) that was not present in the joint model. In the joint model at 1-year, *Sitting Down* (***w***) was associated to *Going Up Stairs* (***t***) and *Pain Severity* (***u***), both of which were associated to pre-operative *Going Up Stairs*. Hence there will have been an indirect association between ***f*** and ***w*** via these intermediaries. However, when we build this predictive model we treat our responses as independent and so the association between ***f*** and ***w*** can no longer occur through the intermediate variables and so the indirect association becomes a direct one.

8.2.2.3 5-year Predictive Model

If we construct another predictive chain graph model as above, though now replacing the 1-year post-operative variables with the 5-year variables we obtain the model with the graph shown in Figure 8.9. Unlike the 1-year predictive model, this graph differs quite substantially from the relationships seen in the joint model. This is due to the fact that we do not include the 1-year data into this particular model and so any indirect associations passing through the 1-year patient status variables will be differently represented.

As mentioned in the context of the joint model, we can see that the number of arrows between the covariates and responses has dropped when compared to the 1-year model. This is likely due to two possibilities, the first being the effect of the

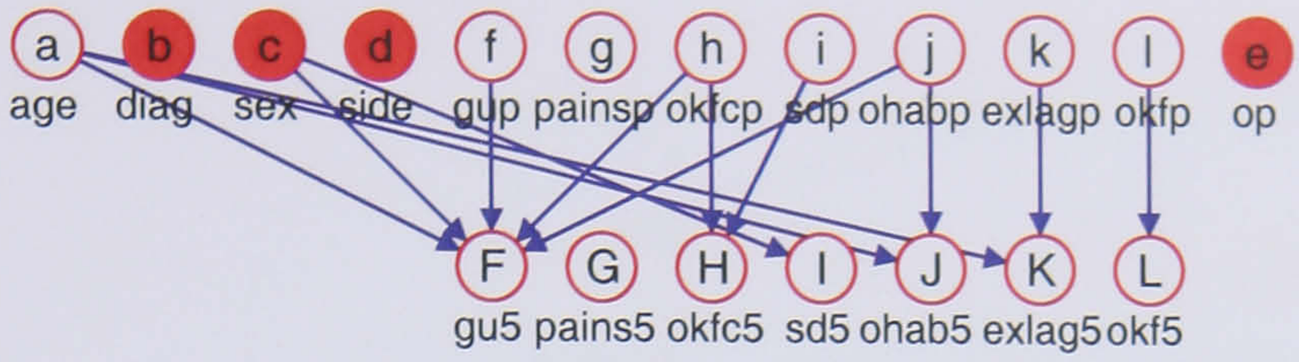


Figure 8.9: The predictive chain graph model for the reduced 5-year knees data.

intervention of treatment and the second being the passage of time causing a divergence from the patient's pre-operative state that increases over time. We can again see several vertical arrows joining pre-operative variables to their 5-year counterparts, though there are fewer than before. *Pain Severity* at 5-years is unassociated with its pre-operative observation since as we saw with the joint model it is best determined by its value at 1 year which is omitted from this model. We also observe that the demographics *Age* and *Sex* have associations with the 5-year variables, but there are no differences due to *Diagnosis*, *Side* or *Operation*.

The variable *Going Up Stairs* ($\boldsymbol{F}$) still has the most arcs and so appears to be the best associated to the pre-operative state of the patient. The appropriate regression equations for $\boldsymbol{F}$ are:

$$\begin{aligned} F_{\text{Male}} &= 4.449 - 0.017a + 0.203f + 0.019h + 0.019j, \\ F_{\text{Female}} &= 3.608 - 0.040a + 0.481f + 0.045h + 0.044j, \end{aligned}$$

In this way we again observe that more elderly patients will be associated with lower scores on *Going Up Stairs*, but good levels on the patient's pre-operative state are associated with similarly good levels of *Going Up Stairs*.

8.2.2.4 10-year Predictive Model

At 10-years, our predictive model has relatively few associations to the patient's pre-operative state, as we can see from Figure 8.10. Only four associations are detected between the pre-operative variables and those at 10-years. This suggests perhaps that since we have performed an intervention and, furthermore, a long period of time has elapsed then the pre-operative state of the patient has become relatively uninformative for us when determining the patients state at 10 years post-

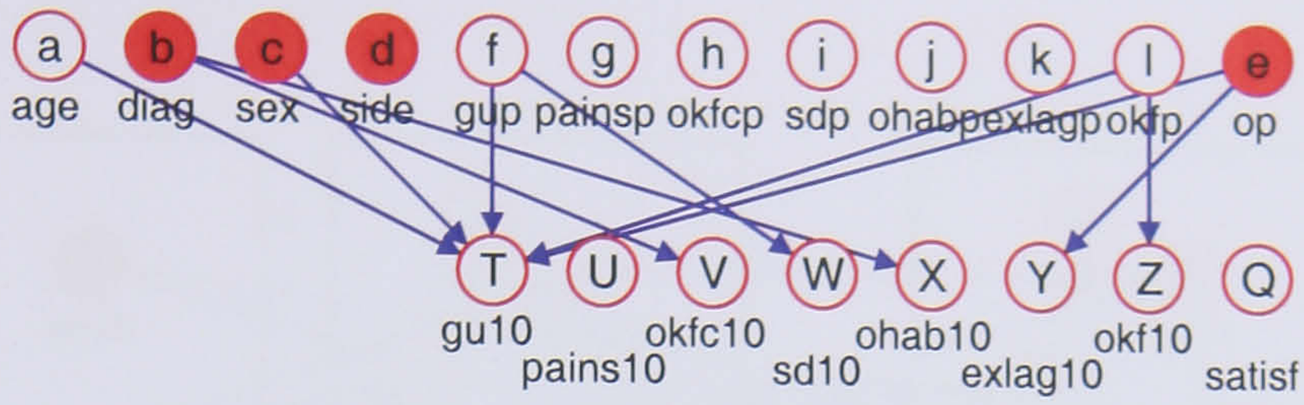


Figure 8.10: The predictive chain graph model for the reduced 10-year knees data.

treatment. Whilst we are detecting various differences and associations due to the demographic variables, the relationships between patient status variables are fairly sparse. Perhaps, we would observe more relationships if we considered the 1-year patient status variables instead of the pre-operative values.

Nonetheless, three out of the four associations between the patient status variables are connecting to *Going Up Stairs* ($\boldsymbol{T}$) at 10-years, which reassures us that this has been a good first choice of variable by the variable selection process. There are also apparent associations between *Diagnosis* and *Operation* and some of the 10-year variables. In the case of the *Diagnosis* associations, caution is advised in interpreting these relationships as the number of patients with osteoarthritis at ten years is 69, whilst the sample for rheumatoid arthritis is only 8. Thus with such a small sample size, these associations could simply be the product of sample variation and more data are required to make firm statements about group differences at this time point. A similar degree of caution is advised with the associations due to *Operation* since both groups have less than 50 cases and, in practical terms, the differences between them seem to be relatively small.

8.2.2.5 Model for Variables Selected Using Utility

During the variable selection process for the knees data we ultimately obtain two distinct subsets of variables - the first was the subset obtained by considering the data alone, and the second was informed by utility information about each variable. The models described above contain those variables identified through the temporal variable selection procedure without using the utility information. If, instead, we chose to use that information and build our chain graphs around this subset instead

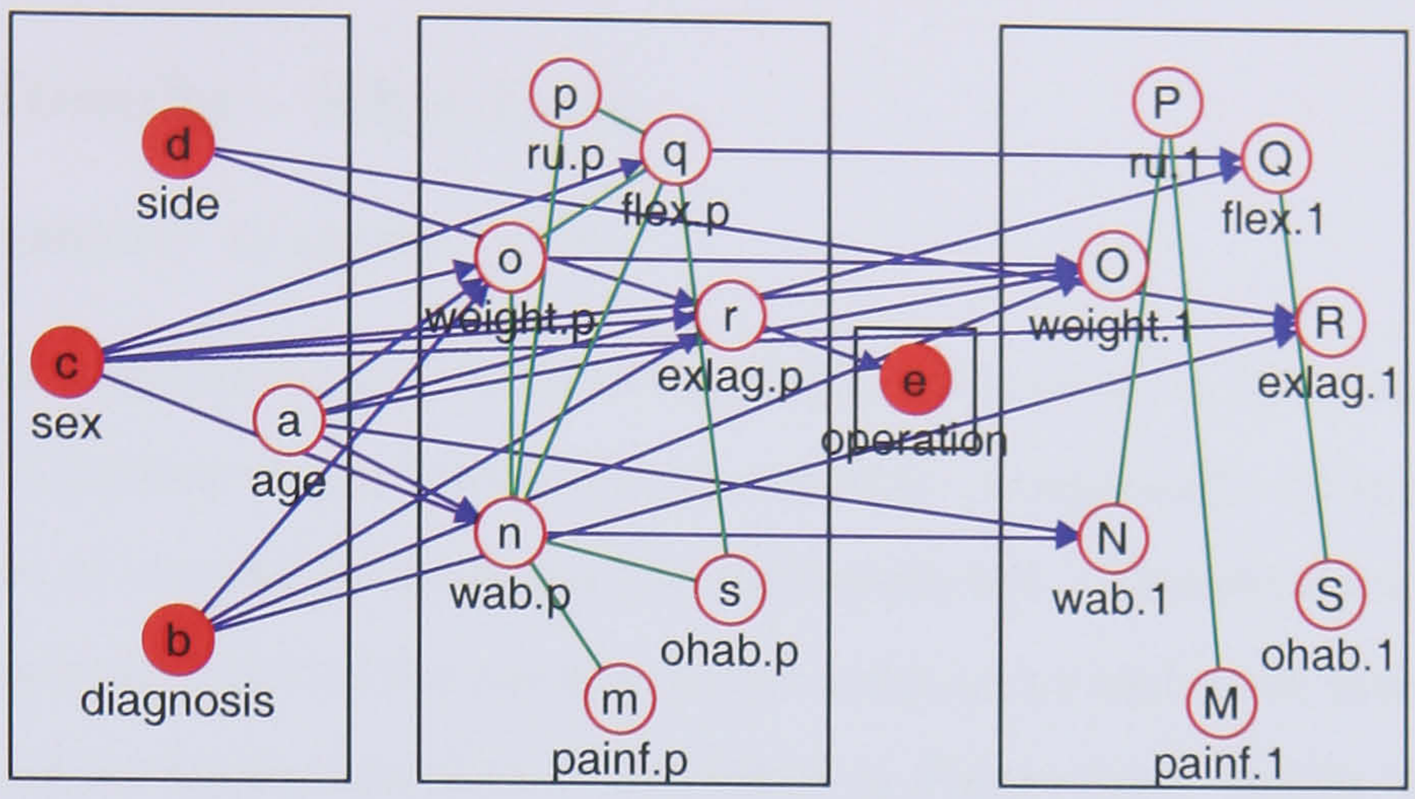


Figure 8.11: The chain graph model for the utility-reduced 1-year knees data.

then models we obtain would be quite different. The chain graph model for the utility-reduced data up to 1-year post-treatment is given in Figure 8.11.

If we consider the internal block structure for the pre-operative and 1-year patient status variables, we observe that there are far fewer associations than are present in the model based on the standard subset of variables. The reason for this is that there are some strong associations within these utility-reduced variables rendering three of the seven conditionally independent of the others at both time points. These three variables are *Pain Frequency* ($\boldsymbol{m}$), *Rising Up* ($\boldsymbol{p}$) and *Other Hip Abduction* ($\boldsymbol{s}$).

There are many associations to the demographic variables within this chain graph model. Closer inspection reveals that the majority of these associations are to *Weight* ($\boldsymbol{o}$) and *Extension Lag* ($\boldsymbol{r}$) at both time points. This further suggests that the three variables above are playing a small role in this model. Furthermore, these three potentially redundant variables are of little use when predicting the 1-year patient state, since there are no arrows either leaving these variables pre-operatively or entering them post-operatively. This suggests that their information is of little use to us if we have observed the other four variables in the group. Thus it is clear that this subgroup is not the best group to work with in terms of the associations within

the data.

8.2.3 Results - Hips Data

8.2.3.1 Complete 12-month Model

The second orthopædic data set concerns hip replacement. It was found in Section 7.3 that the patient status measurements could be substantially reduced due to the high degree of correlation between the measurements. Therefore the hips models include a subset of four of the patient status variables at each time point. The chain graph model for these hips data is presented in Figure 8.12. Initial impressions of the chain graph would suggest the model is less complex than we saw with the knees data, though this is likely only due to the fact we have fewer variables and fewer time points.

Let us first consider the associations present within the individual blocks of variables. Firstly, the block of demographic variables shows us that the three discrete pathology variables are all co-dependent. We also observe an association between *Pathology Osteoarthritis* ($\boldsymbol{c}$) and *Age* ($\boldsymbol{a}$) such that patients with osteoarthritis appear to be older than those without osteoarthritis. The two other demographic variables *Private* ($\boldsymbol{f}$) and *Waiting List* ($\boldsymbol{k}$) are marginally independent of the other variables.

The patient status variables enter the model in blocks 2, 4 and 6 corresponding to pre-operative and 3-month and 12-month post-operative observations. In the pre-operative block all variables are pairwise associated, whereas at three months the edge $[\boldsymbol{vx}]$ is missing from the graph, and at 12 months we only have two associations between the variables. Thus we appear to have a similar decay in the structure of the associations of the patient status variables as we have observed previously with the knees data, though with the hips data the time-scale is far shorter. This could suggest that these variables are becoming less associated to one another as time passes, though the plots of the correlation matrices in Figure 7.8 would contradict this.

The block containing the two treatment variables ($\boldsymbol{l}$, $\boldsymbol{m}$) relating to the use of

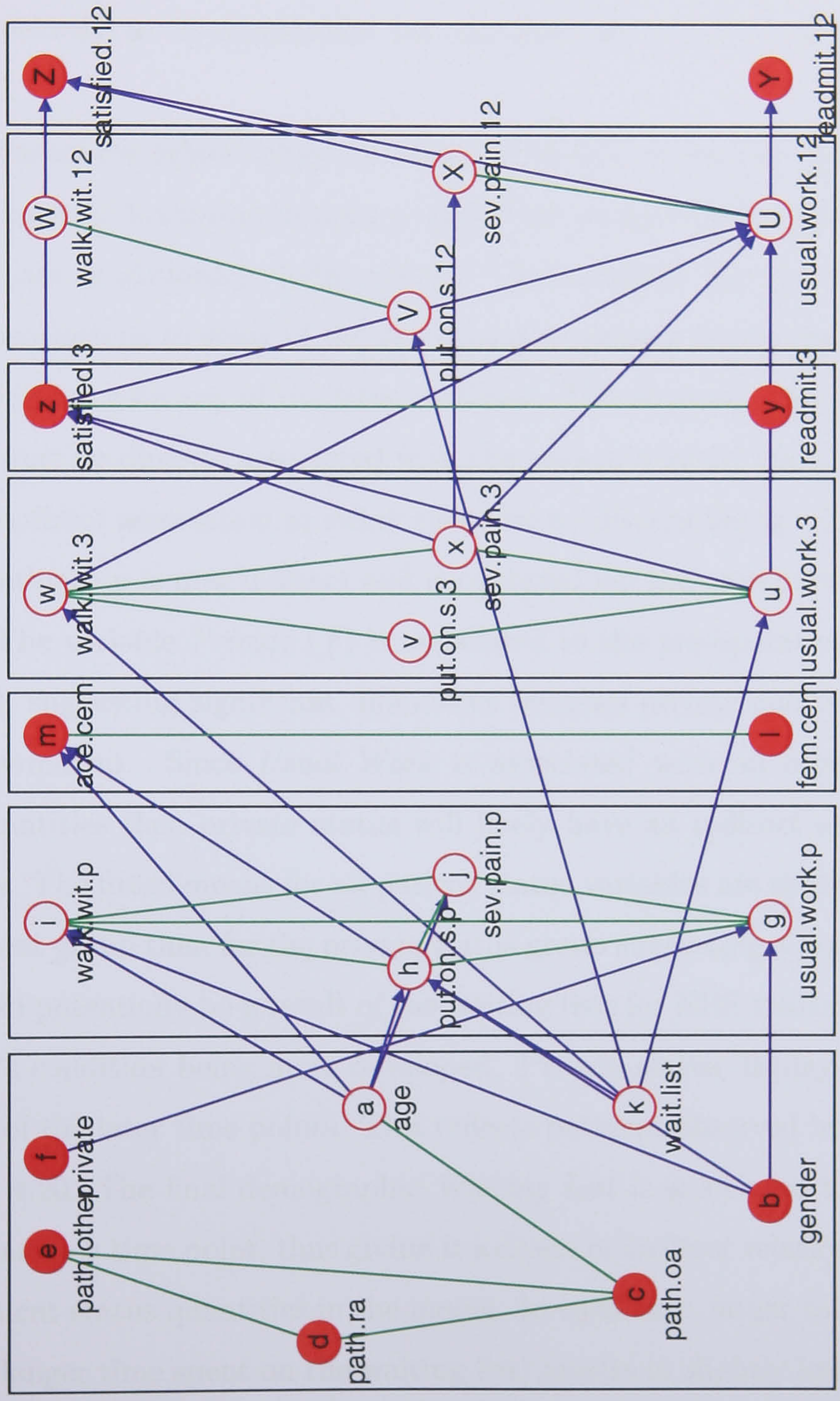


Figure 8.12: The chain graph model for the reduced 12-month hips data.

cement shows that both variables are associated to one another, as we saw in the undirected model in Figure 5.8. Finally, we consider the *Satisfaction* ($\boldsymbol{y}$) and *Readmittance* ($\boldsymbol{z}$) variables which we see are associated to one another at 3 months with patients who were readmitted having a slightly higher probability of low satisfaction. Conversely, at 12 months the two variables are conditionally independent in the model.

Considering the relationships between the blocks, we can see that the pathology variables have no direct associations to any of the patient status variables suggesting that they are conditionally independent. The variables *Age* ($\boldsymbol{a}$) and *Gender* ($\boldsymbol{b}$) display associations to some of the pre-operative patient status variables, but have no relationships with any of the later variables. This suggests that whilst these two variables may be directly associated with the patient's initial state, after treatment there is no direct association so either there are no discernible age or sex differences, or the relationship is now indirect and is captured via the pre-operative state of the patient. The variable *Private* ($\boldsymbol{f}$) is associated to the pre-operative variable *Usual Work* ($\boldsymbol{g}$), suggesting significant differences between private and NHS patients for this measurement. Since *Usual Work* is associated with all other pre-operative status quantities then private status will likely have an indirect association to all quantities. The fitted means for all patient status variables are slightly lower for the NHS patient group than for the private status group suggesting a poorer initial state - this could potentially be a result of the waiting lists for NHS treatment resulting in a patient's condition being more developed. Private status displays no association with any of the later time points. This reflects patterns observed in the profile plot in Figure 4.20. The final demographic *Waiting List* is associated to patient status variables at each time point, thus giving it a direct or indirect relationship with most of the patient status quantities in the model. In each case, larger values for *Waiting List* (i.e. longer time spent on the waiting list) results in slightly lower values of the patient status variables, however the magnitude of this reduction is very small.

The pre-operative block of patient status variables is disconnected from any later variable in the model. This suggests that later patient condition is independent of the patient's initial state, this is likely attributable to a normalising effect of the

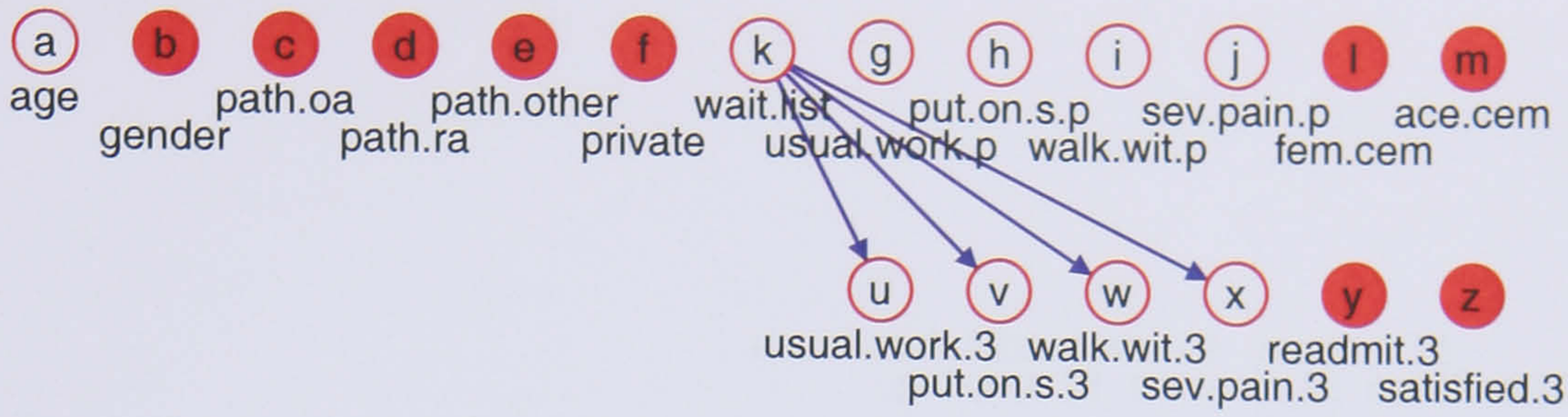


Figure 8.13: The predictive chain graph model for the reduced 3-month hips data.

intervention. The operation types also display no associations with later variables implying that each of the treatment types appear to be equivalent with little difference between patients in any of the treatment groups.

At three months, we note that there are some arrows from the patient status variables pointing to measurements at 12 month, such as those between the *Usual Work* ($\boldsymbol{u}$, $\boldsymbol{U}$) and the *Severe Pain* ($\boldsymbol{x}$, $\boldsymbol{X}$) variables. This suggests that, whilst unrelated to the patient's pre-operative state, the post-operative state of the patient at three months is associated to their state at 12 months.

The patient satisfaction variable ($\boldsymbol{z}$) at 3 months is associated both to *Readmittance* ($\boldsymbol{y}$), and the status variables *Usual Work* ($\boldsymbol{u}$) and *Severe Pain* ($\boldsymbol{x}$) at three months. This seems to be quite sensible that patient satisfaction should depend on their current state and whether they have had to be readmitted to hospital.

At 12 months after treatment, we see relatively few arcs entering or leaving the block of variables suggesting few associations. A consequence of this would be that it would be difficult to make reasonable predictions on the basis of the variables we have available in the current model. The patient satisfaction at twelve months is again associated to *Usual Work* and *Severe Pain*, which suggests they must be key quantities involved in the patient's assessment of their satisfaction with the hip replacement. Satisfaction is also dependent on its previous value at three months.

8.2.3.2 Predictive Models

The predictive models for the hips data are somewhat less interesting than those obtained from the knees data. As we can see from Figures 8.13 and 8.14, there are neither any demographic associations nor pre-operative associations to the response

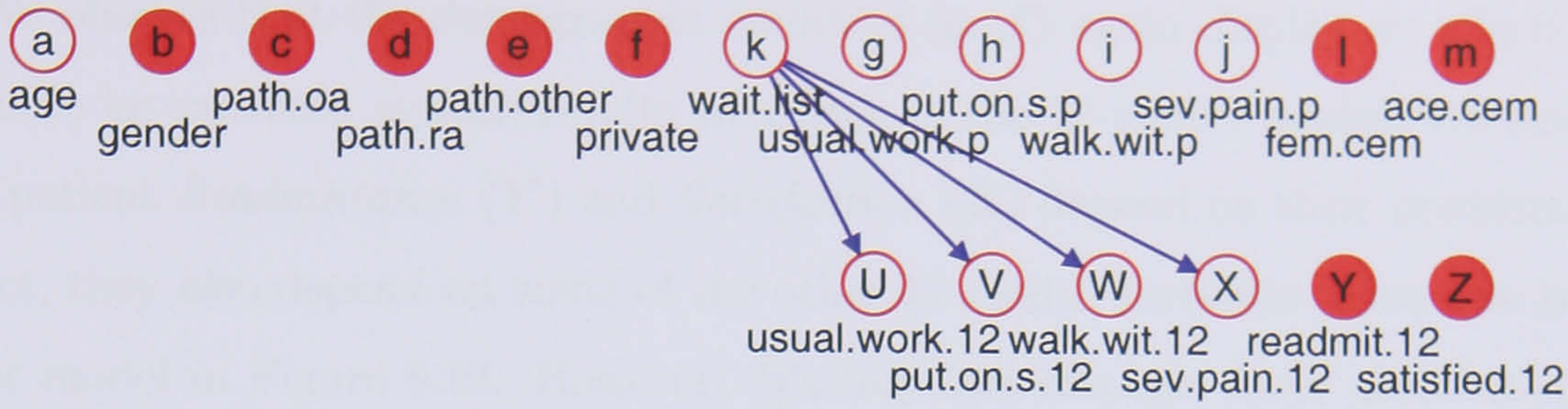


Figure 8.14: The predictive chain graph model for the reduced 12-month hips data.

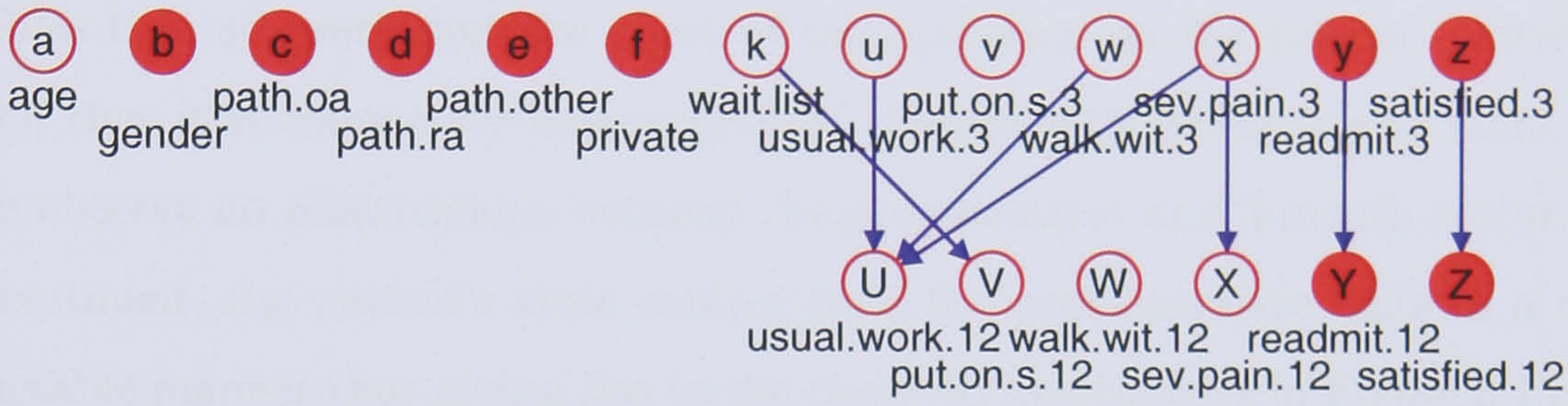


Figure 8.15: The predictive chain graph model for the reduced 12-month hips data given 3-month data.

variables. The only associations we observe are due to *Waiting List* (***k***), whereby longer times spent on the waiting list are associated with very slightly poorer patient states. However, the absence of any associations can still inform us about the structure of the data. We can see firstly that neither the patient's pathology (***c***, ***d***, ***e***) nor the type of their treatment (***l***, ***m***) have any effect on the patient's condition at 3-months and 12-months after surgery. Thus the different treatments appear to be equally effective and all patients respond in a similar way to the treatment regardless of their pathology. We can also observe that none of these demographic or patient status variables are associated with the patient's satisfaction at 3- or 12-months post-treatment. The reason for this is that as we saw in the full joint model, patient satisfaction at a particular time depends only on the patient's state at that time, not on their prior history.

This divorcing of the pre-operative and demographic variables from the post-operative variables could be attributable to the intervention of the operation. If this is the case one may expect to see relationships between the variables at 3- and 12-months. The corresponding graph for such a model is given in Figure 8.15.

We observe that the demographic variables ($\boldsymbol{a}$–$\boldsymbol{f}$) again display no effects and this is to be expected as such results were seen in the 12-month model. We also see that patient *Readmittance* ($\boldsymbol{Y}$) and *Satisfaction* ($\boldsymbol{Z}$) depend on their predecessors; in fact, they also depend on some of the other 12-month variables as we saw in the larger model in Figure 8.12. However, *Waiting List* displays fewer associations to the patient status variables and we see some associations between the 3-month and the 12-month variables. The presence of associations between the patient status variables here suggests that the effect of the operation on the patient's condition is such that it is completely independent of the patient's pre-operative state, and so we observe no relationships between the pre-operative and 3-month states. After treatment, the patient's state evolves from the post-operative state in a more predictable manner thus giving rise to the observed associations in Figure 8.15, but remains independent of the patient's initial condition. Whilst being informative in itself, this relationship is unfortunate as it makes prediction of the patient's future state from their initial state very difficult.

8.3 Prediction from Chain Graphs

Ultimately, the goal of modelling the orthopædic data at each of the various time points has been to predict the future patient state given their current, and typically pre-operative, state. If we are interested in such prediction of these unobserved future quantities on the basis of the past or current data, then to obtain sensible and useful chain graph models we should use the predictive chain graphs described in Section 8.2.1.4 in order to ensure all important predictive/prognostic edges are present in the model.

When we fit the chain graph model to the data set, we obtain a conditional Gaussian distribution representing the joint distribution of all the response and covariate variables in the model. To gain any insight into the nature of the relationships between these two sets of variables, we must examine the values of the parameter estimates obtained through the fitting process. These will typically be in the form of the moments parameters of the distribution $p(i), \mu(i)$, and $\Sigma(i)$. Since we allow the

variance to change freely between cells of the contingency table, we therefore obtain a cell probability $p(i)$, a $(q\times 1)$ mean vector $\mu(i)$ and a $(q\times q)$ variance matrix $\Sigma(i)$ for every cell i in the contingency table that is formed over the discrete variables. This gives a total of $\frac{1}{2}(q+1)(q+2)$ distinct parameters to estimate for each combination of discrete variables. If we have many discrete variables or we have discrete variables with many states, then the number of cells in the contingency table becomes large and the number of parameter estimates even larger. Therefore to investigate these means and variances via mean tables would be a prohibitively complex task, even if disaggregating by choices of subgroups of the discrete variables.

Since the parameter estimates represent the parameters of the joint probability distribution of our data, we can take two steps to simplify our interpretation of the fitted parameters. The first step is that we can condition on the covariates in the model, which will give us the parameter estimates of the distribution of the response variables given the covariates. One of the features of the CG distribution is that its conditional distribution is also conditionally Gaussian. The formulae for obtaining the parameters of the conditional distribution, $f_{A|B}$, of the set of responses A given the covariates B from the moments parameters $p(i)$, $\mu(i)$ and $\Sigma(i)$ of the joint distribution, f_{AB}, are:

$$\begin{aligned} p^{A|B}(i_A) &= \{\Sigma(i)_{BB}\}^{-1}\mu(i)_B y_B - y_B^T\{\Sigma(i)_{BB}\}^{-1}y_B/2 \\ &\quad -\frac{1}{2}\left[\log|\Sigma(i)_{BB}| + \mu(i)_B^T\{\Sigma(i)_{BB}\}^{-1}\mu(i)_B\right] \\ &\quad +\log p(i) - \#(\Gamma\cup B)\log(2\pi)/2 - \log\kappa(i_B, y_B) \qquad (8.7) \\ \mu^{A|B}(i_A) &= \mu(i)_A + \Sigma(i)_{AB}(\Sigma(i)_{BB})^{-1}(y_B - \mu(i)_B), \qquad (8.8) \\ \Sigma^{A|B}(i_A) &= \Sigma(i)_{AA} - \Sigma(i)_{AB}(\Sigma(i)_{BB})^{-1}\Sigma(i)_{BA}, \qquad (8.9) \end{aligned}$$

where y_B are the values of the continuous covariates and $\kappa(i_B, i_A)$ is a constant.

We can see in (8.8) we have an expression for the mean of the response variables, $\mu^{A|B}(i_A)$, as a linear function of the covariates y_B with coefficients given by the values of $\Sigma(i)_{AB}(\Sigma(i)_{BB})^{-1}$. We also obtain an estimate for the conditional variance matrix, which on closer inspection is simply the partial variance matrix of the responses given the covariates obtained from the fitted variance matrix of the joint model, $\Sigma(i)$. The linear regression equations may often contain some terms

with zero coefficients when we include the set of all covariates, these correspond to variables that are unassociated with the responses and so in the fitted distribution are modelled as being independent. Equally the equations may contain other covariates or responses that are of no interest to us. In both cases we can marginalise the conditional distribution of our responses to eliminate these terms. However, care must be taken when marginalising over arbitrary sets of variables as, unlike the conditional distribution, the marginal distribution of the CG distribution is not always Gaussian. This is most obviously true if we are marginalising over discrete variables which are associated with variables in the model as this will leave us with a complex Gaussian mixture distribution.

For a subset of variables A with $B = V \setminus A$, we can use Lauritzen's *weak marginal* [81] $f_{[A]}$ of f which has the properties that it has the same moments as the correct moments of the joint distribution and when B contains no discrete variables (or when the discrete variables satisfy certain conditions) it corresponds exactly to the true marginal distribution. The formulae for the parameters of this weak marginal given the parameters $p(i)$, $\mu(i)$, $\Sigma(i)$ of the joint distribution are:

$$p_{[A]}(i_A) = \mathrm{P}[I_A = i_A] = \sum_{j:j_A=i_A} p(j) \tag{8.10}$$

$$\mu_{[A]}(i_A) = \mathrm{E}[Y_A|I_A = i_A] = \sum_{j:j_A=i_A} \frac{p(j)}{p_{[A]}(i_A)} \mu(j)_A \tag{8.11}$$

$$\begin{aligned} \Sigma_{[A]}(i_A) &= \mathrm{Var}[Y_A|I_A = i_A] \\ &= \sum_{j:j_A=i_A} \frac{p(j)}{p_{[A]}(i_A)} \{\mu(j)_A - \mu_{[A]}(i_A)\}\{\mu(j)_A - \mu_{[A]}(i_A)\}^T \\ &\quad + \sum_{j:j_A=i_A} \frac{p(j)}{p_{[A]}(i_A)} \Sigma(j)_A \end{aligned} \tag{8.12}$$

So via appropriate marginalisation and conditioning of the joint distribution we can obtain equations such as those given in Section 8.2.2, and we can use them as a basis for prediction. If our data were in block-recursive form, i.e. with blocks $B_1, B_2, ..., B_k$, and we seek to predict variables in a later block, B_j, from variables in an earlier block, B_i with $i < j$, then care must be taken when obtaining the prediction equations. It is not appropriate to calculate a full joint model involving

all blocks and then marginalise out variables which are not in $B_i \cup B_j$ for two reasons. Firstly, any indirect associations between variables in B_i and B_j which pass through intermediary blocks will be lost and so potentially important terms in the regression equation may not be present. Secondly, the structure over the response variables may be inappropriate and preventing the inclusion of some predictive relationships. To address this it would be best to construct a two-block predictive chain graph model over only B_i and B_j and then condition on B_i and marginalise if necessary.

8.4 Results and Validation

Having obtained chain graph models for the data sets, we can now consider evaluating the adequacy of the fit of these models. This is a large and complex task as we have a linear equation associated with each of our response variables in each of our models. If we were to consider only the predictive chain graph models this gives us a total of 29 regressions to examine in both data sets. For simplicity we shall focus only on a few of the response variables - for the knees data we shall consider the regressions of the most important variable *Going Up Stairs* at 1 and 5 years and the 10-year *Satisfaction* score; for the hips data we consider only the relationships of the 12-month data to that at 3 months. Similarly, the evaluation of the regression will focus on simple methods such as the visual examination of residuals and the calculation of the coefficient of determination in order to keep the general analytic process manageable. Whilst more advanced techniques, such as calculating deletion residuals, and performing more rigorous model validation methods will provide a more thorough analysis, their computationally intensive nature would preclude their application to such a large number of models.

8.4.1 1-year predictive knees model

For the 1-year knees data model (see Figure 8.8), we can extract the regression equations for *Going Up Stairs* as described above. Since *Going Up Stairs* is associated

with *Sex*, we obtain separate equations for male and female patients:

$$\begin{aligned} t_{\text{Male}} &= 3.447 - 0.025a + 0.267f + 0.233g + 0.028h + 0.011l, \\ t_{\text{Female}} &= 2.781 - 0.031a + 0.327f + 0.285g + 0.034h + 0.013l, \end{aligned}$$

where again the variable letters correspond directly to the labelled nodes in the chain graph model. These equations are slightly different from those given in Section 8.2.2 since they correspond to the predictive models rather than the complete joint model.

We can see that the association between *Age* ($\boldsymbol{a}$) and *Going Up Stairs* ($\boldsymbol{t}$) is a negative one with more elderly patients being associated with lower scores of for both sexes. We can also see that the other variables in the equation, i.e. pre-operative *Going Up Stairs* ($\boldsymbol{f}$), *Pain Severity* ($\boldsymbol{g}$), *Other Knee Fixed Contracture* ($\boldsymbol{h}$) and *Other Knee Flexion* ($\boldsymbol{l}$), all contribute to an improvement in the state of *Going Up Stairs* at 1 year, which suggests that a good pre-operative state over these variables is associated with a good state at 1 year. Both of these relationships seem intuitively sensible in the context of the data.

Unfortunately, a major drawback of the graphical modelling approach in this type of regression analysis is that the standard errors associated with these regression coefficients are unknown. Therefore it is not possible to gauge statistically the importance of each of these coefficients in the regression model. For example, at first glance the coefficients for *Other Knee Flexion* ($\boldsymbol{l}$) seem fairly small in (0.011 and 0.013), however it is unclear whether this is likely due to the fact that that variable is measured on a large scale relative to *Going Up Stairs* rather than an intrinsic insignificance of the coefficient. The unavailability of numerical statements of the significance of these coefficients impedes the analysis of these regressions. Whilst we would expect that these coefficients are significant as the inclusion of their corresponding edges in the graphical model were deemed to be significant the relationship between these edge inclusion tests and the significance of these regression coefficients is not apparent.

	Est.	SE	|Std. Est.|	Efron CI		Hall CI	
Intercept	3.447	0.4810	7.1669	[2.919,	4.837]	[2.057,	3.975]
Age	−0.025	0.0055	4.5570	[−0.040,	−0.018]	[−0.032,	−0.010]
Go Up Stairs	0.267	0.0347	7.6506	[0.227,	0.364]	[0.170,	0.307]
Sev. Pain	0.233	0.0550	4.3240	[0.070,	0.284]	[0.182,	0.396]
OK F.Cont.	0.028	0.0061	4.5802	[−0.004,	0.020]	[0.036,	0.060]
OK Flex.	0.011	0.0020	5.3954	[0.007,	0.015]	[0.007,	0.015]
Sex	−0.666	0.1711	3.8920	[−0.952,	−0.304]	[−1.028,	−0.380]
Sex:Age	−0.006	0.0036	1.6853	[−0.014,	0.000]	[−0.012,	0.002]
Sex:GUStairs	0.060	0.0333	1.8036	[0.005,	0.134]	[−0.014,	0.115]
Sex:Sev. Pain	0.052	0.0226	2.3026	[0.002,	0.090]	[0.014,	0.102]
Sex:OKFC	0.006	0.0019	3.2137	[−0.001,	0.006]	[0.006,	0.013]
Sex:Flex	0.002	0.0014	1.3903	[0.000,	0.006]	[−0.002,	0.004]

Table 8.1: Parameter estimates for the 1-year knees model with bootstrapped standard errors and confidence intervals.

8.4.1.1 Bootstrapping Standard Errors

A possible method for resolving the problem of unknown standard errors is to apply bootstrapping methods discussed in Section 8.1.3. We can use these bootstrap estimates to learn about the distribution of the original coefficients and various properties thereof, such as standard errors and confidence intervals. The results of bootstrapping the regression coefficients for this model are shown in Table 8.1 where we have taken 5000 bootstrap samples.

The results of the bootstrapping process are presented in Table 8.1. The estimates from the two equations have been combined by considering main effect and interaction terms from *Sex*. The main effects (rows 1 to 6) thus correspond to values when *Sex*=Male, and remainder correspond to the differences from these parameter values when *Sex*=Female. The columns of the table contain the original parameter estimates, the bootstrap standard error, the absolute value of the original estimates standardised by the bootstrap error, the value for both Efron and Hall's 95% confidence intervals as given in (8.3) and (8.4) respectively.

Examination of the standardised main effects estimates and the confidence inter-

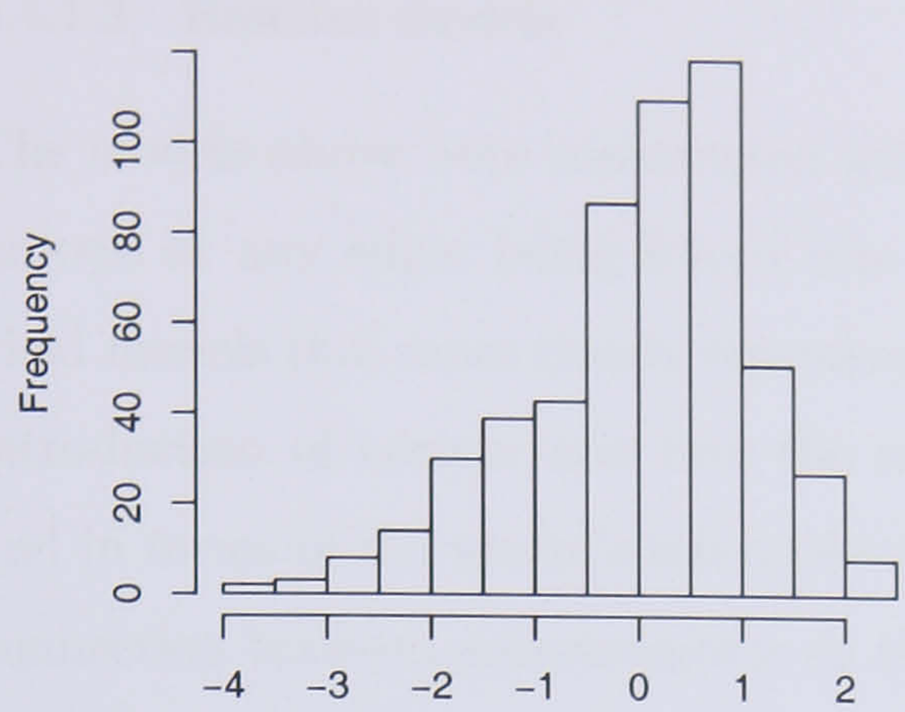
vals for those estimates show that the majority appear significant with large absolute standardised estimates (> 3) and confidence intervals that do not cover zero. The only exception here is perhaps the term for *Other Knee Fixed Contracture*, which despite having a large standardised value one of its confidence intervals contains the origin. This casts some doubt over the importance of this term in the regression. In terms of the interactions, we can see several of these are perhaps non-significant with small standardised coefficients and confidence intervals containing zero. It should be noted here that interaction terms are automatically present in the model for a given set of main effects in order to retain the models graphical property. These terms are not explicitly tested for importance or significance and do not correspond to arcs in the model graph.

In this case the estimates for the standard deviations of t are 0.996 and 1.103 for male and female patients respectively. These standard deviations are relatively large in size, particularly when the variable *Going Up Stairs* is measured on a five-point scale. This suggests that the data are very noisy or there is still a large degree of variability in the residuals that has not been accounted for by the model.

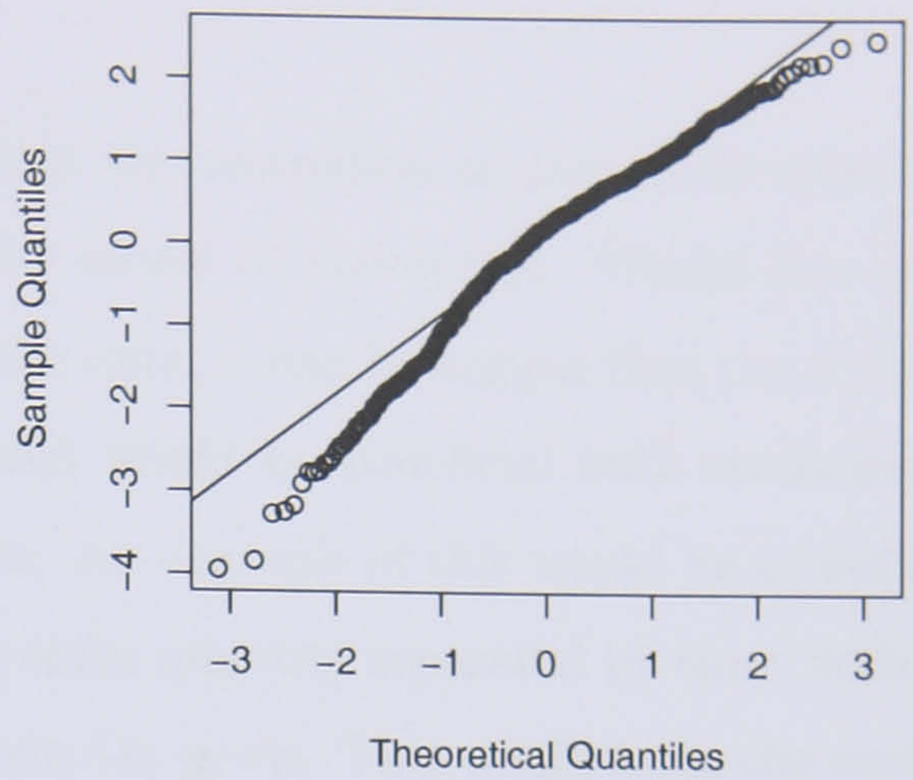
8.4.1.2 Model Evaluation

A histogram and quantile plot of the residuals for this model are shown in Figure 8.16. We can see from the histogram that the residuals are skewed with more positive residuals than we would expect under Normality and a long tail to the left. The quantile plot corroborates this skewness, though the residuals could be assumed to be approximately Normal. Irrespective of this however, it is clear that the distribution of the residuals is not ideal and is likely due to the fact that the response variable is recorded on a 5-point Likert scale which was approximated as continuous. The apparent smoothness of the quantile plot is mainly due to the fact that we include continuous covariates such as *Age* in the regression model. If these were not present the departure from Normality would likely be more severe.

In terms of the goodness of fit of the regression, we can calculate the proportion of variance explained by the model. In this case we obtain a value of $R^2 = 0.2889$, which informs us that only approximately 30% of the variation of the original response



(a) Histogram of residuals



(b) Normal Quantile plot of residuals

Figure 8.16: Histogram and quantile plot from model of *Going Up Stairs* at 1-year.

variable is explained by the predictions from the model. This is quite low and it is not directly obvious whether this is due to the model being poor, or the data being noisy and inherently difficult to predict. To determine which of these cases is most likely we could try fitting a larger model with all of the pre-operative covariates. This would represent the best linear fit we could obtain using the available data. The value of R^2 for this model could then be used as an approximate baseline against which we could compare the performance of the model under consideration. This larger regression model has an R^2 value of 0.3150 and so represents the best proportion of the variation that we could capture in a graphical model for the response and all seven covariates. This is close to the value we obtained from the original model, so this would suggest that the data are difficult to predict on the basis of a linear model over the pre-operative variables and further that adding extra terms provides only small improvement. The original R^2 value can also be adjusted to compensate for the fact that the larger model includes more terms in the regression equation, thus allowing for a comparison on a more even footing. In this case the adjusted values are $\overline{R}^2 = 0.2727$ for the original model and $\overline{R}^2 = 0.2928$ for the larger model. This shows that both models have a similar level of performance though the larger model still performs slightly better, but the margin of the difference between the two is small.

8.4.1.3 Baseline models

The models above were constructed without any restriction on the model selection process or any edges being forced into the model at the outset. Whilst this may yield models that more closely represent the data, it can be argued that the a priori introduction of certain arcs into the model would be beneficial both statistically and in terms of the model's interpretation. An example of this would be to force a connection between measurements on the same quantity separated by time, such as *Pain Severity* pre-operatively and at 1 year, i.e. $\boldsymbol{g} \rightarrow \boldsymbol{u}$. This would force the earlier observation to act as a baseline for the later quantity and other arcs would then be added into the model if they made an additional contribution.

In practical terms this will result in models that are somewhat different in appearance, however application of this technique did not give any appreciable improvement in performance. For the 1-year model discussed above, if we consider the variable *Pain Severity* we obtain a model with connections to pre-operative *Other Knee Fixed Contracture* ($\boldsymbol{h}$), and *Pain Severity* ($\boldsymbol{g}$) which replaces the connection to *Going Up Stairs* ($\boldsymbol{f}$). The performance of the two models are both very similar and equally poor with $\overline{R}^2 \simeq 0.01$ for both models. This pattern was repeated for all of the variables in the 1-year model. Whilst the introduction of arcs to represent baseline measurements into the model is on a sound statistical footing, the benefits here appear to be negligible for these data. However, this method should not be discounted entirely as it likely that introducing baselines will be beneficial for other data sets.

8.4.2 5-year predictive knees model

For *Going Up Stairs* at 5 years, the situation is quite similar to that for 1-year. Again, due to a sex dependence we obtain separate regression equations for the different sexes though this time we have fewer covariates in the equations:

$$\begin{aligned} F_{\text{Male}} &= 4.559 - 0.018a + 0.209f + 0.006h + 0.020j, \\ F_{\text{Female}} &= 3.876 - 0.043a + 0.493f + 0.014h + 0.047j. \end{aligned}$$

Again we observe a negative association with *Age* ($\boldsymbol{a}$) suggesting a poorer pa-

tient state for more elderly patients. We also retain a positive association to the pre-operative values of *Going Up Stairs*, *Other Knee Fixed Contracture* and *Other Hip Abduction* suggesting that good patient performance on these variables pre-operatively is associated with good patient states at 5 years. In terms of the associated standard deviations for *Going Up Stairs* ($\boldsymbol{F}$) for the different sexes, we find these values to be 0.664 and 1.566. Unlike the values at 1-year, this appears to suggest that there is notably less variability in the Male subgroup, leaving the Female group harder to predict.

8.4.2.1 Bootstrapping Standard Errors

The results from bootstrapping these regressions are presented in Table 8.2. Inspection of the estimates and confidence intervals demonstrate that the coefficient for *Other Knee Fixed Contracture* is not a significant term in the regression. This is illustrated by the small value of the standardised coefficient and by both confidence intervals covering zero. The reason for the inclusion of this term in the model is unclear as both *Other Knee Fixed Contracture* and its interaction with *Sex* appear non-significant and so have little notable contribution to the model. Additionally, the main effect terms for *Sex* itself has a relatively low standardised value and one of the associated confidence intervals contains zero suggesting a possible low significance here. However, since *Sex* and all its interactions are essentially considered together in the examination of an edge connecting *Sex* with *Going Up Stairs* and there are a number of highly significant interaction terms then this is likely the reason for its inclusion in the model.

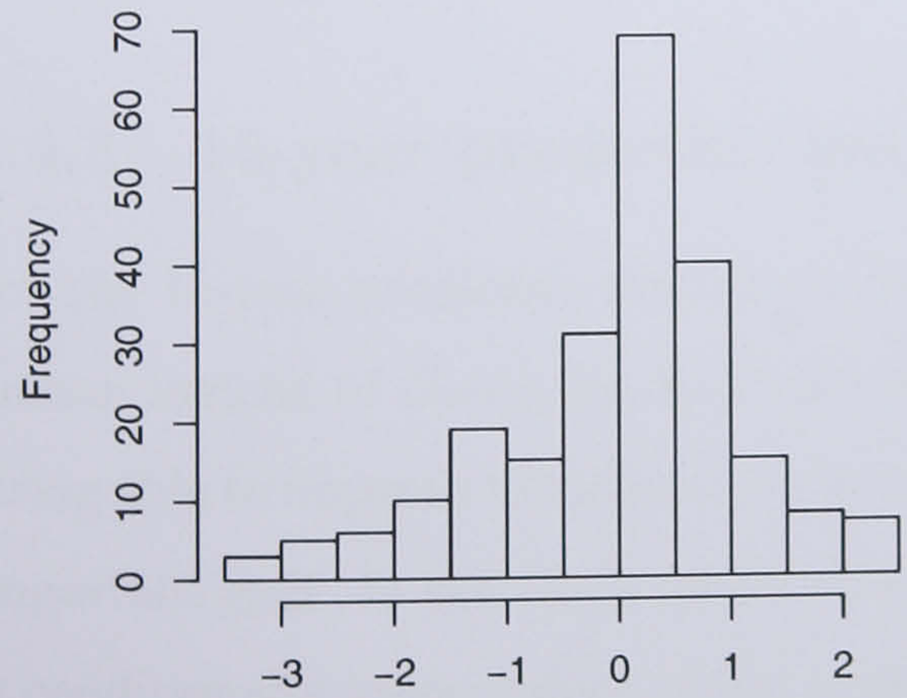
8.4.2.2 Model Evaluation

An examination of the distribution of the residuals via a histogram and quantile plot is presented in Figure 8.17. We can see again a slight skewness to the residual distribution and a curvature to the quantile plot which suggests deviation from Normality. The situation appears to be, again, quite similar to the 1-year results.

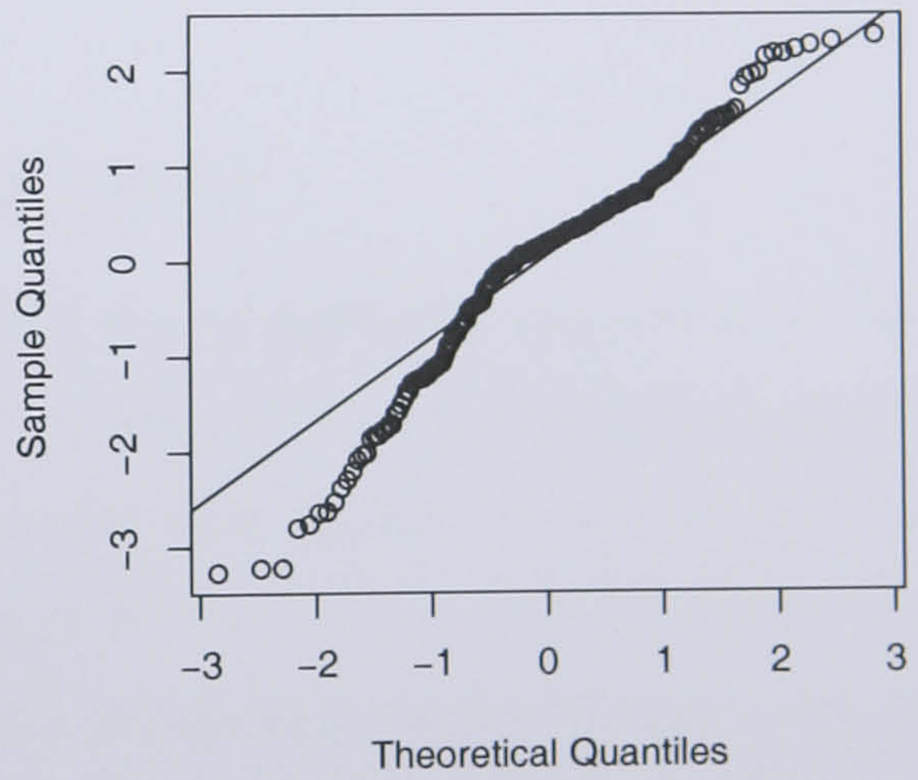
The proportion of variation explained for this model is $R^2 = 0.3870$ which indicates that roughly 40% of the variability of the response variable is being explained

	Est.	SE	|Std. Est.|	Efron CI	Hall CI
Intercept	4.559	0.4423	10.3085	[3.686, 5.429]	[3.689, 5.432]
Age	−0.018	0.0061	2.9447	[−0.032, −0.008]	[−0.028, −0.004]
Go Up Stairs	0.209	0.0448	4.6671	[0.125, 0.303]	[0.115, 0.293]
OK F.Cont.	0.006	0.0071	0.8410	[−0.006, 0.022]	[−0.010, 0.018]
OK Flex.	0.020	0.0069	2.9146	[0.009, 0.035]	[0.005, 0.031]
Sex	−0.683	0.6000	1.1383	[−1.8161, 0.562]	[−1.928, 0.4501]
Sex:Age	−0.025	0.0082	3.0570	[−0.043, −0.011]	[−0.039, −0.007]
Sex:GUStairs	0.284	0.0559	5.0833	[0.174, 0.395]	[0.173, 0.394]
Sex:OKFC	0.008	0.0095	0.8417	[−0.009, 0.029]	[−0.013, 0.025]
Sex:Flex	0.027	0.0071	3.8010	[0.014, 0.041]	[0.013, 0.040]

Table 8.2: Parameter estimates for the 5-year knees model with bootstrapped standard errors and confidence intervals.



(a) Histogram of residuals



(b) Normal Quantile plot of residuals

Figure 8.17: Histogram and quantile plot from model of *Going Up Stairs* at 5-years.

by the model. In comparison with the 1-year model, we can see that despite having fewer terms in the regression equation we are capturing 10% more of the variability of the data suggesting that the 5-year state is better predicted by the pre-operative data. For comparison, the R^2 value for the prediction model including all seven covariates is 0.4065. This larger model only captures slightly more variability and again appears to indicate that it remains difficult to accurately predict the response using the pre-operative variables as covariates. However, if we consider the adjusted R^2 values we find that the original model has a value of 0.3587, whilst the parent model scores 0.3584 suggesting that, in fact, the two models have a similar level of performance.

Additionally, since we are predicting the patient's 5-year state on the basis of their pre-operative state alone we could, no doubt, improve our performance if we also knew and included the patient's state at 1-year post-operation. This would require the construction of a new chain graph model including both pre-operative and 1-year values as covariates. However, the accuracy and performance of the predictions will likely be improved since the post-operative states are more closely related to one another than they are to the pre-operative state. However, since our aim is typically to predict from pre-operative information performing such predictions may not be possible.

8.4.3 10-year predictive knees model

For the 10-year predictive model, we sought to investigate the regression of *Satisfaction* instead of *Going Up Stairs* as this quantity is of great interest to clinicians. Being able to improve the long-term satisfaction of the patients is considered to be an important goal. In the chain graph in Figure 8.10 we can see that *Satisfaction* ($\boldsymbol{Q}$) is conditionally independent of the patient's pre-operative state. Indeed, according to the model, the only variables associated with *Satisfaction* are the patient's *Pain Severity* at 5 and 10 years. The corresponding equation underlying this modelled relationship is given below:

$$Q = 2.937 + 0.216G + 0.212U,$$

	Est.	SE	Std. Est.	Efron CI	Hall CI
Intercept	2.937	0.7445	3.9448	[1.526, 4.251]	[1.623, 4.348]
Pain Sev. 5Y	0.216	0.1292	1.6721	[0.010, 0.480]	[−0.048, 0.422]
Pain Sev. 10Y	0.212	0.0791	2.6812	[0.078, 0.392]	[0.032, 0.346]

Table 8.3: Parameter estimates for the 10-year knees model with bootstrapped standard errors and confidence intervals.

where $\boldsymbol{G}$ and $\boldsymbol{U}$ are *Pain Severity* at 5 and 10 years respectively. We can see that higher (i.e. better) levels of both quantities result in a better satisfaction score. The associated standard deviation here is 0.429, which is a reasonable value suggesting 95% of our residuals lie in the region ± 0.858 implying that predictions are usually accurate to within 1 point of the satisfaction score.

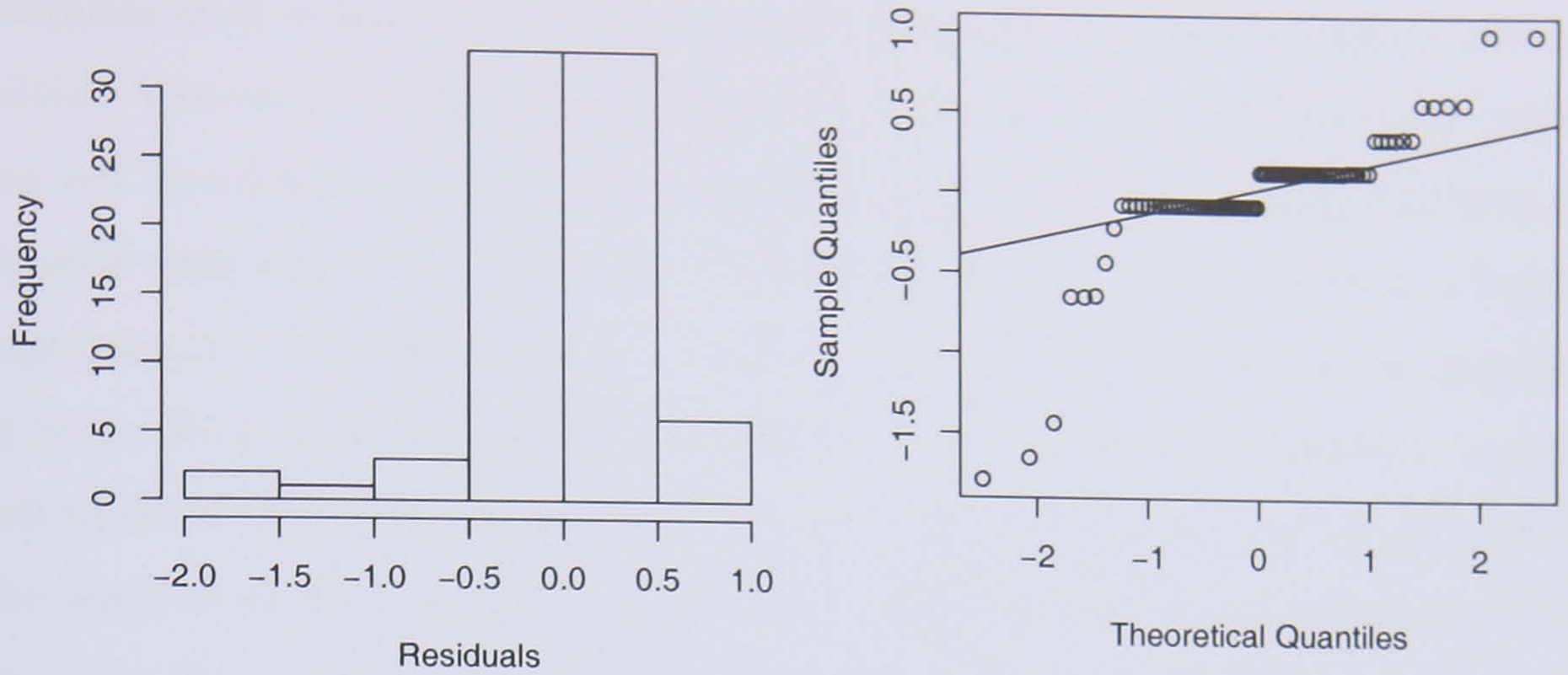
8.4.3.1 Bootstrapping Standard Errors

The results from the bootstrapping of these parameter values in Table 8.3 show that two of these terms in the regression appear to be significant, but *Pain Severity* at 5-years appears to be more questionable with a relatively low standardised estimate value and a confidence interval covering zero.

8.4.3.2 Model Evaluation

If we now examine the residual distribution for this variable we observe a somewhat different situation to that previously seen. The histogram and quantile plot are shown in Figure 8.18. The histogram again appears skewed in a similar way to those seen above with a long tail to the left; though since we have less than 40% of the cases present at 5 years in the 10-year model the detail of the distribution is more coarse and the similarity is not quite as apparent. The quantile plot however is noticeably different - the clear step pattern is reminiscent of the quantile plots in Chapter 3. This will be due to both covariates and response in this case being originally ordinal and approximated by continuous quantities and so were not originally Normally distributed.

The goodness of fit of the model in terms of the proportion of variation explained



(a) Histogram of residuals

(b) Normal Quantile plot of residuals

Figure 8.18: Histogram and quantile plot from model of *Satisfaction* at 10-years.

by the model is given by $R^2 = 0.2968$, and the adjusted value is $\overline{R}^2 = 0.2682$. This value is of a similar order as those obtained previously which seems a reasonable result given that these data appear inherently difficult to predict. For comparison, the graphical model where all of the pre-operative, 1-year, 5-year and 10-year variables were associated to *Satisfaction* were calculated. The resulting model gave a value of $R^2 = 0.8487$, which is relatively large compared to the original model. However, the adjusted R^2 value for this parent model is only $\overline{R}^2 = 0.3867$, showing that the standard R^2 value has artificially inflated the performance by the sheer number of parameters in the model, which highlights the importance of making this adjustment. Nonetheless, we can see that the addition of an extra 26 parameters has only resulted in improving $\overline{R}^2$ by around 0.12. This suggests that despite being less prognostically powerful, the original model given above does appear to be a suitable yet parsimonious choice.

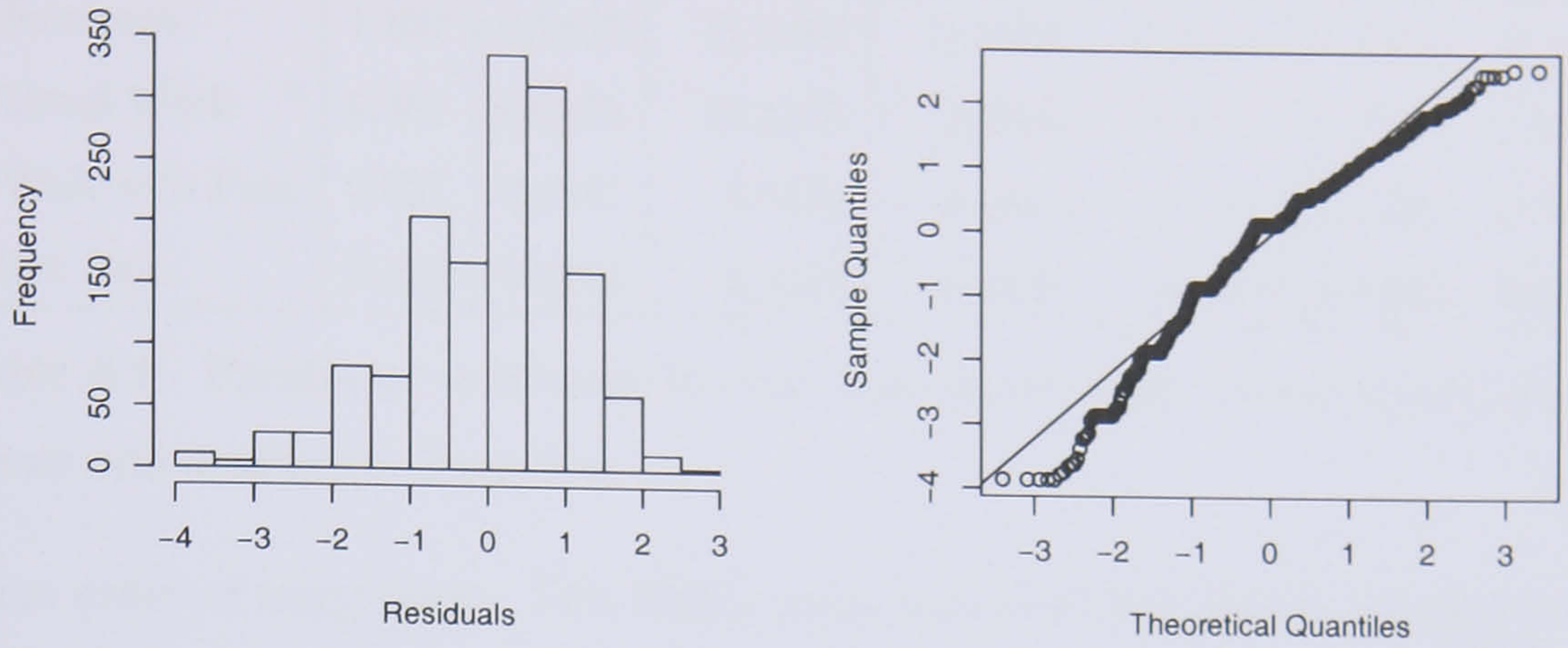
It is clear from the quantile plot in Figure 8.18 that such an analysis of these data is not appropriate. The problem here is that the data were originally ordinal, but were subsequently treated as continuous quantities due to constraints on the analysis. The variables were then analysed as if they were continuous, and the regressions and the residual analyses assume this. However this assumption is invalid in cases such as this. When considering data which are discrete, the analysis of the

residuals from a particular model and the diagnosis of model adequacy are not trivial. Various texts have addressed these problems, [1, 85], however those results are not directly applicable. The reason for this is that the existing methods for discrete data rely on us having fitted particular discrete models, such as a logistic regression or a proportional odds model. These methods then assess the goodness of fit in the context of the discrete model by examining various quantities - e.g. the cell probabilities or the associated log odds - quantities which are meaningless in the world of continuous data. Therefore, it would be equally inappropriate to try to apply the results from assessing discrete regression models to the output of a continuous regression, as it would be to perform the standard assessments of a continuous regression whose variables were initially discrete. This appears to be a problem that cannot be easily circumvented and which hinges on the assumption of continuity for the main variables. The best way to resolve this would be to leave all variables in their original discrete/continuous states, however this returns us to the problem of the overwhelming consequent dimensionality of the problem. The resolution of this recurrent problem is unclear and is a possible area for future research.

8.4.4 Hips model

As we have seen in Figures 8.13 and 8.14, there are no relationships between the pre-operative patient status variables and those observed at later times. The only detectable associations in these models are attributable to the effect of waiting time on patient state. These associations manifest themselves in the form of a slightly poorer patient state. However, for prognostic purposes this separation of the patient states is not useful and is most likely attributable to the effect of the operation. However, if we consider the relationships between the post-operative patient states, as in Figure 8.15, we find evidence of some associations between the patient state at three and twelve months. If we consider *Usual Work* at 12 months ($\boldsymbol{U}$), we find it is associated with *Usual Work* ($\boldsymbol{u}$), *Walking Without Pain* ($\boldsymbol{w}$), and *Severe Pain* ($\boldsymbol{x}$) at 3 months. The relationship is of the following form:

$$U = 1.837 + 0.401u + 0.081w + 0.123x.$$



(a) Histogram of residuals

(b) Normal Quantile plot of residuals

Figure 8.19: Histogram and quantile plot from model of *Usual Work* at 12 months for the hips data.

We see, as before, that good levels of *Usual Work* at 12 months are associated with similarly high scores on the patient's three-month state. The estimated standard deviation for $\boldsymbol{U}$ is 0.762, which reflects the noisy nature of the hips variables. The histogram and quantile plot for the residuals are given in Figure 8.19 and show that the residuals could be considered to be approximately normal. In fact, all the variables in this model are ordinal as with the satisfaction model for the knees data in 8.4.3. However, whilst there is evidence of a slight step pattern to the quantile plot which reflects this granularity of the data we can see a pronounced difference in the nature of the residuals of the two models. This could suggest that the problems with the residuals observed with the satisfaction model for the knees data due to the ordinality of the variables could be attributable to its small sample size with only 78 cases. The problems seen with the knees data are not as evident with this hips model despite being purely ordinal again, though here we have at total of 1496 cases.

The adequacy of the model fit as assessed by the R^2 value is poor with $R^2 = 0.0086$. This is likely due to the noisy nature of the data and suggests that our model barely explains any of the variation of the data. Looking at the larger model including all the covariates we find the R^2 value barely changes and remains of the

	Est.	SE	Std. Est.	Efron CI	Hall CI
Intercept	1.837	0.1633	11.2489	[1.5339, 2.1570]	[1.5170, 2.1401]
Usual Work	0.401	0.0265	15.1178	[−0.0550, 0.0470]	[0.7550, 0.8570]
Walk w/o Pain	0.081	0.0141	5.7566	[−0.0430, 0.0120]	[0.1500, 0.2050]
Sev Pain	0.123	0.0182	6.7528	[−0.0430, 0.0280]	[0.2180, 0.2890]

Table 8.4: Parameter estimates for the hips model with bootstrapped standard errors and confidence intervals.

same order of magnitude. This highly noisy nature of the data is a feature of this hips data set and poses a significant problem when we are seeking to predict these data.

8.5 Limitations and Discussion

One of the main limitations of applying this chain graph modelling approach to the data is the absence of standard errors for the parameter estimates of the conditional Gaussian distribution (p, μ and Σ). Since the coefficients in the regression equations of the response variables given the covariates are functions of these parameters, the standard errors associated with these coefficients are also unknown. However, the associations that are modelled between these variables are known to be important and significant since the corresponding edges were included into the graphical model via the selection process. Therefore we could conclude that the significance of the coefficients is implicit due to this selection. However, when considering the regression equations for the responses there is no direct statement of the value of the standard errors and hence the associated t-values and significance probabilities cannot be calculated. This lack of standard errors leaves a worrying gap in the graphical modelling framework and presents a major limitation when compared to other statistical techniques.

Expressions for the standard errors for the parameters of pure continuous models are available in the paper by Roverato and Whittaker [108] and can be expressed in terms of the Isserlis matrix of the variance [64]. However, the results are specific to the covariance selection models and are not directly extensible to the mixed

data situation. Furthermore if the chain graph model under consideration were the saturated chain graph model, then the regression equations we obtain would correspond directly to the standard least-squares linear regression model and so the results from this methodology would be directly applicable. When the model is not saturated, the problem becomes far less tractable - even more so if we leave the framework of decomposable models.

An alternative approach to the model selection process was taken by Pigeot *et al* [98] in their study of the careers of sociologists. Instead of applying the model selection procedures that were discussed in Chapter 5 and were applied to the orthopædic data, Pigeot *et al* constructed their graphical model explicitly in terms of the regression equations of variables in later blocks given those other variables in the same block and all variables in prior blocks. Each response variable was considered in terms of its corresponding univariate regression and edges were present in the model if their corresponding terms were included in the regression equation through standard forward selection based on an increase in the likelihood of the regression equation. This strategy, whilst directly interpretable in the context of regression, suffers from the fact that it does not strictly equate to a graphical model with the associated Markovian properties. Hence interpretation of the presence or absence of arcs are complicated and typically only valid for special cases.

This absence of standard errors for parameter estimates is slightly problematic as it potentially leaves a gap in the final analysis of the regressions obtained via the chain graph models. However, the methods of bootstrapping [41, 43] can provide a mechanism for obtaining estimates of these quantities. It should be noted that bootstrapping is a computationally intensive approach and could be infeasible in a practical context. Nevertheless, the ability to calculate such values and the associated significance probabilities has made the modelling and prediction analyses more thorough. Bootstrapping does provide a remedy, though imperfect, for some of the limitations of this methodology, and its application here also appears to be novel in the context of graphical modelling. This area is a ripe area for future research and would complete the analysis of the graphical modelling methodology.

The problem of the exclusion of significant edges from the model due to the

constraints of decomposability are still present in the chain graph method, though to a lesser degree. As explained previously, it is desirable to remain within the framework of decomposable models as to venture outside those boundaries would result in models that were harder to interpret and fit. As before, backward selection would definitively address this issue however the large number of variables and the mixture of data types renders this a currently impossible avenue of investigation.

The construction of the models was performed by forward edge selection from the independence model. However, it has been discussed above that it may be reasonable to introduce some structure into the model before beginning the selection process in order to force certain edges into the model. The main case for this would be the introduction of edges between measurements on the same quantity that are separated by time in order to use the earlier observation as a baseline value for the latter. There is a compelling statistical argument for doing this with the types of repeated measures data investigated in this thesis, however application of the technique showed a negligible change in model performance. Therefore for the purposes of the models in this thesis this baseline method was not used, though it should not be discounted for the analysis of future data sets.

In terms of the adequacy of the fit of the graphical models to the data a number of points can be made. First, we can see from the R^2 values for the regression equations we have considered here that we captured between 20% and 40% of the variation of the response variables by the fitted model. When compared with a larger model including more of the pre-operative covariates, the performance could only be slightly improved suggesting that the data are intrinsically difficult to predict on the basis of the pre-operative data and that the original models performed quite well relative to these parent models. This low prognostic capacity in the models is not unreasonable, as the patient's pre-operative state was commonly seen to be somewhat different from their post-operative condition due to the normalising effect of the treatment. It may be the case that we would be better able to predict the patient's future state on the basis of their immediate post-operative condition since the post-operative data seem more closely associated. However, whilst we would obtain benefits in terms of the prognostic power of the model it would not be helpful

in deciding what treatment to suggest for a patient.

The effects of outliers on the selection of models can be considerable as mentioned earlier in this chapter. The work by Kuhnt *et al* [77] on the effects of contamination of the data by outliers on the model selection process showed that outliers can seriously affect the final choice of model. Therefore outlying observations in the data will have a notable effect on the final graphical model, so data should be screened for outliers before constructing these models. Additionally, this has implications for variables which are heavily skewed, such as *Extension Lag*, since they will display a large number of outliers and could have pronounced effects on the model.

Additionally, the effect of the violation of the normality assumptions of the data is evident in the distribution of the residuals from the regressions. This does raise some concerns since, ideally, we would expect Normal residuals and this is not the case for some response variables. The effect of this on the performance is not directly apparent though it may perhaps account for some of the lack of fit to the data. However, the goal of this project is to construct, combine and apply methods that will enable the analysis and interpretation of general orthopædic data sets. By applying an overarching framework in terms of the structural assumptions as in Chapter 2, or distributional assumptions such as the assumption of Normality we then enable ourselves to analyse general or arbitrary orthopædic data. It is almost certainly possible that we could construct better models by considering each data set individually and applying any of the many varied statistical techniques available today. However, the specific results and methods applied to one data set would not necessarily translate to other data as yet unseen. Therefore, in order to remain within a domain of generality, we must make such assumptions as these and accept the fact that performance will likely be slightly less than if we were to apply specific individual solutions to every data set encountered.

Chapter 9

Discussion, Problems and Limitations

This chapter begins in Section 9.1 with a discussion and evaluation of the methods presented in the previous chapters of this thesis. The applicability of the methods and techniques to the problem of clinical decision support is also discussed. In Section 9.2, the main unresolved problems encountered throughout this thesis are reviewed and discussed with some possible solutions being proposed. Finally the chapter ends in Section 9.3 with some suggestions for future development of the ideas and techniques presented in this thesis.

9.1 Discussion and Evaluation

The focus of this research has been on data resulting from total joint replacement surgery, but a further goal is that the results be sufficiently generic to be portable to other areas of orthopædics, or medicine. Working only within a general framework has substantially increased the complexity of the project as it required the development or application of methods and techniques not just to the two example data sets, but also to any arbitrary unseen data set of a similar nature. Further, as the research performs a technology translation between statistics and medicine an additional goal has been to ensure that the results and output of the statistical analysis are easily interpretable by non-experts. The problem has been further complicated

by the highly multivariate nature of the data which has significantly increased the size and complexity of the problem.

The research itself can be broken down into four main strands:

1. Data generalisation

2. Visualisation and graphical methods

3. Variable selection

4. Modelling and prediction

First, the purpose of the data generalisation process was to provide and abstract structural framework for the data in which they could be analysed. To this end the method and the abstraction itself have been most successful as they provide a general structural context within which the two example data sets could be investigated using the various statistical methods presented. This structural skeleton for the data is necessary in order to be able to propose and answer questions about prospective data sets. If there were no such assumed overarching structure then the data could be formless aggregations of disparate variables that would be impossible to reason about in generality. The framework identified the group of replicated patient status variables as a set of key observations, and this was further corroborated by orthopædic consultants - hence the focus of the subsequent analysis was to investigate the relationships within this group of variables and to determine the effects of the passage of time, the treatment and other factors. Suggesting such questions without the context of the meta-structure for the data would have been difficult, and subsequently proposing methods of analysis would have been impossible. Thus this stage of generalising the structure of the data set has proven to be an invaluable first step.

The second strand of research was focussed on the exploration of the data sets via graphical methods. The goal and purpose of these visualisations was to provide the clinician with a simple intuitive overview of the data, especially the differences between two or more subgroups of patients, e.g. those who received different treatments. They also summarised the evolution of the patient status variable over time

and provided a summary of the associations between the variables. In attempting to achieve these goals, the methods presented in Chapter 4 have been reasonably successful. The t-test plots served as a simple summary for the differences between two groups of variables, however they were limited by the fact that they can only be used to compare two subgroups of the data at a time and their application to ordinal data may be questionable.

The correlation plots (corrgrams) of Friendly [50] are an efficient way to depict the association structure between groups of variables, especially the patient status variables. They allow for an immediate assessment of possible correlations structure to the variables, as with the knees data, and easy identification of variables which are tightly correlated to others and those which appear unassociated. The interpretations of these behaviours among the variables would be of some interest to clinicians. These associations between groups of variables are of particular interest as they play a governing role in the variable selection process with the variable with the highest average squared correlation being identified as the most important. The correlations are also important in terms of the graphical models selected in the later stages of the analysis.

The profile plots display the evolution in the standardised mean of the patient status variables over time. They are a simple graphic that is both informative and intuitive and presents the viewer with an immediate overview of the data set. Orthopædic consultants have expressed notable interest in these methods. The ability to divide the profile plots into groups and present all variables simultaneously extends its usefulness significantly. The first limitation of these plots is that the plots can become easily swamped when displaying profiles from multiple subgroups. This problem could perhaps be addressed via interactive methods on a computer allowing the user to view and compare only certain profiles of interest. The second limitation of these plots is that points in the profile with small sample sizes and large associated uncertainty can appear as extreme values on the plot with substantial deviations from the other profiles on the graph. Attempts to display sample sizes by the colour intensity on the graph provided a partial solution, alternatively line thickness could be used to illustrate the same concept. A preferable method for displaying the

level of associated uncertainty would be the addition of error bars to each point, however this could be immensely confusing when we are displaying many profiles simultaneously.

In terms of the applicability to the clinical decision support problem, these visual methods are very useful tools. The profile plot and the t-test plots require only a small amount of statistical knowledge to read and interpret and both provide compact and intuitive summaries of the data and have been well received by clinicians. The correlation plots require more statistical knowledge to interpret and so may be of less interest in a clinical setting, but they remain an informative visualisation and are especially useful since notions of correlation and association are intrinsically linked to the variable selection and modelling methods.

The third strand of research addressed the variable selection problem where the goal was to identify a subgroup of the most important variables in the data set. The variables were chosen in such a way as to attempt to prevent multiple variables conveying the same information from being selected in the final set, thus the chosen variables all conveyed novel information in terms of the overall variability. The motivation for pursuing this avenue of research was to reduce the dimensionality of the data and thus, ultimately, reduce the complexity of the modelling problem.

The variable reduction methodology has been very successful at its goal and the final procedure is an efficient method with excellent performance for a stepwise procedure. The h statistics prove to be useful indicators for variable importance and the combination using partial variance to eliminate the effects of selected variables yields an effective technique. The extension of the process to include the repeated measures data sets has proven to be extremely valuable in the context of the modelling problem for the orthopædic data as it allowed the extraction of a subset that was, on average, the best at all time points in the data. This allowed for the simultaneous reduction of the multiple replications of the patient status variables thereby rendering the modelling problem significantly more tractable. The extension of the process to include variable utilities as weightings to guide the selection process could prove a powerful tool allowing the clinician to incorporate their expertise into the selection process to yield results that are potentially more practically useful.

The variable selection approach has some direct and potentially very useful applications to the clinical setting. For example, we have seen strong redundancies among the patient status variables in the hips data set and could use only a third of the patient status variables and yet retain 65% of the original variability. This could have consequences for the Oxford Hip Score [27] which is formed by summing these quantities and now can been seen to compound these redundant relationships. Furthermore there are notable implications for the future collection of data of such measurements which could result in smaller patient questionnaires and cheaper studies.

The final strand of research addresses the problem of constructing an appropriate model for the orthopædic data and obtain predictions from these models. The chain graph modelling approach provides an ideal framework for handling the repeated measures aspect of the data with its block-recursive formulation. Using this block structure we can investigate the associations among the patient status measures and associations between these variables at different time points. The framework also allows for the inclusion of both discrete and continuous data allowing for the inclusion and the assessment of potential effects of other categorical factors.

The methodologies applied to investigate these problems have been broadly successful, though that success has been slightly limited. The association of a graphical model with an independence graph gives a useful visual aid that is easily interpretable by clinicians. The associations and relationships between the variables can be directly read from the independence graph, which has been much appreciated by orthopædic consultants. The regression equations can be easily extracted from the chain graph and so the associations and relationships represented in the graph can be quantified via regression coefficients. The parsimonious linear models obtained from the chain graphs appear to be reasonable, however it does appear that the example data sets are complicated by the noisy responses. Furthermore, the analysis of these fitted regressions is confounded by the unavailability of standard errors for the regression coefficients preventing an assessment of their significance. Bootstrapping the parameters of the regression provides us with appropriate estimates of these values, though whether it is practical or reasonable to rely on bootstrapping

as a preferred method in these circumstances is questionable.

Another limitation of the modelling strategy is the assumption of Normality for the patient status variables in the model. The chain graph models require continuous variables to be multivariate Normal, however by assuming that the ordinal patient status variables were continuous we violate this assumption. This has implications for the fitting of the models to the data and their subsequent criticism and analysis.

A further limitation is the restriction of working with graphical and decomposable models. This places restrictions on the model selection process which can prevent the inclusion of potentially significant edges into the model as it would result in a violation of these properties.

9.2 Unresolved Problems and Limitations

9.2.1 Distributional Assumptions

As mentioned above, at several places in this thesis we have made assumptions on the distributions of the variables in the data. Typically this has taken the form of assuming that the many ordinal patient status variables can be approximated as continuous quantities. The motivation for this was that to retain ordinality of the many patient status variables would require working with contingency tables of prohibitively high dimension which would consequently be mostly sparse due to the sheer number of possible combinations (the 'curse of dimensionality'). The retention of the ordinal nature of the data would require working with and modelling these contingency tables which would heavily complicate the problem and be computationally intractable. Therefore, the ordinal variables were assumed to be continuous quantities as they could be considered as the discretisations of a latent continuous quantities.

These assumptions have many connections to the methods presented here. For example, the standardised profile plots assume continuity of the variables in order to plot the (conditional) means of the status variables. The variable selection methods require a correlation matrix in order to identify the best subset. If the data were treated as ordinal this reduction may not be directly possible - a matrix of ordinal as-

sociations could be used instead, but the interpretation of the results in terms of the variability of the measurements would be lost. If we considered the ordinal variables to be ordinal discretisations of Normal distributions, then the polychoric correlation is an estimate of the correlation between these latent continuous quantities[33] and could perhaps be used as an alternative here.

Finally, the modelling and prediction process is probably most affected by these assumptions of continuity. Since the distribution of an ordinal variable with only 5 levels is markedly different from Normal, then the assumptions of Normality are not upheld. The direct effects of this are not immediately apparent, though the effects of including heavily skewed continuous variables will have a detrimental effect due to the increased number of outlying observations affecting the model selection process [77]. These assumptions also affect the interpretation of the fit of the regression models themselves as we have constructed continuous regression models for approximated ordinal responses and covariates. This assessment of goodness of fit becomes problematic as it is clear that the standard methods from continuous linear regression are not appropriate, whereas similar methods for models of discrete data are simply not transferable or meaningless in the context of continuous regression. In such cases alternative analyses are required.

As discussed in Chapters 3 and 5, it could be possible to improve the Normality of the continuous variables by transformation. This could then have a consequent effect of improving the regression models, however this strategy would not be applicable to the ordinal data, even when approximated as continuous. However, the main limitation, as previously discussed, is that we lose the ability to reason directly about the quantities of interest and instead must operate in terms of functions of these variables. This sacrifices the interpretability of the final models and was seen to reap little reward.

Initially, we have applied a general structure to the data in terms of the data abstraction framework from Chapter 2, which gives the arbitrary unseen orthopædic data sets form and structure that forms the basis for analysis. However in order to model these data in generality it is also necessary to make some distributional assumptions to tackle the problem. Without such assumptions we would further suffer

from the dimensionality problems associated with the ordinal variables. Doubtlessly, considering each data set individually and creating specific tailored and, likely, entirely different solutions to each one would give us better models. However, in doing so we lose any notion of generality of the analysis as the methods would not be applicable or extensible beyond the scope of that particular data set. Hence, to continue working with an eye to the general case or a future unseen data set we must make assumptions about the nature and distribution of the data in order to be able to envisage methods to analyse the data. Consequently, the model performance may be lower than the specific tailored solutions and there will likely be cases where the validity of the underpinning assumptions are in question; however they are crucial when working in generality. However in the future, it could be possible to incorporate ordinal methods and analyses that could easily accommodate ordinal variables.

9.2.2 Dimensionality

The problem of the high dimensionality of the data was encountered extensively in Chapter 5 where it was seen that the graphical modelling methodology struggled with large numbers of variables. However, the problems attributable to high dimensionality are not confined to the modelling of the data. For example, when discussing the profile plots in Section 4.3 it was observed that the graphs become overcrowded when viewing profiles for multiple subgroups of the data. This problem of dimensionality is due to the large number of potentially informative discrete factor variables in the data and their consequently large number of possible interactions. One possible method of combating this would be to use dynamic computer visualisations for such investigations, which would be more flexible than considering static graphics and would allow for a sensible exploration of the data.

In terms of the problems of dimensionality associated with the modelling of the data, they were chiefly attributable to the fact that we had many patient status variables which were replicated at several time points giving rise to large and complex models. Many of the patient status variables were initially ordinal, but these were approximated as continuous quantities. This simplifies the problem considerably

as the analysis of continuous variables is significantly less sensitive to high dimensionality. Furthermore, the development and application of the variable selection methods to the patient status measurements at all time points provided a substantial reduction in the size of the problem which simplifies the model very effectively. The combination of these two methods has been successful at preventing dimension problems from affecting the modelling process. However, it is conceivable that were we to have a large number of time points in the data giving many replications of the patient status variables in the model then the original problem may recur. In such cases, we could proceed in one of two directions. First, if we were seeking to construct a joint model for the entire data set then the variable selection procedure would have to be more strict by returning a smaller variable subset, thereby reducing the number of variables at each time point. Secondly, if we were seeking to investigate only two time points such as the covariate/response structure for obtaining predictions then the models will be substantially smaller than the joint model anyway and so will not be as susceptible to problems of dimension.

9.2.3 Model Selection

One problem associated with the model selection process was that there were insufficient data to determine the parameters of the saturated graphical model. This was due to the high number of variables in the model and since for every possible combination of the discrete variables we must estimate $(q+1)(q+2)/2$ parameters, where q is the number of continuous variables, this can rapidly becomes problematic. Hence with large data sets we could have many more parameters to estimate than we have data available. This had the consequent effect of preventing model selection by backwards elimination forcing forward selection to be the method of determining the final model. If the saturated model was able to be determined, then it would be best to use a backward selection approach as we begin with a model that is consistent with the data and prune away any unimportant relationships in order to arrive at a final model. With forward selection, the initial model is that of independence, which will likely be inconsistent with the data and we seek to include edges into the model to improve the models representation of the data.

Another problem encountered during the selection of the graphical models was that certain significant edges were not included in the model since this would result in the model no longer being decomposable, or graphical or both. Whilst it is necessary to retain the graphical property of the model in order to interpret the model graph correctly, sacrificing decomposability to include such relationships is possible. The disadvantage of doing so is that the model could no longer be fitted exactly and an iterative method would be required such as via the MIPS [53] algorithm. Whilst not necessarily a barrier to using non-decomposable methods, with many variables in the model this could dramatically slow down the model selection process as for potential edge we would have to iteratively fit the prospective model in order to test it. This problem of important yet missing edges is averted when we use a backward selection strategy. However, this is not possible with these data as the models are too large to estimate the huge number of parameters of the saturated model, rendering it impossible to fit.

An alternative solution to this problem is to examine all edges which give a model that is graphical and either decomposable or not. If that edge is decomposable then we could add that edge to the model as normal. If not, then we could attempt to find the minimal parent model that contains the edges in the current model and the non-decomposable edge. If the edge is not decomposable because the resulting model is no longer triangular, then we could attempt to triangulate the graph by introducing additional 'fill-in' edges [73]. However, the problem of triangulating a given graph is an NP-complete problem and this is not a trivial process and would complicate the model selection process. These fill-in edges should be clearly marked as such as they will (typically) represent non-significant relationships that are present in the model for convenience of retaining decomposability. If the candidate edge were to introduce a forbidden path from a discrete node to another discrete node passing through continuous nodes, then we would be unable to obtain a decomposable model from this and so the edge would have to remain excluded from the selection process. A related problem associated with the model selection concerns the effects of having responses and covariates that included discrete variables. With this situation where discrete variables constitute some of the covariates and responses, it becomes

very easy to introduce a forbidden path into the model via connections with other continuous covariates or responses.

A further alternative is, of course, to consider performing backward selection from the saturated model, as the saturated model contains all edges significant or otherwise. Backward selection from here would retain all the significant edges in the model and so we would not suffer from the problems detailed above which apply exclusively to forward selection. However, there is a converse problem in that in order to remain decomposable certain *non-significant* edges will ineligible for *deletion*. These edges will approximately correspond to the fill-in edges introduced in the triangulation process mentioned above.

9.2.4 Regression Analysis

Having obtained a reasonable graphical model for our orthopædic data, we can investigate the relationships between covariates and responses by considering the corresponding conditional distributions. This gives us a linear regression-type formula for the response variables in terms of the covariates. However, as has been established in Chapter 8, the analysis for these regressions is limited by the absence of standard errors for the fitted regression coefficients. This absence is a consequence of a similar lack of expressions for the direct calculation of the standard errors for the fitted moments parameters (p, μ and Σ) of the joint CG distribution. Consequently, without statements about the errors associated with our regression parameters it is very difficult to assess the importance of the individual coefficients in the linear regression. Whilst we could conclude that since the graphical model has added an arc between each of the covariates and the response variable in this regression then there is an implicit significance attached to the coefficients in the conditional distribution. However, without standard errors we are unable to calculate the corresponding t-values and significance probabilities and obtain a numerical statement of the coefficient's importance. This leaves a substantial gap in the theoretical basis of the graphical modelling framework and is prominently absent when compared to many other standard statistical techniques.

The absence of theoretical results for the expressions of the standard errors posed

a problem to the analysis, however application of bootstrapping [41, 43] gave an alternative route for obtaining estimates of these values. As the bootstrap process is computationally intensive, bootstrapping all the parameters of the joint distribution would require a phenomenal amount of computation that would be unreasonable and impractical. Bootstrapping the smaller conditional distributions for standard errors for the regression parameters however was feasible and provided an indication of the importance of the regression coefficients that partially resolved one of the prime limitations of the methodology. Whether it is reasonable to routinely use bootstrapping to obtain such estimates is debatable as the computation may prove to be impractical. This area appears to be a novel application of the bootstrap paradigm and is a ripe area for future research which would complete the analysis of the graphical modelling methodology.

Finally, many of the variables were actually ordinal in nature rather than the continuous form that would typically be required for these analyses. Consequently, this raises some potential questions over the validity of the interpretation of the resulting regression, and also the analysis of the residuals, for example. Whilst the pragmatic application of continuous methods to the ordinal data has substantially simplified the complexity of the problem, it does pose these additional problems. Furthermore, results from the analysis of discrete or ordinal regressions are of little use to us here as it is impossible to use such techniques to interpret a continuous regression. The ideal solution would be to leave the variables on their original scales and treat them as ordinal or categorical as appropriate, however that is simply not possible given the constraints of the current statistical and computer technologies.

9.2.5 Low Predictability

An intrinsic feature of both data sets analysed in this thesis has been that the effect of the intervention has altered the state of the patient in such a way that the immediate post-operative state is relatively unrelated to their prior condition. When combined with the fact that the variables that are often considered as responses are noisy, this means that generating accurate predictions of the patient's post-operative state is difficult. Whilst not a problem with the methodology, if this feature were a common

and recurrent pattern among orthopædic data sets then we may be restricted to models with low prognostic power.

This issue of low predictability is attributable to the data and not the model selection. The reason for this conclusion is that in Section 8.4 larger parent models including all possible covariates were also fitted for comparison with the smaller models chosen via the model selection process. Comparison of the associated R^2 values for these pairs of models showed that there was typically only a small difference in the proportion of variation captured by the models. This suggested that the selected models were reasonable and parsimonious and the inclusion of further terms into those models would have only a negligible effect on model performance.

Transformation of the variables to improve Normality could improve the predictive power of these models. However, it would be unreasonable to apply transformations such as Box-Cox to the ordinal variables, even though they were approximated as continuous. As discussed above, this approach was dismissed as little improvements to the Normality of the data were attained via transformation and the reduction of the interpretability of the model was undesirable.

9.3 Possible Future Development

Despite these outstanding problems with and limitations of the work presented herein, it is conceivable that these methods could form the basis for a software tool providing clinical decision support. The tool would be required to fulfil two roles, the first of which being a data entry and storage system. By initially configuring the system, variables could be directly associated to components of the data abstraction. This association could simply take the form of specifying a temporal partial ordering to the data, or perhaps directly associating each variable to an element in the data generalisation. This would then allow the clinician to input the data recorded on patients into the system, which could then be stored in a manner that would facilitate future analysis.

With the key groups of variables being identified, the second role of the application becomes simpler. This role is the support of the clinical decision-making via the

methods and techniques discussed previously. The data could be easily explored by intelligent application of the visualisations, especially the profile plots which would give the clinician an overview of the data set as a whole. The profile for a single patient could then be overlayed onto the graph to inform the clinician where that patient lies with respect to the other patients in the data.

With the patient status variables identified, the variable selection process could be applied almost automatically requiring little user intervention. However, the clinician could specify utilities for the variables which would inform the selection process. The system would then be ready to model the data via the graphical modelling or chain graph methodologies. Obviously, the type of model generated depends on the specific questions being asked of the data. If the clinician were seeking to model the entire data set then a full chain graph model of all variables would be appropriate. If, however, they were seeking to predict one group of variables from another then they could identify the covariates and responses and then a simple 2-block predictive chain graph could be fitted. The problem of the standard errors for coefficients would still remain however, and bootstrapping the estimates is a time-consuming approach which could be infeasible within this setting. Nonetheless, these methods can provide the basis for such a data analysis package that would be of specific use in a clinical setting.

The methods proposed in this thesis do not explicitly consider the monitoring of patients. Instead the focus is directed towards understanding the relationships between the various patient status measures and exploiting any prognostic capabilities to obtain estimates of future patient state. Nonetheless, the monitoring of patient state is an area which could be addressed by some of the methods presented. For example if a patient were to present in a condition that strongly differed from that expected under the model then this may be a cause for further investigation. Similarly, methods such as the profile plots could perhaps be modified to present a method of process control.

A further avenue of research for the project would be to consider taking the graphical models returned from the fitting process and then using them as a basis for the construction of a Bayesian Belief Network (BBN) [21, 80]. Whilst this is a

complex problem, the graphical models we obtain from our data set could be used as a prior distribution for the BBN. This would allow us an entirely different avenue of prediction via the network that would not suffer from the problems of the chain graph method. This is a ripe area for future research.

Chapter 10

Conclusions

This chapter contains a summary and review of the conclusions drawn from the work presented in this thesis. Section 10.1 discusses some of the specific medical conclusions and implications made from the analysis of the two available data sets, whereas Section 10.2 presents a brief review of the general conclusions that were covered in detail in Chapter 9.

10.1 Medical Implications

Whilst the making of domain-specific conclusions of particular relevance to orthopædics is best left in the hands of the experts in the field, we can make some broad statements on the results we have obtained thus far and the methods that have been presented here.

10.1.1 Composite Scores

First, let us consider the issue of the composite scores. Composite scores are widely used in orthopædics to condense multiple variables which encode aspects of a patient's condition into a single quantity. This quantity is then used to assess treatment performance, compare patients and monitor patient condition. The fundamental limitation of this methodology is that it is a gross over-simplification with the large amounts of information contained in the individual status variables being lost when they are amalgamated into this single number. The individual variables convey de-

tails such as which areas of the patient's status may be changing, i.e. pain levels or walking ability - these individual features are lost when the variables are combined.

Seeking to retain this multivariate patient representation, it was determined that due to the high number of constituent status variables it would be necessary to perform some form of dimension reduction. To this end the techniques of Chapter 6 were developed. Investigation of the original correlation matrices showed (in the case of the knees data) that there was strong evidence of structural groupings within the data, suggesting that a degree of redundancy was present in the data. For example, strong association of the walking ability measures suggested that including all these variables into the composite measure was merely replicating the same information multiple times and would likely be counter-productive. With the hips data the variables exhibited a strong and almost homogeneous correlation which suggested that a dimension reduction would be particularly effective in this case as all the variables appeared to be closely related. Extensions to the selection process enabled the selection to be performed over all time points in the data.

Performing the variable selection procedure on the variables of the knees data (using the Nottingham scoring system [118]) enabled a reduction of 63% of the number of original variables (from 19 to 7) at a loss of only 31–43% of the information in the data (assuming that all the information was actually genuine and noise free.) Variable selection with the hips data was similarly effective, giving a reduction of 66% of the variables (from 12 to 4) with an associated loss of 36–45%. Whilst these losses were deemed to be acceptable here, it is trivial to repeat the analysis and extract larger subsets which cover a greater percentage of the variation of the data.

Despite being substantially different data sets, the variable selection strategy managed to reduce the knees and hips data sets to approximately one third of their original size whilst retaining approximately two-thirds of the information of the data. The ability to perform such a dramatic reduction has significant medical implications. First, it raises questions about the validity of operating with these composite scores; the ability to perform such reductions is indicative of redundancies of the component variables. Introducing groups of similar variables into such a score will mean that one particular aspect of the patient's status will be over-represented

in the composite score, and may even lead to the score being dominated by a single block of variables. This behaviour is undesirable.

Secondly, it identifies those variables which can be labelled the 'most informative' about a patient's status. These principal variables are most indicative of the patient's status and all other variables are typically associated with this pivotal group. Whilst these composite scores are limiting and disadvantageous it could be quite feasible to construct a better score using the principal variables extracted via these methods. An appropriate linear combination for expressing the patient's status as a single summary could be obtained by examining the first principal component over these variables. Given a suitable data set this would represent the variance-maximising linear combination of the principal variables and would give a measure that was most sensitive to the patient's condition.

Thirdly, by effectively reducing the number of informative variables in the data sets there will be consequent implications for other data collection endeavours. Since the principal variables represent the most informative set of variables in the data, in future studies information could be gathered on these variables alone. This could mean a significant reduction in the size of the questionnaires given to patients, making the process less burdensome. Additionally, there would be associated benefits in terms of cost and time. Furthermore, if we are using an 'Ease of Measurement' utility as in Section 7.2.4.3 then we could obtain the most informative subset of variables that cause minimal discomfort to the patient. Indeed, a utility could be constructed to choose the measurements that were such that the need for a physical examination was eliminated and the information could be given remotely via telephone or the Internet. Whilst not advocating a replacement for a face-to-face consultation, this could prove an efficient data collection mechanism that causes the least inconvenience to the patient.

10.1.2 Plots and Data Exploration

An exceptionally useful tool for the initial exploration of the orthopædic data is the profile plot. This plot has been well received by clinicians as it presents an immediate overview of the changes in a patient's condition over time. Dispensing with the

single composite measure, the profile plot allows for the visual representation of all the constituent variables of the composite score to be individually examined.

Examination of the profile plots showed a recurrent pattern across many of the variables, and indeed between the two data sets. Initially patients present with a poor condition as measured by the available variables. The subsequent treatment however appears to provoke a dramatic improvement in the mean of the patient states to a level far above their initial position. The hips data looked at a finer time-scale and it was seen that the even with the initial post-operative improvement being evident at 3-months, the mean state continued to improve up to the 1-year point albeit at a slower rate. This suggests that the treatment may have a persistent and continual effect over this period. The time points of the knees data were more separated with the first post-operative point being at 1-year. However, these data also illustrated the dramatic improvement as seen with the hips data, however due to the spacing of the observations the continued improvement was not visible. Instead we saw a slow decline over the subsequent 9 years, though with the final recorded state being, on average, still better than the patient's initial presentation.

10.1.3 Modelling and its Results

The graphical modelling methodology has been seen to be most effective in the orthopædic setting; the duality between the model itself and the associated independence graph is especially useful. The ability to display a graphical representation of the model substantially improves the interpretability of the relationships it represents. Displaying variables as nodes on a graph which are joined if there is a significant relationship between them is a highly useful and informative visual aid that depicts the model structure. These associations and relationships can be read directly from the model graph - a feature which has been much appreciated by orthopædic consultants.

In the course of the modelling of the temporal sequence of the patient status variables, one feature was observed that was common to both data sets. That feature was that there was a clear separation of the patients' pre-operative and post-operative states. In other words, the patient state after the operation was

(relatively) independent of the initial state. This has two implications; the first is that prediction of future patient condition on the basis of their initial state becomes difficult and suffers from low accuracy. The second is that if we infer that the treatment is responsible for this change in patient states, then this treatment is affecting patients in a manner that does not depend on the severity of their initial condition. Indeed, the treatment appears to act independently - possibly having a normalising effect across all patients.

In both data sets, the interventions investigated concerned the use of cement during the joint replacement procedure. In both cases of knee and hip replacement it was seen that there were no significant differences between these two types of intervention. This factor had no effect at any of the post-operative time points for both data sets suggesting that the use of cement during the operation has no discernible benefits or shortcomings in terms of the patient status.

The hips data set recorded the length of time that the patient was on a waiting list. Including this variable into the model displayed an association with variables at each of the time points. This suggests a possible relationship between the time spent on the waiting list and the patient's condition. Investigation of the coefficients of the modelled relationships showed that the waiting list was having a significant, yet small, negative effect on the patient's condition.

Again looking at the hips data set, a variable was recorded which indicated whether the patient was being treated via the NHS or privately. Including this information into the model had some interesting results. First, it was observed that this factor was associated with several of the pre-operative variables. These variables typically had a poorer state for patients in the NHS group when compared with the private group. This could be a potential consequence of the waiting times associated with the NHS treatment process resulting in NHS patients presenting in a state that was slightly more advanced than the private group. These differences were present only pre-operatively there was no significant effect on the post-operative data.

10.2 General Conclusions

The attention of this thesis has been focused on the analysis of clinical outcome data from total joint replacement. However, a key purpose of this work was to tackle such data sets in generality. The analysis of an unseen, unstructured and arbitrary data set is a prohibitively complex, if not impossible, task. This necessitated the construction of a structural framework at an early stage, which was imposed on the types of data sets to be considered. This framework gave structure to the possible analyses one may seek to perform on these data and enabled subsequent development of methodologies.

The remainder of the work in this thesis fell into three distinct categories. The first of which addressed exploratory visualisations of the data such as the t-test plots and standardised profile plots. These methods were seen to be both effective and intuitive in a clinical setting. The profile plots were especially well received as they give an immediate overview of the mean patient's evolution over time on many different measurements. The plots also enabled an initial comparison of multiple subgroups of patients, which can illustrate future possible research questions.

The second strand of the thesis dealt with the issue of variable selection. This area played a pivotal role due to the high dimensionality of the data and the replication of measurements over time. The work in this area forms the main novel theoretical contribution of this thesis and has been demonstrated to be highly efficient and effective in comparison to other methods in the literature. The ability to easily extend the selection process also emphasises its flexibility with extensions for longitudinal data and utilities attached to variables playing a prominent role in the subsequent analyses.

The final component was that of the modelling of the data. The techniques of graphical modelling and chain graph models were used to model the data due to their intuitive interpretation in terms of statements of conditional independence and the repeated measures structure of the data could be readily accommodated in a chain graph. However the use of this methodology was not without its limitations such as restrictive distributional assumptions and the unavailability of closed-form expressions for standard errors.

Overall, the methods presented herein have met with success, enabling the almost routine analysis of a generic orthopædic data set which conforms to the abstract structure of Chapter 2. The combination of the various strands of research provides tools for exploratory analysis through dimension reduction to modelling and prediction. As a future development of this work, all of these methods could be combined with a simple data management system to form the basis of a software package providing statistical support for clinical decision-making in orthopædics.

Appendix A

Implementation

In this appendix, an overview of the computer implementations of various aspects of this thesis is presented. This is contained in two sections: the former addresses the computer package MIM [38] that was used for some of the graphical modelling and a re-implementation of that package using C# [44]; the latter discusses some of the R [101] functions and their packages that have been created to perform some the methods discussed in this thesis.

A.1 MIM

A.1.1 Original

The MIM package is a command-driven Windows application designed for performing graphical modelling. MIM was written by David Edwards and is thoroughly documented in [38]. The software is freely available under the GNU Public Licence from `www.hypergraph.dk`. The package MIM has been used extensively to perform the analyses presented in Chapters 5 and 8.

MIM is operated via issuing commands to a terminal window such as that shown in Figure A.1. MIM allows for the specification of variables and models of mixed data types and performs edge deletion tests, model fitting, model selection procedures, chain graph modelling and a variety of other ancillary functions. It includes a mechanism for graphically depicting the independence graph associated with a

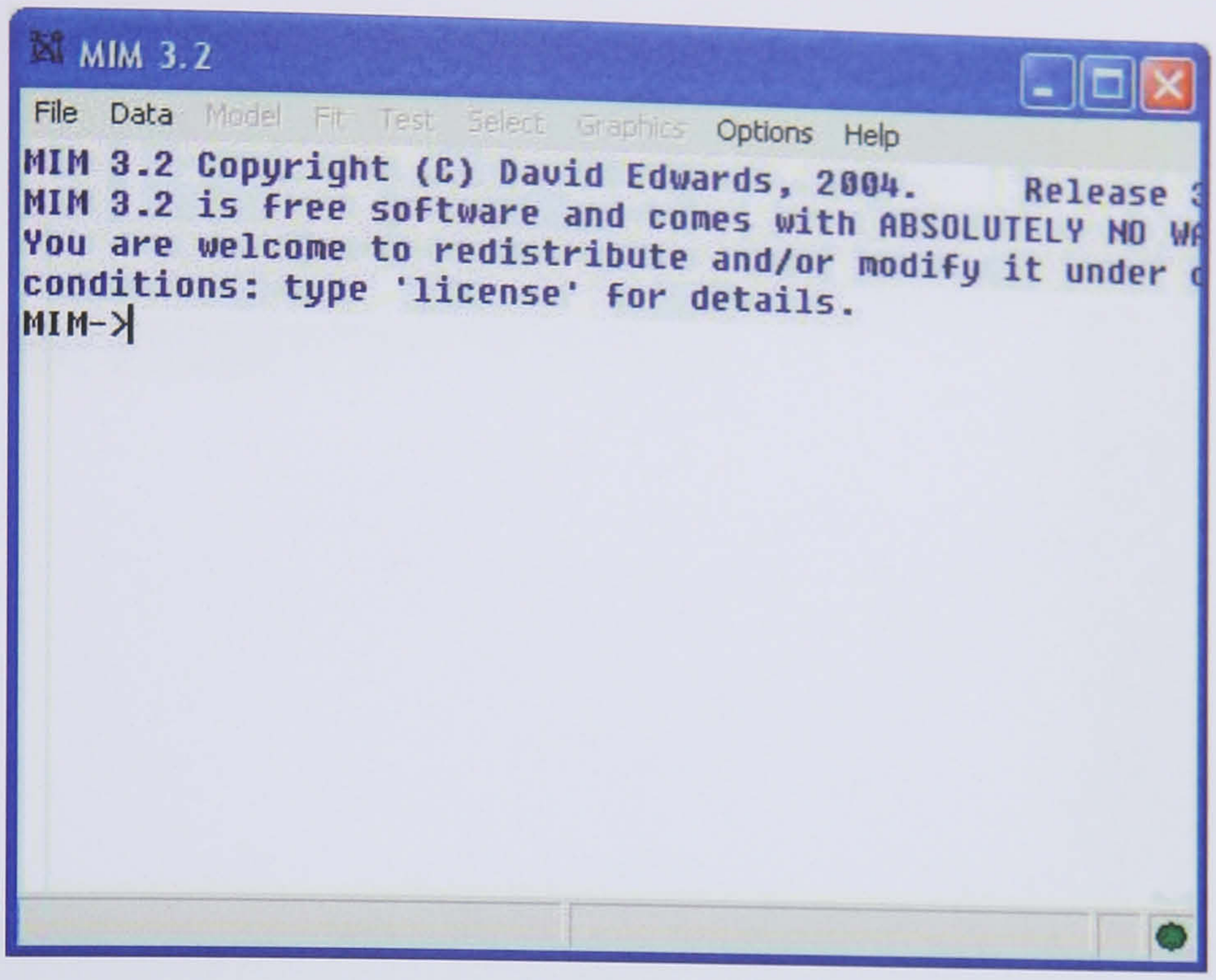


Figure A.1: The MIM interface.

particular model and allows for an interactive exploration of that model.

A.1.2 MIM#

The source code for MIM version 3.2 is freely available under the GNU Public License, therefore some time was spent investigating the possibilities of extending or customising the MIM graphical modelling framework. The original application was written using Borland Delphi, an object-oriented programming language similar in style to C++. To consider modifying the MIM application, the relatively new programming language C# was chosen instead of the original Delphi code. The reasons for this are that C# is widely portable in a similar way to Java, is also object-oriented and is a new language with growing support and popularity.

As a part of the research for this thesis, the core of the original MIM application was re-implemented in C# in order to provide scope for customisations of the application and to exploit some of the properties of this new programming language. This re-implementation will be termed MIM# in order to distinguish it from the original application. Whilst only a fraction of the functionality of the original appli-

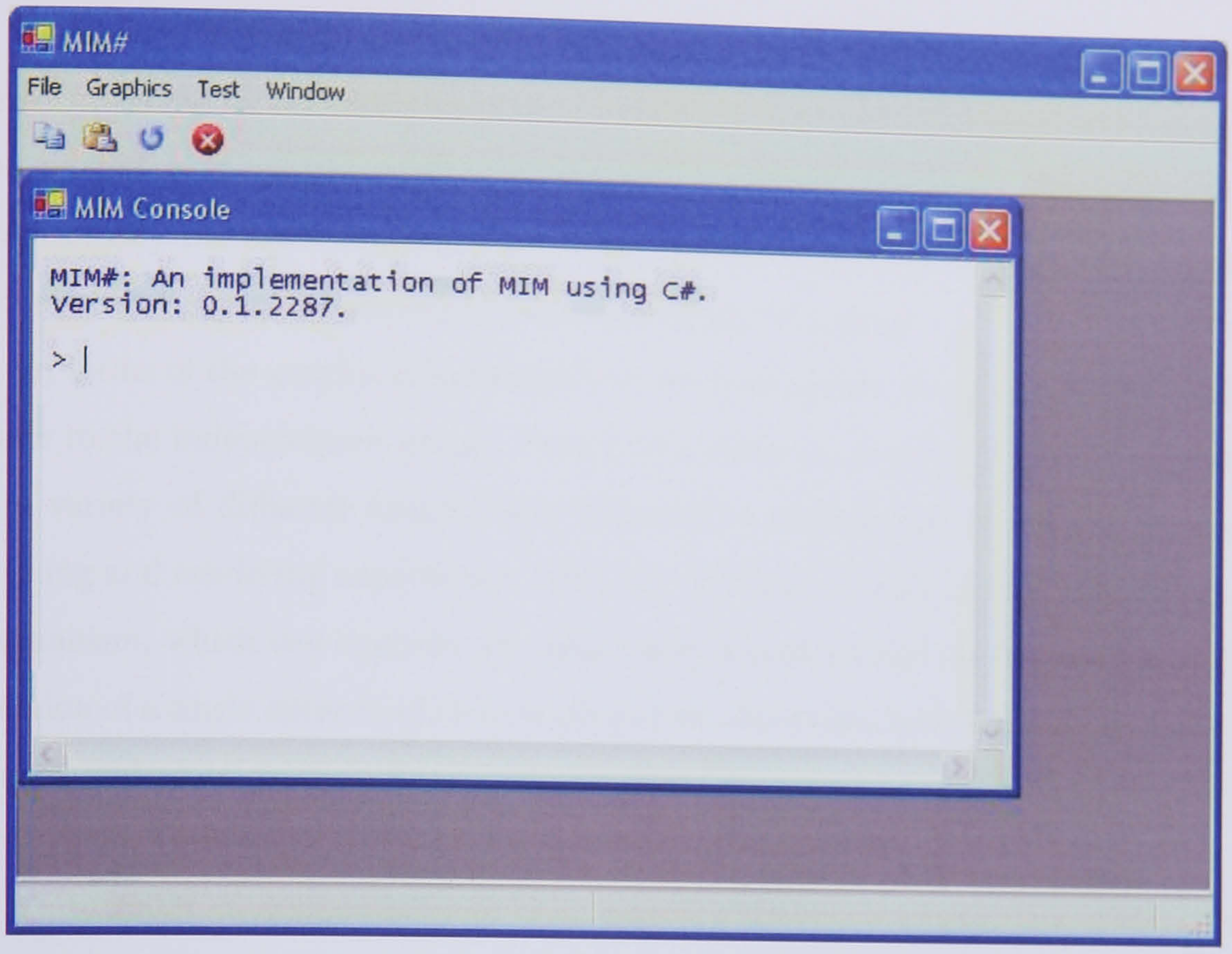


Figure A.2: The MIM# interface.

cation was implemented in this alternative version, it was sufficient to perform basic modelling and model selection with both undirected and chain graphical models.

A.1.3 Enhancements

This new implementation of MIM uses the same command syntax as the original MIM and features a graphical interface that is closely based on the original, though with some minor cosmetic and functional alterations. This should allow users familiar with the original system to use MIM# with ease.

The user interface for the new package MIM# is shown in Figure A.2 and differs slightly in presentation from that of the original in Figure A.1. The MIM# application uses a Multiple Document Interface (MDI) which allows the various child windows of the application to be contained inside a single parent window. For example the command terminal and one or more graph windows are all contained within the MIM# window. This prevents the multiple component windows from appearing independently on the desktop.

Several modifications addressed some of the difficulties with the original terminal-based interface. Specifically, support was included for copy and paste of commands and output in a manner similar to that of the terminal window of R [101]. This dramatically increases the usability of the terminal interface and allows for easier submission of terminal commands and extraction of output.

In terms of the graphical capabilities of the application, some modifications were made to the independence graph. Firstly, the ability to export all graphical output to a variety of different formats was included by radically re-designing the graph drawing and rendering subsystems. This has enabled a totally flexible visual output mechanism, which can support any additional output format by requiring only the creation of a single extra module to support basic drawing capabilities. Consequently graphs and plots can now be directly produced on-screen or exported to Postscript (PS, EPS), Bitmap (BMP) and Windows MetaFile (WMF).

Only small modifications have been made to the core functionality of the application since the efforts were to replicate the original application's main capabilities. The most significant adjustment in this area would be the inclusion of support for models containing more variables. In the original MIM, variables are specified via a single letter in the ranges A to Z and a to z thereby restricting models to 52 variables with simple names. In MIM#, support was included for multi-character variable names thus hugely expanding the scope for variable names and the number of variables. This has implications for the specification of model formulae for models with multiple character variable names. In these cases using the clique specification of MIM, i.e. *abc* to denote the clique over the three variables a, b, c. However, with multi-character variable names it could be possible to have a single variable named *abc* which would be confusing. Therefore, where longer variable names are present they are separated by ':' giving a clique specification of the form $var1 : var2 : var3$.

A small number of novel commands were also implemented in MIM# to support functions that were previously not provided. The first command is `predictions` which works in a similar way to the original `residuals` command, but instead calculates the predicted values of a variable under the current graphical model rather than the model residuals.

```
> statdisplay fg,ab
Empirical parameters:
Parameters of conditional distribution of fg given ab.
Conditional means and covariances.
 b
1        f    6.347   -0.051    2.104
         g    1.593    0.007    0.083    0.445
                   a        f        g
2        f    2.062    1.915    1.619
         g    2.020   -0.001   -0.046    0.247
                   a        f        g
```

Figure A.3: Example of output from the `statdisplay` command.

A second command is `statdisplay`, which also mirrors an existing command - in this case `display`. The purpose of `display` is to display sets of fitted statistics (typically counts, means and correlations) for a group of variables, often conditional on a second group. The purpose of `statdisplay` is to produce similar tables of statistics, but in this case the values are the (conditional) empirical counts, means and correlations rather than fitted values. This command allows for the direct calculation of these values for the data, and also enables the comparison of the empirical values (as input) to the fitted values (as output). Sample output from `statdisplay` is shown in Figure A.3.

The command `anovaform` also addresses the area of displaying parameter estimates. As mentioned above, this is typically achieved via the `display` command which produces tables of counts, means and correlations for each possible group of the discrete factors. When there are many factors with several levels this can produce reams of tables for scrutiny. The command `anovaform` represents these values as interaction terms as would be associated with an ANOVA analysis. Thus `anovaform` will calculate the value of the fitted main effects associated with each

```
> anovaform m

Calculating ANOVA decomposition...

Overall mean effects and slope:

  m  2.183   -0.024  0.266  0.006  0.028  0.010  0.042      536
     Means        a      f      i      j      k      p  Count

Main effects for discrete variables:

c0

  m  0.252    0.000  0.000  0.000  0.000  0.000  0.000      238
     Means        a      f      i      j      k      p  Count

c1

  m -0.202    0.000  0.000  0.000  0.000  0.000  0.000      298
     Means        a      f      i      j      k      p  Count
```

Figure A.4: Example of output from the `anovaform` command.

discrete variable, in addition to the relevant interaction effects. This can be used as a simple method for scrutinising the differences between the levels of a different factor and provides a more efficient representation than that of the output of `display`. Example output is given in Figure A.4

The final command is particularly applicable to the predictive chain graphs where we have a 2-block structure of covariates and responses. In such cases, we may be interested in the marginal relationships that exist between a particular response and variable the covariates in the model. To do this, the `explore` command was created to present a graphical interface for such an investigation. Calling this command opens an `explore` window which allows for such an interactive examination (see Figure A.5). The covariates are presented in the left column of variables and the responses in the right. Arrows between the two columns indicate the relationships in the model as usual. Firstly the user selects a response of interest by clicking it with the mouse. This populates the covariate list with appropriate variables. The user can then select a variable and a plot of the modelled marginal relationship

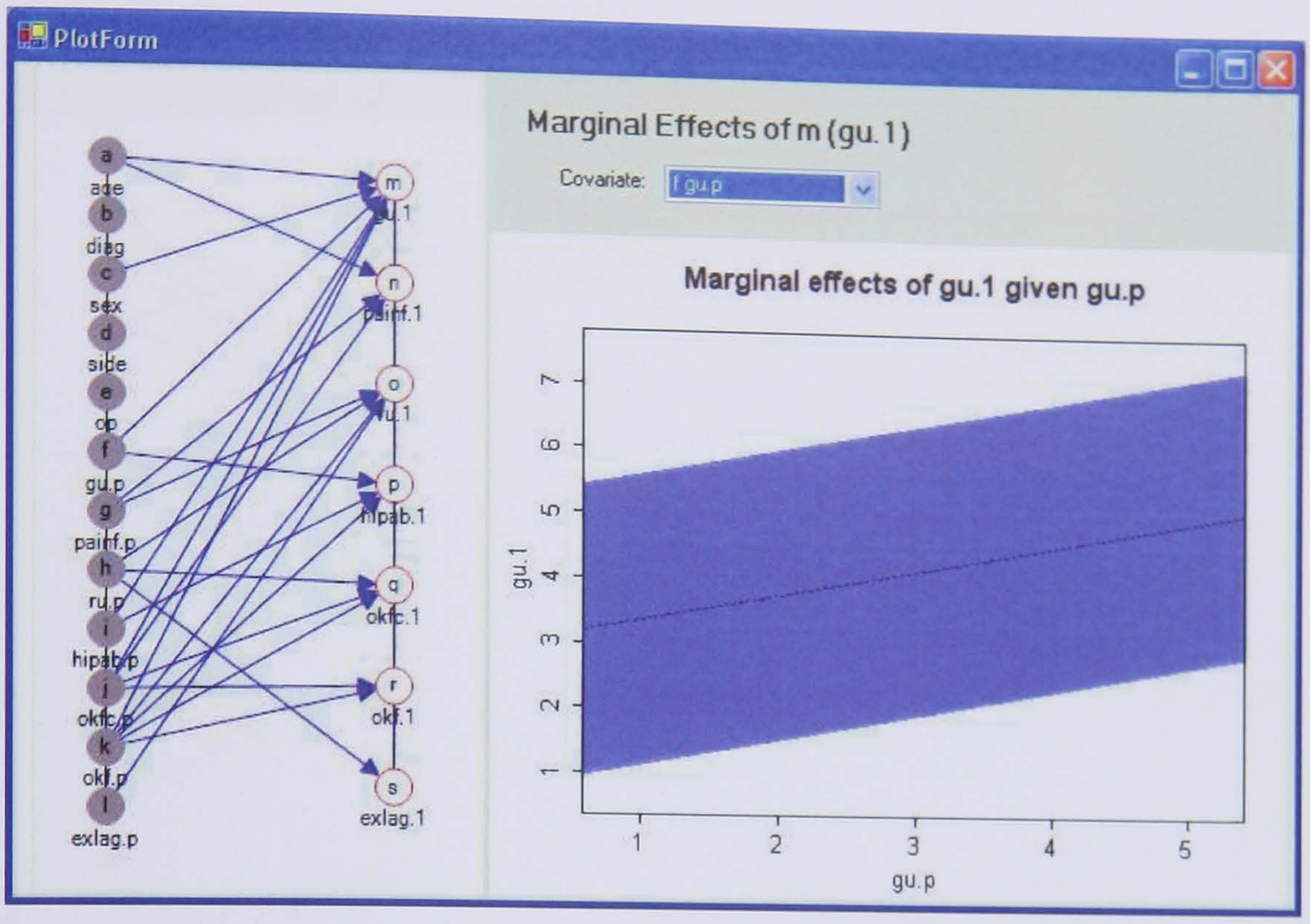


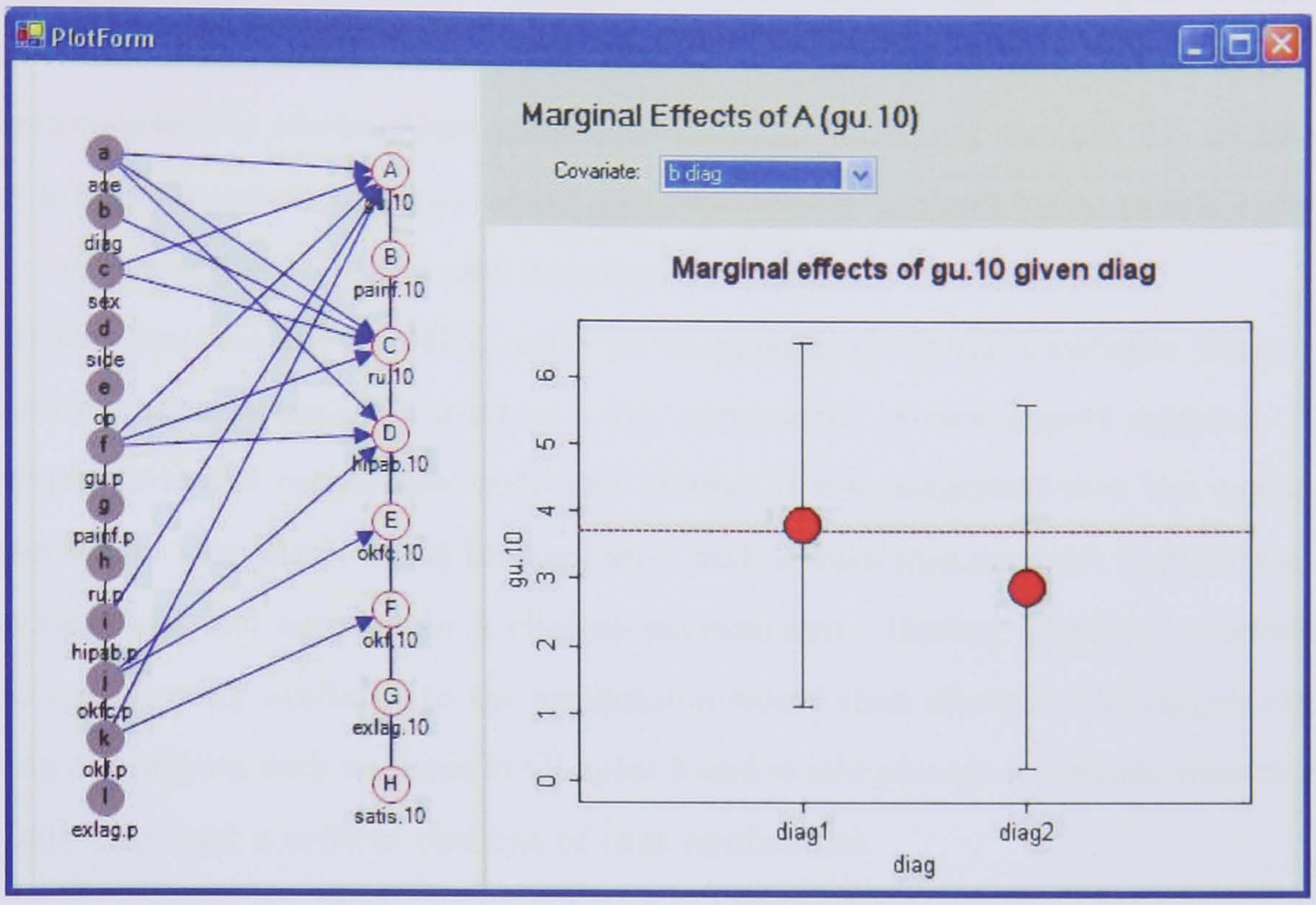
Figure A.5: The `explore` interface for continuous covariate and response

is given in the window beneath. In the continuous-continuous case a plot of the fitted regression equation with an envelope of $\pm 2\hat{\sigma}$ is drawn (as in Figure A.5). For discrete-continuous situations, error bar plots are drawn; each level of the discrete covariate is represented by the corresponding plot of $\mu \pm 2\hat{\sigma}$ (see Figure A.6). Discrete-discrete relationships are indicated by the display of the corresponding contingency table, with values in the table coloured in a manner similar to Friendly's mosaic plots [48] to indicate deviations from independence (see Figure A.6). Finally, continuous-discrete plots show the fitted logistic regression function.

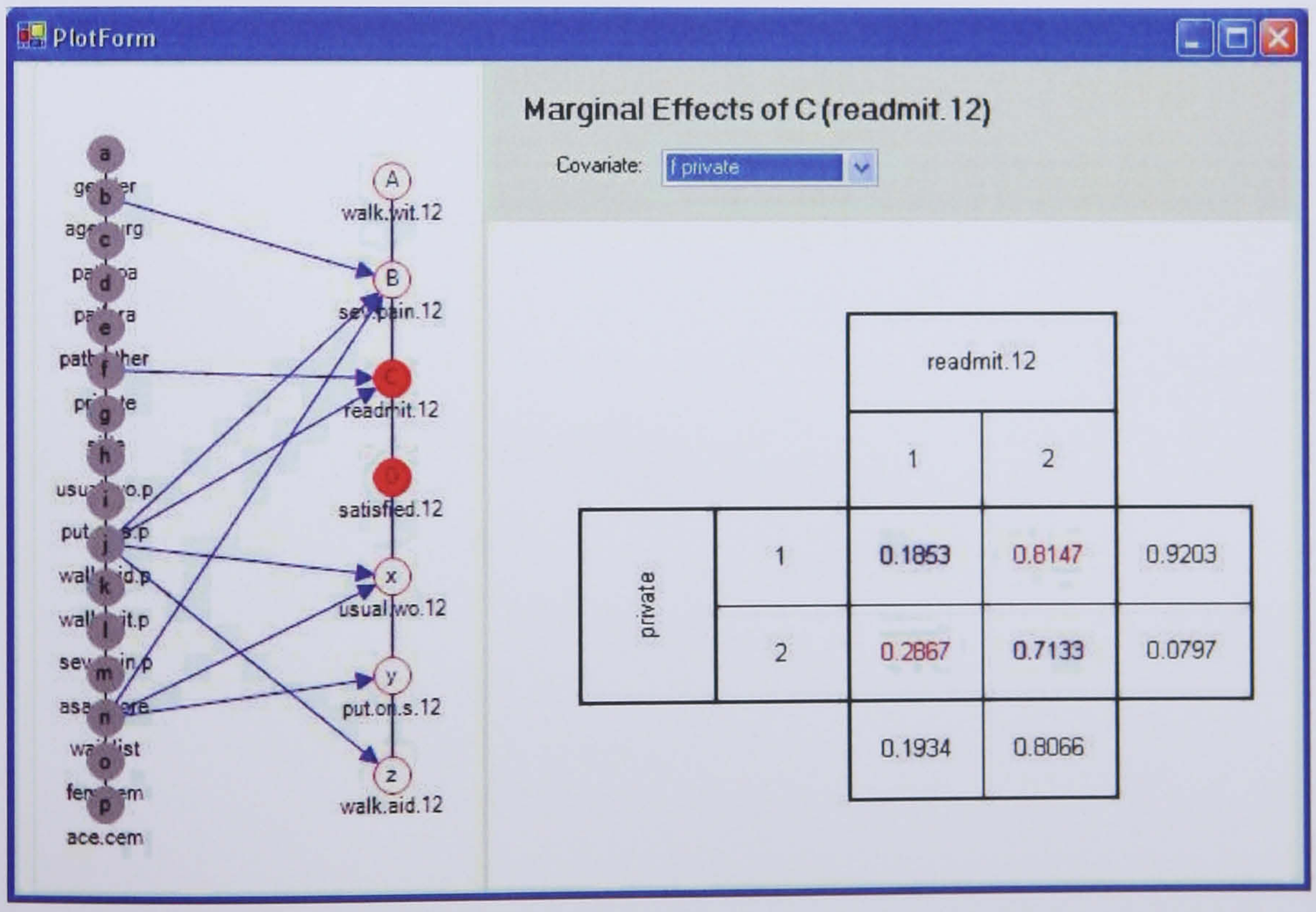
A.1.4 Potential Future Development

First, as mentioned above, only a skeleton group of MIM functions were implemented in MIM#. Whilst sufficient for basic modelling needs, it in no way provides the level of functionality supported by the original MIM package. Therefore a logical development of the MIM# application is to provide these as yet un-implemented capabilities.

An obvious development of this application would be to combine all of the func-



(a) Discrete covariate, continuous response



(b) Discrete covariate and response

Figure A.6: The `explore` interface for discrete-continuous and discrete-discrete covariate and response.

tionality provided by MIM into a single freeware library. Aggregating all of MIM's capabilities into a single library file, such as a DLL, would allow future users to programmatically access these graphical modelling functions directly. By exposing core functions and capabilities would allow for MIM's methods to be usable outside the confines of the standard user interface.

If the functionality of MIM could be encapsulated within a suitable library of functions, then it could be used as a key component of new expert systems. For example, with the orthopædic data sets studied it was suggested that the methods presented in this thesis could be combined with a data management application to provide statistical support in a clinical environment. Having graphical modelling functions directly available to the application would then allow for the simple replication of analyses such as those in Chapter 8 and would provide a suitable framework for inference and a critical element of that application.

A.2 R code

A.2.1 Variable Selection

The R functions in this package (provisionally titled `varsel`) perform the various variable selection functions described and investigated in Chapter 6. The functions are documented following the format of the standard R function documentation.

A.2.1.1 Basic selection routines

Description:

The variable selection procedures of Beale (**A1**, **A2**), de Falguerolles (**DF**), Jolliffe (**B1**, **B2**, **B4**), Krzanowski (**KP**)and McCabe (**M1**, **M2**, **M3**) are implemented in the following functions:

Usage:

```
varsel.beale(data, n, method, retIdx, diagnostic)
varsel.defalg(R, n, retIdx, diagnostic)
varsel.jolliffe(R, n, method, retIdx, diagnostic)
varsel.krzproc(data, n, retIdx diagnostic)
varsel.mccabe(R, n, method, retIdx, diagnostic)
```

Arguments:

`data`	The data matrix.
`R`	A correlation matrix.
`n`	The number of variables to select.
`method`	Some functions provide multiple selection methods which are specified via this argument. See below for details. Defaults to `"All"`
`retIdx`	Whether to return selected variable names or column indices. Defaults to `FALSE`.
`diagnostic`	Whether to show diagnostic output. Defaults to `FALSE`.

Details:

The input to the variable selection procedures differs between the selection methods. Some functions require the entire data set in order to operate (`varsel.beale` and `varsel.krzproc`), whereas others operate using the correlation matrix of the data. The required input is specified via the argument `data` for the data matrix, or `R` for the matrix of correlations. This argument and `n`, the number of variables to select, are the only required arguments of the selection procedure.

The functions `varsel.beale`, `varsel.jolliffe` and `varsel.mccabe` provide support for multiple variable selection methods. The method to be performed is specified via the `method` argument in the form of a character string containing the method's abbreviation as given in Table 6.3, e.g. `"M1"` or `"B2"`. More than one method can be performed by setting `method` to be a vector of such strings or assigning it the value `"All"` to perform all available selection methods.

Where the methods require performing an exhaustive search of all possible variable subsets of size `n`, computation can be long and involved. This is particularly true when the total number of variables, p is large and `n`$\simeq p/2$. If the number of subsets requiring evaluation is greater than 1000, the user is prompted to confirm that wish to proceed.

Return value:

An $(b \times n)$ matrix of the selected variable names (or column indices if `retIdx` is `TRUE`), where b is the number of selection methods performed and n is the size of the variable subset specified via the argument of the same name.

A.2.1.2 h-based selection function

Description:

The variable selection method based on the h values of the variables is implemented in `varsel.hmethod`. This method also allows for the specification of numeric utilities for each variable to guide the selection process.

Usage:

```
varsel.hmethod(R, n, utils, retIdx, diagnostic, scree.plot, cor.plot)
```

Arguments:

`R`	A correlation matrix
`n`	Number of variables to select.
`utils`	An optional vector of variable utilities. Defaults to `NULL`.
`retIdx`	Whether to return selected variable names or column indices. Defaults to `FALSE`.
`diagnostic`	Whether to show diagnostic output. Defaults to `FALSE`.
`scree.plot`	Whether to draw scree plots when finished. Defaults to `TRUE`
`cor.plot`	Whether to draw correlation plots of $\tilde{\boldsymbol{S}}_{22\cdot 1}$ at each stage. Defaults to `FALSE`.

Details:

At each stage the h values for remaining variables are calculated. The variable that is selected is the one with the highest value of h (or uh). The correlation matrix ($\boldsymbol{R}$) is then replaced by the unscaled partial correlation matrix given the variables selected so far ($\tilde{\boldsymbol{S}}_{22\cdot 1}$), and the process iterates. The method favours variables that are uncorrelated with one another, but are correlated to the unselected variables.

Numeric utilities for the merit of selecting a particular variable can be specified via the `utils` argument to guide the selection process.

Return value:

If `retIdx=TRUE`, then a vector of the indices of the selected variables. Otherwise a list with three elements:

`scores`	The h scores for every variable at each stage of the selection process.
`vals`	Values of h (and possibly uh) for each variable at the point at which it is selected. Also contains the squared norm and trace of the remaining unscaled partial correlation matrix at each step.
`vars`	The names of the selected variables.

A.2.1.3 General selection function

Description:

All of the above variable selection procedures can be accessed via a single method, which calls the relevant selection procedures from those given above and also calculates appropriate information about the performance of the returned subsets.

Usage:

```
varsel(data, n, R, method, calc.Var, calc.Time, diagnostic)
```

Arguments:

`data`	A data frame over which to perform the selection.
`n`	Number of variables to select.
`R`	If `data` is omitted, selection is based on this correlation matrix.
`method`	One or more method codes. See Table 6.3. Defaults to `"All"`.
`calc.Var`	Whether to calculate the trace and squared norm of the remaining partial correlation matrix. Defaults to `TRUE`.
`calc.Time`	Whether to record the time taken to perform each selection (for comparison of relative speeds of the methods). Defaults to `TRUE`.
`diagnostic`	Whether to print diagnostic output. Defaults to `FALSE`.

Details:

`varsel` serves as a wrapper for the other specific variable selection functions. Supplying `varsel` with `data` or `R` depends on the selection method you wish to use. All methods can operate using `data`, but Beale's and Krzanowski's methods cannot perform variable selection using `R`.

Return value:

A data frame whose first `n` columns contain the selected variables. The rows correspond to the selection methods. If `calc.Var` is `TRUE` then columns $(n+1)$ to $(n+4)$ correspond to the values of $||\tilde{\boldsymbol{S}}_{22\cdot 1}||^2$, $||\tilde{\boldsymbol{R}}_{11}||^2$, the percentage trace and percentage squared norm. If `calc.Time = TRUE`, then these values are appended in

the final column.

A.2.2 Graphics

A.2.2.1 Correlation Plot

Description:

The `corplot` produces a correlation mosaic plot for a given correlation matrix. See Section 4.2 for details.

Usage:

```
corplot(R, useAbs, main, sort, colDiag)
```

Arguments:

`R`	The correlation matrix to plot.
`useAbs`	Whether cells in the mosaic show $\|r_{ij}\|$ (`TRUE`, default) or r_{ij} (`FALSE`).
`main`	Main label for the plot. Defaults to `""`.
`sort`	Whether to sort the variable based on their angles on the biplot of the data. Defaults to `FALSE`.
`colDiag`	Whether show the values of diagonal elements of `R`, for use when `R` is in unscaled correlation form. Defaults to `FALSE`.

Details:

The correlation plot displays a matrix of rectangles corresponding directly to `R`. The rectangles on the diagonal contain the variable labels. Off-diagonal rectangles are coloured with intensity equal to the corresponding $|r_ij|$, where a zero correlation results in an empty (white) cell, and a correlation of ± 1 gives a solid red (or blue if `useAbs` is `FALSE`) cell.

A.2.2.2 Profile Plot

Description:

The function `profile.plot` produces a profile plot over one or more variables.

Usage:

```
profile.plot(data, time, index, group1, group2, main, colSample)
```

Arguments:

`data`	Data frame of variables to be plotted as profiles. See below for details on the form of `data`.
`time`	Vector of factors indicating to which time point the cases of `data` belongs.
`index`	Vector of unique indices indicating to which observation each case belongs.
`split1`	A factor variable representing sub-groups of `data`.
`split2`	A second factor variable representing sub-groups of `data`.
`main`	The main label of the plot. Defaults to `""`.
`colSample`	Whether to indicate sample sizes at each point via colour intensity. Defaults to `FALSE`.

Details:

The idea behind this plot is that we have a number of observational units each uniquely identified via a distinct ID number (or similar) in `index`. Several measurements are repeatedly made on these units at several points in time, as indicated by `time`. The measurements are recorded in `data` where each column represents a different variable. Thus each observational unit can be represented in several rows in `data`, where each corresponding row in `index` gives the same identifier value. The various entries for this unit will have different values for `time`, representing a set of measurements taken on the same unit at different times.

A.2.2.3 *t*-Test Plots

Description:

Plots the *t* statistics obtained for comparing the means of one or more variables between two groups in the independent or paired-sample cases.

Usage:

```
ttest.plot(data, split, signif, fixed, main, probs)
paired.ttest.plot(data1, data2, signif, fixed, main, probs)
```

Arguments:

`data`	A data frame containing the variables whose sub-groups are to be compared.
`split`	A factor variable defining two sub-groups of `data` for the independent sample case.
`data1`	Data frame containing the first observation of the paired sample.
`data2`	Data frame containing the second observation of the paired sample.
`signif`	The significance level for the tests. Significant results are coloured red. Defaults to 0.05.
`fixed`	Whether to fix the vertical limits of the plot to ± 6 and crop any values beyond. Defaults to `FALSE`.
`main`	Main label for the plot. Defaults to `""`.
`probs`	Whether to plot significant probabilities, rather than *t* statistics. Defaults to `FALSE`.

Details:

`ttest.plot` simply performs an independent sample *t*-test for each of the variables in `data`, comparing the means of the groups given by `split`.

Conversely, `paired.ttest.plot` performs a paired *t*-test for each of the variables in the `data` arguments, where the pairs are defined by the matching rows of `data1` and `data2`. Consequently `data1` and `data2` are expected to be of identical dimensions.

Value:

The function draws a t-test plot as per Section 4.1. The function also returns a matrix of size $(p \times 6)$ where p is the number of variables in `data`. For `ttest.plot` the first three columns contain the difference in the means of the two groups, the pooled variance, and the standard error. The first three columns for `paired.ttest.plot` contain mean difference, the variance of the differences, and the standard error. In both cases the final three columns contain t, $|t|$ and the corresponding p-value.

References

[1] A Agresti. *Categorical Data Analysis.* Wiley, New York, 1990.

[2] H Aikake. A new look at the statistical model identification. *IEEE Transactions in Automatic Control*, 19:716–23, 1974.

[3] E Anderson. The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

[4] H P Andrews, R D Snee, and M H Sarner. Graphical display of means. *The American Statistician*, 34(4):195–199, 1980.

[5] ANSI/HL7, Michigan, USA. *HL7 Version 3 Reference Information Model*, 2003.

[6] J H Badsberg. Model search in contingency tables by CoCo. In Y Dodge and J Whittaker, editors, *Computational Statistics, CompStat 1992 Neuchâtel*, pages 251–256, Heidelberg, 1992. Physica-Verlag.

[7] E M Beale, M G Kendall, and D W Mann. The discarding of variables in multivariate analysis. *Biometrika*, 54:357–366, 1967.

[8] C S Berkey, N M Laird, I Valadian, and J Gardner. Modelling adolescent blood pressure patterns and their prediction of adult pressures. *Biometrics*, 47(3):1005–1018, 1991.

[9] A Blauth, I Pigeot, and F Bry. Interactive analysis of high-dimensional association structures with graphical models. *Metrika*, 51:53–65, 2000.

[10] W Bossert and J A Weymark. Utility in social choice. In S Barberà, P J Hammond, and C Seidl, editors, *Handbook of Utility Theory*, volume 2: Extensions, chapter 20, pages 1099–1177. Kluwer Academic Publishers, 2004.

[11] G E P Box and D R Cox. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26:211–246, 1964.

[12] G E P Box, W G Hunter, and J S Hunter. *Statistics For Experimenters: An Introduction To Design, Data Analysis And Model Building.* Wiley, New York, 1978.

[13] J Cadima, J O Cerdeira, and M Minhoto. Computational aspects of algorithms in variable selection in the context of principal components. *Computational Statistics & Data Analsysis*, 47:225–236, 2004.

[14] J Cadima and I T Jolliffe. Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22(2), 1995.

[15] J Cadima and I T Jolliffe. Variable selection and the interpretation of principal subspaces. *Journal of Agricultural, Biological and Environmental Statistics*, 6(1):62–79., 2001.

[16] R B Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1:245–276, 1966.

[17] CEN/TC251 WG I. *ENV12265: Electronic Healthcare Record Architecture*, 1995.

[18] J M Chambers, W S Cleveland, B Kleiner, and P A Tukey. *Graphical Methods For Data Analysis.* Wadsworth, 1983.

[19] W S Cleveland. *Visualizing Data.* Hobart Press, Summit, New Jersey, 1993.

[20] J Cohen. *Statistical Power Analysis for the Behavioral Sciences.* Academic Press, New York, revised edition, 1977.

[21] R G Cowell, A P Dawid, S L Lauritzen, and D J Spiegelhalter. *Probabilistic Networks and Expert Systems.* Springer-Verlag, New York, 1999.

[22] D R Cox and N Wermuth. Linear dependencies represented by chain graphs (with discussion). *Statistical Science*, 8:204–283, 1993.

[23] D R Cox and N Wermuth. *Multivariate Dependencies: Models, Analysis and Interpretation.* Chapman and Hall, London, 1996.

[24] M J Crowder and D J Hand. *Analysis Of Repeated Measures.* Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1990.

[25] J N Darroch, S L Lauritzen, and T P Speed. Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics*, 8:522–539, 1980.

[26] A P Dawid. Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B*, 41:1–31, 1979.

[27] J Dawson, R Fitzpatrick, A Carr, and D Murray. Questionnaire on the perceptions of patients about total hip replacement. *Journal of Bone & Joint Surgery - British Volume*, 78(2):185–190, 1996.

[28] A de Falguerolles and S Jmel. Un critère de choix de variables en analyse en composantes principales fondé sur des modèles graphiques gaussiens particuliers. *Canadian Journal of Statistics*, 21(3):239–256, 1993.

[29] A P Dempster. Covariance selection. *Biometrics*, 28:157–75, 1972.

[30] T DiCiccio and B Efron. Bootstrap confidence intervals. *Statistical Science*, 11:189–212, 1996.

[31] P J Diggle. An approach to the analysis of repeated measures data. *Biometrics*, 44:959–971, 1988.

[32] P J Diggle, K Liang, and S L Zeger. *Analysis Of Longitudinal Data.* Oxford University Press, Oxford, 1994.

[33] F Drasgow. Polychoric and polyserial correlations. In L Kotz and N L Johnson, editors, *Encyclopedia of statistical sciences*, volume 7, pages 69–74. Wiley, New York, 1988.

[34] P Drineas, R Kannan, and M W Mahoney. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. Technical Report YALEU/DCS/TR-1271, Yale University, 2004.

[35] M Drton and M D Perlman. Model selection for Gaussian concentration graphs. *Biometrika*, 91(3):591–602, 2004.

[36] H T Eastment and W J Krzanowski. Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, 24(1):73–77, 1982.

[37] D Edwards. Hierarchical interaction models. *Journal of the Royal Statistical Society, Series B*, 52(1):3–20, 1990.

[38] D Edwards. *Introduction to Graphical Modelling.* Springer, New York, 2nd edition, 2000.

[39] D Edwards and T Havránek. A fast model selection procedure for large families of models. *Journal of the American Statistical Association*, 82:205–213, 1987.

[40] E Eells. *Probabilistic Causality.* Cambridge University Press, Cambridge, 1991.

[41] B Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979.

[42] B Efron and R Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1:54–96, 1986.

[43] B Efron and R Tibshirani. *An introduction to the bootstrap.* Chapman and Hall, 1993.

[44] European Computer Manufacturers Association (ECMA), Geneva, Switzerland. *Standard ECMA-334: C# Language Specification*, 3rd edition, 2005.

[45] P M Fayers and D Machin. *Quality of Life: Assessment, Analysis and Interpretation.* Wiley, Chichester, 2000.

[46] K A Fisher. Application of "Student's" distribution. *Metron*, pages 90–104, 1925.

[47] R A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[48] M Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89:190–200, 1994.

[49] M Friendly. Conceptual and visual models for categorical data. *The American Statistician*, 49:153–160, 1995.

[50] M Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4):316–324, 2002.

[51] S Frontier. Étude de la decroissance des valeurs propers dans une analyze en composantes principales: comparison avec le modèle de baton brisé. *Journal of Experimental Marine Biology and Ecology*, 25:341–347, 1976.

[52] M Frydenberg. The chain graph Markov property. *Scandinavian Journal of Statistics*, 17:333–353, 1989.

[53] M Frydenberg and D Edwards. A modified iterative scaling algorithm for estimation in regular exponential families. *Computational Statistics & Data Analsysis*, 8:142–153, 1989.

[54] W Gibbs. *Elementary Principles of Statistical Mechanics.* Yale University Press, 1902.

[55] L A Goodman and W H Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):734–764, 1954.

[56] T G Gregoire and B L Driver. Analysis of ordinal data to detect population differences. *Psychological Bulletin*, 101(1):159–165, 1987.

[57] R J Harris. *Primer of Multivariate Statistics.* Academic Press, New York, 1975.

[58] J A Hartigan and B Kleiner. Mosaics for contingency tables. In W F Eddy, editor, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 268–273, New York, 1981. Springer-Verlag.

[59] J A Hartigan and B Kleiner. A mosaic of television ratings. *The American Statistician*, 38:32–35, 1984.

[60] J A Hartigan and M A Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[61] R R Hocking. The analysis and selection of variables in linear regression. *Biometrics*, 32:1–49, 1976.

[62] H Hofmann. Constructing and reading mosaicplots. *Computational Statistics & Data Analsysis*, 43(4):565–580, 2003.

[63] D Ingram. The good european health record. In M F Laires, M F Ladeira, and J P Christensen, editors, *Health in the New Communication Age*, pages 66–74. IOS Press, 1995.

[64] L Isserlis. On a formula for the product-moment correlation of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12:134–139, 1918.

[65] J N R Jeffers. Two case studies in the application of principal component analysis. *Applied Statistics*, 16:225–236, 1967.

[66] I T Jolliffe. Discarding variables in principal component analysis. I: Artificial data. *Applied Statistics*, 21(2):160–173, 1972.

[67] I T Jolliffe. Discarding variables in principal component analysis. II: Real data. *Applied Statistics*, 22(1):21–31, 1973.

[68] I T Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 2nd edition, 2002.

[69] H F Kaiser. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20:141–151, 1960.

[70] L Kaufman and P J Rousseeuw. *Finding Groups In Data: An Introduction To Cluster Analysis.* Wiley, New York, 1990.

[71] M G Kendall. *A Course In Multivariate Analysis.* Griffin, London, 1957.

[72] F M Khaw, L M G Kirk, R W Morris, and P J Gregg. A randomised, controlled trial of cemented versus cementless press-fit condylar total knees replacement. *Journal of Bone & Joint Surgery - British Volume*, 84:658–666, 2002.

[73] U Kjaerulff. Triangulation of graphs - Algorithms giving small total state space. Technical Report R 90–09, Aalborg University, Denmark, March 1990.

[74] S Kreiner. Computerized exploratory screening of large-dimensional contingency tables. *Compstat*, 7:43–48, 1986.

[75] W J Krzanowski. Selection of variables to preserve multivariate data structure, using principal components. *Applied Statistics*, 36(1):22–33, 1987.

[76] W J Krzanowski and F H C Marriott. *Multivariate Analysis I: Distributions, ordination and inference*, volume I of *Kendall's Library of Statistics.* Arnold Publishers, 1994.

[77] S Kuhnt and C Becker. Sensitivity of graphical modeling against contamination. In M Schader, W Gaul, and M Vichi, editors, *Between Data Science and Applied Data Analysis*, pages 279–287. Springer, New York, 2003.

[78] R Largo, J Caflisch, F Hug, K Muggli, A Sheehy, T Gasser, and L Molinari. Neuromotor development from 5 to 18 years, part I: Timed performance. *Developmental Medicine and Child Neurology*, 43:436–443, 2001.

[79] S Lauer. Interactive modelling of categorical data. In B Marx and H Friedl, editors, *Proceedings of the 13th International Workshop on Statistical Modeling, New Orleans*, pages 443–446, 1998.

[80] S L Lauritzen. Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87:1098–1108, 1992.

[81] S L Lauritzen. *Graphical Models.* Oxford University Press, Oxford, 1996.

[82] S L Lauritzen and T S Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society, Series B*, 64(3):321 361, 2002.

[83] S L Lauritzen and N Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17:31–57, 1989.

[84] R J A Little. Regression with missing X's: A review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.

[85] C J Lloyd. *Statistical Analysis of Categorical Data.* Wiley, New York, 1999.

[86] T A Louis. General methods for the analysis of repeated measures. *Statistics in Medicine*, 7:29–45, 1988.

[87] G P McCabe. Principal variables. *Technometrics*, 26(2):137–144, 1984.

[88] A W McCaskie, D J Deehan, T P Green, K R Lock, J R Thompson, W M Harper, and P J Gregg. Randomised, prospective study comparing cemented and cementless total knee replacement: Results of press-fit condylar total knee replacement at five years. *Journal of Bone & Joint Surgery - British Volume*, 80(6):971–975, 1998.

[89] R McGill, J W Tukey, and W A Larsen. Variations of box plots. *The American Statistician*, 32:12–16, 1978.

[90] W N Mohamend, I Diamond, and P W F Smith. The determinants of infant mortality in Malaysia: a graphical chain modelling approach. *Journal of the Royal Statistical Society, Series A*, 161:349–366, 1998.

[91] D J Murdoch and E D Chow. A graphical display of large correlation matrices. *The American Statistician*, 50(2):178–180, 1996.

[92] G Neil-Dwyer, D Lang, P Smith, and F Ianotti. Outcome after aneurysmal subarachnoid haemorrhage: The use of a graphical model in the assessment of risk factors. *Acta Neurochirurgica*, 140:1019–1027, 1998.

[93] M Okamoto. Optimality of principal components. In P R Krishnaiah, editor, *Multivariate Analysis II*, pages 673–685. Academic Press, 1969.

[94] P Gregg P and B C Reeves. *National Total Hip Replacement Outcome Study.* Royal College of Surgeons of England and British Orthopaedic Association, London, 2000.

[95] J Pearl. Causal diagrams for empirical research (with discussion). *Biometrika*, 82(4):669–710, 1995.

[96] J Pearl and A Paz. Graphoids: A graph based logic for reasoning about relevancy relations. In B D Boulay, D Hogg, and L Steel, editors, *Advances in Artificial Intelligence II*, pages 357–363. North-Holland, Amsterdam, 1987.

[97] P R Peres-Neto, D A Jackson, and K M Somers. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analsysis*, 49(4):974–997, 2005.

[98] I Pigeot, A Heinicke, A Caputo, and J Brüderl. The professional career of sociologists: a graphical chain model reflecting early influences and associations. *Allgemeines Statistiches Archiv*, 84(1):3–21, 2000.

[99] G Pison, A Struyf, and P J Rousseeuw. Displaying a clustering with CLUSPLOT. *Computational Statistics & Data Analsysis*, 30:381–392, 1999.

[100] T Prvan and A W Bowman. Nonparametric time dependent principal component analysis. *The Australian & New Zealand Industrial and Applied Mathematics Journal*, 44:C627–C643, 2003.

[101] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2005.

[102] J O Ramsay and B W Silverman. *Applied Functional Data Analysis: Methods and Case Studies.* Springer-Verlag, New York, 2002.

[103] J L Rasmussen. Analysis of Likert-scale data: A reinterpretation of Gregoire and Driver. *Psychological Bulletin*, 105(1):167–170, 1989.

[104] J A Rice. *Mathematical Statistics and Data Analysis.* Duxbury, Belmont, California, 2nd edition, 1995.

[105] R Rosenthal and R L Rosnow. *Essentials of behavioral research: Methods and data analysis.* McGraw Hill, New York, 2nd edition, 1991.

[106] V Rousson and T Gasser. Simple component analysis. *Applied Statistics*, 53(4):539–555, 2004.

[107] A Roverato and S Paterlini. Technological modelling for graphical models: An approach based on genetic algorithms. *Computational Statistics & Data Analsysis*, 47:323–337, 2004.

[108] A Roverato and J Whittaker. Standard errors for the parameters of graphical Gaussian models. *Statistics and Computing*, 6:294–302, 1996.

[109] D B Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[110] M Ruggeri, A Biggeri, P Rucci, and M Tansella. Multivariate analysis of outcome of mental health care using graphical chain models. *Psychological Medicine*, 28:1421–1431, 1998.

[111] J Rumbaugh, I Jacobson, et al. *The Unified Modeling Language Reference Manual.* Addison-Wesley, Reading, Massachusetts, 1999.

[112] G Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–4, 1978.

[113] S C Shapiro. *Encyclopedia Of Artifical Intelligence.* Wiley, New York, 1990.

[114] L B Sheeber, E D Sorensen, and S R Howe. Data analytic techniques for treatment outcome studies with pretest/posttest measurements: An extensive primer. *Journal of Psychiatric Research*, 30(3):185–199, 1996.

[115] A Stuart, J K Ord, S Arnold, et al. *Kendall's Advanced Theory of Statistics: Classical Inference and the Linear Model*, volume 2A of *Kendall's Library of Statistics*. Arnold Publishers, 6th edition, 1998.

[116] M Studený and R R Bouckaert. On chain graph models for description of conditional independence structures. *The Annals of Statistics*, 26(4):1434–1495, 1996.

[117] B G Tabachnik and L S Fidell. *Using Multivariate Statistics*. Allyn & Bacon, Boston, Massachusetts, 4th edition, 2001.

[118] M Tew and W Waugh. *Guide to recording information about knee replacements: a manual for use in outpatient clinics and hospitals*. Department of Orthopædic Surgery, University of Nottingham, 1980.

[119] E R Tufte. *The Visual Display Of Quantitative Information*. Graphics Press, Cheshire, Connecticut, 1983.

[120] E R Tufte. *Envisioning Information*. Graphics Press, Cheshire, Connecticut, 1990.

[121] J W Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts, 1977.

[122] W F Velicer. Determining the number of components from a matrix of partial correlations. *Psychometrika*, 41:321–327, 1976.

[123] C Waternaux, N M Laird, and J H Ware. Methods for analysis of longitudinal data: blood lead concentrations and cognitive development. *Journal of the American Statistical Association*, 84:33–41, 1989.

[124] N Wermuth and S L Lauritzen. On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *Journal of the Royal Statistical Society, Series B*, 52:21–72, 1990.

[125] J Whittaker. *Graphical Models In Applied Mathematical Multivariate Statistics*. Wiley, Chichester, 1990.

[126] L Wilkinson. *The Grammar of Graphics*. Springer, New York, 2nd edition, 2005.

[127] S Wold. Cross-validatory estimation of the number of components in factor and principal component analysis. *Technometrics*, 20:397–405, 1978.

[128] S Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

[129] J Wyatt and D Spiegelhalter. Field trials of medical decision-aids: potential problems and solutions. In *Proceedings of the 15th Symposium on Computer Applications in Medical Care*, pages 3–7, Washington DC, 1991. McGraw Hill.